Current Topics in Microbiology and Immunology

Volume 363

Current Topics in Microbiology and Immunology

Previously published volumes

Further volumes can be found at www.springer.com

Vol. 332: **Karasev A. (Ed.)**:
Plant produced Microbial Vaccines. 2009.
ISBN 978-3-540-70857-5

Vol. 333: **Compans, Richard W.**; **Orenstein, Walter A. (Eds.)**:
Vaccines for Pandemic Influenza. 2009.
ISBN 978-3-540-92164-6

Vol. 334: **McGavern, Dorian**; **Dustin, Micheal (Eds.)**:
Visualizing Immunity. 2009.
ISBN 978-3-540-93862-0

Vol. 335: **Levine, Beth**; **Yoshimori, Tamotsu**; **Deretic, Vojo (Eds.)**:
Autophagy in Infection and Immunity. 2009.
ISBN 978-3-642-00301-1

Vol. 336: **Kielian, Tammy (Ed.)**:
Toll-like Receptors: Roles in Infection and Neuropathology. 2009.
ISBN 978-3-642-00548-0

Vol. 337: **Sasakawa, Chihiro (Ed.)**:
Molecular Mechanisms of Bacterial Infection via the Gut. 2009.
ISBN 978-3-642-01845-9

Vol. 338: **Rothman, Alan L. (Ed.)**:
Dengue Virus. 2009.
ISBN 978-3-642-02214-2

Vol. 339: **Spearman, Paul**; **Freed, Eric O. (Eds.)**:
HIV Interactions with Host Cell Proteins. 2009.
ISBN 978-3-642-02174-9

Vol. 340: **Saito, Takashi**; **Batista**, **Facundo D. (Eds.)**:
Immunological Synapse. 2010.
ISBN 978-3-642-03857-0

Vol. 341: **Bruserud, Øystein (Ed.)**:
The Chemokine System in Clinical and Experimental Hematology. 2010.
ISBN 978-3-642-12638-3

Vol. 342: **Arvin, Ann M. (Ed.)**:
Varicella-zoster Virus. 2010.
ISBN 978-3-642-12727-4

Vol. 343: **Johnson, John E. (Ed.)**:
Cell Entry by Non-Enveloped Viruses. 2010.
ISBN 978-3-642-13331-2

Vol. 344: **Dranoff, Glenn (Ed.)**:
Cancer Immunology and Immunotherapy. 2011.
ISBN 978-3-642-14135-5

Vol. 345: **Simon, M. Celeste (Ed.)**:
Diverse Effects of Hypoxia on Tumor Progression. 2010.
ISBN 978-3-642-13328-2

Vol. 346: **Christian Rommel**; **Bart Vanhaesebroeck**; **Peter K. Vogt (Ed.)**:
Phosphoinositide 3-kinase in Health and Disease. 2010.
ISBN 978-3-642-13662-7

Vol. 347: **Christian Rommel**; **Bart Vanhaesebroeck**; **Peter K. Vogt (Ed.)**:
Phosphoinositide 3-kinase in Health and Disease. 2010.
ISBN 978-3-642-14815-6

Vol. 348: **Lyubomir Vassilev**; **David Fry (Eds.)**:
Small-Molecule Inhibitors of Protein-Protein Interactions. 2011.
ISBN 978-3-642-17082-9

Vol. 349: **Michael Kartin (Eds.)**:
NF-kB in Health and Disease. 2011.
ISBN 978-3-642-16017-2

Vol. 350: **Rafi Ahmed**; **Tasuku Honjo (Eds.)**:
Negative Co-receptors and Ligands. 2011.
ISBN 978-3-642-19545-7

Vol. 351: **Marcel, B. M. Teunissen (Ed.)**:
Intradermal Immunization. 2011.
ISBN 978-3-642-23689-1

Vol. 352: **Rudolf Valenta**; **Robert L. Coffman (Eds.)**
Vaccines against Allergies. 2011.
ISBN 978-3-642-20054-0

Vol. 353: **Charles E. Samuel (Ed.)**:
Adenosine Deaminases Acting on RNA (ADARs) and A-to-I Editing. 2011.
ISBN 978-3-642-22800-1

Vol. 354: **Pamela A. Kozlowski (Ed.)**:
Mucosal Vaccines. 2012.
ISBN 978-3-642-23692-1

Vol. 355: **Ingo K. Mellinghoff**; **Charles L. Sawyers (Eds.)**:
Therapeutic Kinase Inhibitors. 2012.
ISBN 978-3-642-28295-9

Vol. 356: **Cornelis Murre (Ed.)**:
Epigenetic Regulation of Lymphocyte Development. 2012.
ISBN 978-3-642-24102-4

Vol. 357: **Nicholas J. Mantis (Ed.)**:
Ricin and Shiga Toxins. 2012.
ISBN 978-3-642-27469-5

Vol. 359: **Benhur Lee, Paul Rota (Eds.)**:
Henipavirus. 2012.
ISBN 978-3-642-29818-9

Vol. 360: **Freddy Radtke (Ed.)**:
Notch Regulation of the Immune System. 2012.
ISBN 978-3-642-24293-9

Vol. 361: **Klaus Aktories; Joachim H. C. Orth; Ben Adler (Eds.)**:
Pasteurella multocida. 2012.
ISBN 978-3-642-31016-4

Vol. 362: **Marco Falasca (Ed.)**:
Phosphoinositides and Disease. 2012.
ISBN 978-94-007-5024-1

Michael G. Katze
Editor

Systems Biology

Responsible series editor: Hilary Koprowski

Springer

Editor
Michael G. Katze
Department of Microbiology
University of Washington
Seattle, WA
USA

ISSN 0070-217X
ISBN 978-3-642-33098-8 ISBN 978-3-642-33099-5 (eBook)
DOI 10.1007/978-3-642-33099-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012953557

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

You hold in your hand a volume devoted to systems biology of infectious disease. If you are new to the field, you may be asking "what is systems biology?" If you think you already know the answer, you may be wondering how such an approach can be applied to a problem as complex as infectious disease. Our goal is to address both of these questions, and we anticipate that this volume will be of great interest to investigators already engaged in systems biology research as well as to those scientists and clinicians who may be seeking an introduction to the field.

What is Systems Biology?

As you read through this volume, it will become apparent that while there is no single concise definition of systems biology, most authors will settle on several key points. First, systems biology is an inter-disciplinary approach, requiring the combined talents of biologists, mathematicians, and computer scientists. Second, systems biology is holistic, with the goal of obtaining a comprehensive understanding of the workings of biological systems. This is achieved through the acquisition of massive amounts of data by high-throughput technologies—oligonucleotide microarrays, mass spectrometry, and next-generation sequencing—and the analysis of this data through sophisticated mathematical algorithms (Fig. 1). It is perhaps the use of mathematics, to integrate abundant and diverse types of data and to generate models of interconnected molecular networks, that best characterizes systems biology.

An additional characteristic often attributed to the approach is the use of an iterative cycle of experimental perturbations. Once a model has been developed, subsequent perturbations of the biological system are used to yield refinements to the model and increase its predictive capacity. While the value of a clear understanding of complex molecular networks may seem readily apparent, proponents of systems biology argue that the approach is also the only way to understand the "emergent properties" of biological systems. As described in

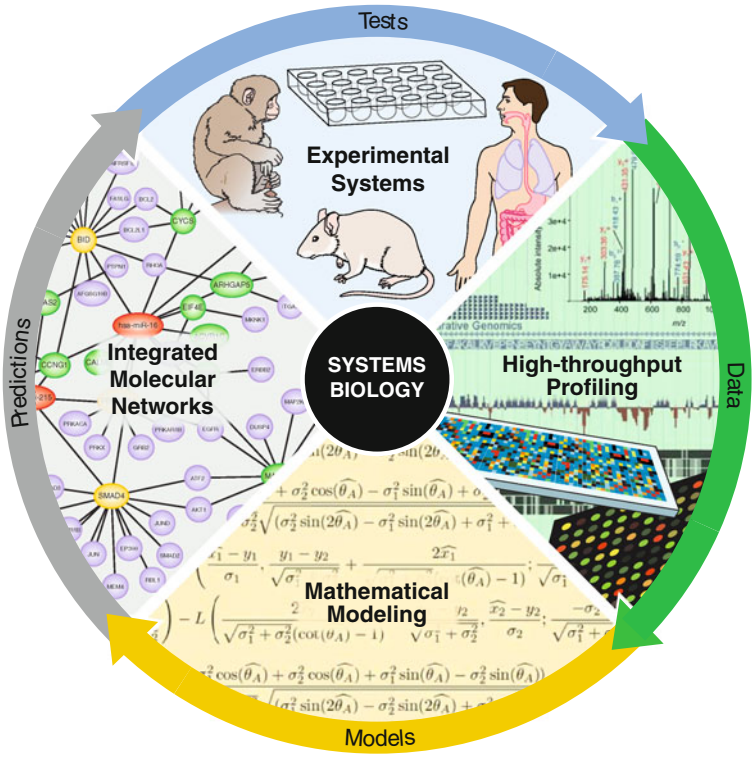


Fig. 1 The systems biology paradigm viewed as an iterative cycle of events leading to the generation of integrated models of molecular networks that serve to generate predictions for subsequent testing, model refinement, and a deeper understanding of biological processes

"Systems Approaches to Dissecting Immunity", these are properties—or biological outcomes—that cannot be predicted by an understanding of the individual parts of a system alone. Finally, systems biology typically seeks to capture information about changes in a biological system over time, providing unique insights into the dynamic nature of the system, a property that has particular relevance to infectious disease.

Why Focus on Infectious Disease?

Systems biology as we know it today was made possible by the human genome project and the advent of high-throughput technologies to measure global changes in gene transcription and protein and metabolite abundance. The first uses of this approach just over a decade ago focused on the systematic perturbation of yeast and the mathematical modeling of metabolic pathways (Ideker et al. 2001). Given

the complexity of even a single-cell organism, many would argue (and some still do) that the approach is ill-suited for multi-cellular organisms or mammalian systems. Yet the cancer field rapidly embraced the approach and has proven its utility for network-based classification and prognosis of breast cancer, the identification of oncogenes in B-cell lymphomas, and improvements to radiation therapy (Laubenbacher et al. 2009).

The infectious disease field, in contrast, has come rather late to the game. Although our own group published the first genomic analysis of HIV-infected cells in culture (Geiss 2000), and numerous reports of transcriptional profiling of virus-infected cells and tissues have followed, the application of a true systems biology approach to infectious disease has until only recently been considered too daunting. What has brought about the change in attitude? Recent and dramatic improvements to mathematical modeling (see "Studying Salmonellae and Yersiniae Host–Pathogen Interactions Using Integrated 'Omics and Modeling" and "Insights into Proteomic Immune Cell Signaling and Communication via Data-Driven Modeling") and the success of the approach in other fields are certainly contributing factors, but perhaps most important is the growing realization that the infectious disease field desperately needs to take new approaches to solve long unanswered challenges, particularly in the areas of vaccine and drug development.

Trying to understand the countless and complex pathogen–host interactions and intra- and inter-cellular signaling events that occur during the course of infectious disease is indeed a formidable task. Historically, a reductionist approach was both the most tractable and only available line of attack. But clearly a new approach is needed. Vaccines against numerous deadly diseases, most notably AIDS, malaria, and tuberculosis are still lacking. Drug-resistant viruses and bacteria continue to emerge, a trend that is likely to endure as long as microbial targets remain the focus of new drug development, and the focus on microbial targets also yields drugs that are typically narrow in spectrum. As described throughout this volume, systems biology offers a new and holistic approach to understanding pathogen–host interactions, the innate immune response, and the mechanisms that lead to disease, immunopathology, or protective immunity. The approach holds enormous potential, but there are challenges as well.

Risks and Rewards

No doubt everyone engaged in systems biology research has heard the criticism that the approach is nothing more than an expensive fishing expedition that takes funding away from individual investigators. The relative merits of big versus small science aside, if systems biology is a fishing expedition, the chapters in this volume show that the approach is beginning to make some nice catches. For example, systems biology is accelerating vaccine development by increasing our understanding of how protective immune responses are elicited "Systems Biology

of Vaccination in the Elderly". Similarly, by providing a better understanding of the host response to infection, the approach is facilitating the development of drugs that target the host side of the pathogen–host interaction "Systems Biology Analyses to Define Host Responses to HCV Infection and Therapy", an approach that will yield drugs that are broader in spectrum and less prone to microbial resistance. Moreover, systems biology is beginning to deliver on its much touted potential for yielding biomarkers for new diagnostic and prognostic applications "Systems Biology Approach for New Target and Biomarker Identification".

Nevertheless, the approach is expensive, and with ever-tightening budgets, more money for systems biology means less money elsewhere. Moreover, because the approach has been extensively hyped as being revolutionary, expectations have been set high, and many are understandably disappointed with the pace of progress. The extent to which systems biology represents a true paradigm shift has also been called into question (Bothwell 2006). And there are still plenty of technical, scientific, and mathematical hurdles to overcome. Even the choice of experimental systems can be a challenge. The jump from cell culture systems to nonhuman primates, for example, represents an enormous leap in system complexity that taxes every aspect of the approach, particularly computational and modeling techniques. Yet, as discussed in "The Role and Contributions of Systems Biology to the Non-Human Primate Model of Influenza Pathogenesis and Vaccinology" and "'Omics Investigations of HIV and SIV Pathogenesis and Innate Immunity", significant progress is being made, and the analysis of biologically relevant infection models is essential if we are to understand the processes of disease and immunity and translate findings into rational drug design and vaccine development.

In This Volume

We begin this volume with an engaging editorial by Dr. Valentina Di Francesco and colleagues, who oversee a broad portfolio of systems biology research contracts at the National Institute of Allergy and Infectious Diseases (NIAID). NIAID has made a substantial commitment to systems biology through the sponsorship of genomic, proteomic, and bioinformatic resource centers, and more recently through the funding of a systems biology for infectious disease research program. This program is aimed at using experimental and computational approaches to analyze, model, and predict the architecture and dynamics of the molecular networks underlying the initiation and progression of infectious disease (Aderem et al. 2011). Each of the primary investigators associated with this program have provided material for this volume.

The chapters of *Systems Biology* provide the reader with cutting-edge research from leaders in the systems biology field. The initial chapter provides both a concise overview of the systems biology paradigm as well as an excellent discussion of how this approach is being used to dissect the innate immune system.

Subsequent chapters are devoted to systems biology approaches to bacterial–host interactions (including *Salmonella*, *Yersinia*, and *Mycobacterium*), where molecular events within the pathogen are as important as the host response to the invading microbe; the application of high-throughput and computational approaches to nonhuman primate models of influenza and AIDS; and an overview of the emerging field of systems vaccinology, where systems biology is changing the way we think about vaccine design and testing. Final chapters are dedicated to defining the host response to hepatitis C virus infection and therapy, to drug target and biomarker identification, and to new computational approaches, including data-driven modeling. By assembling a diverse spectrum of perspectives and expertise, it is hoped that the information provided here will serve as a catalyst for additional innovative approaches that will continue to drive the field forward and that will ultimately transform how we view, treat, and protect against infectious disease.

Seattle, Washington, July 2012

Marcus J. Korth
Michael G. Katze

References

Aderem A, Adkins JN, Ansong C, Galagan J, Kaiser S, Korth MJ, Law GL, Mcdermott JE, Proll SC, Rosenberger G, Schoolnik G, Katze MG (2011) A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. mBio 2:e00325-00310

Bothwell JH (2006) The long past of systems biology. New Phytol 170:6−10

Geiss GK, Bumgarner RE, An MC, Agy MB, Van 'T Wout AB, Hammersmark E, Carter VS, Upchurch D, Mullins JI, Katze MG (2000) Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. Virology 266:8−16

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292:929−934

Laubenbacher R, Hower V, Jarrah A, Torti SV, Shulaev V, Mendes P, Torti FM, Akman S (2009) A systems biology view of cancer. Biochim Biophys Acta 1796:129−139

Acknowledgments

We thank Cynthia Baker for her invaluable help in assembling this volume, Sean Proll for assistance with figure production, and Patrick Lane for the final systems

biology graphic. Research in the Katze laboratory is supported by Public Health Service grants R2400011172, R2400011157, P30DA015625, P51RR00166, and U54AI081680 and by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract HHSN272200800060C.

Introduction: Embracing Complexity in Infectious Disease Research

The concept of systems biology is not new, in fact reflecting on work done in the 1960s, British biologist Denis Noble described systems biology as "… putting together rather than taking apart, integration rather than reduction. … It requires that we develop ways of thinking about integration that are as rigorous as our reductionist procedures, but different … it means changing our philosophy, in the full sense of the term" (Noble 2006).

Infectious disease research seems an ideal target on which to apply a systems biology approach to understand an infectious agent, its host biology in response to infection, and the dynamic nature of the pathogen and host interactions over the course of disease. Traditional experimental approaches to infectious disease research have focused mostly on subsets of the virulence process or on particular events that occur during infection limiting both the speed and ability to understand the complex process as a whole. For example, the study of individual genes, operons and regulons provides necessary insights into the workings of infection, but it does not offer a comprehensive framework for the interaction of the regulatory networks of the infectious agent and the host cells. Comprehensive identification of the cellular and molecular components of the pathogen and its host, and characterization of the functional role and mechanisms involved in the interaction require the use of advanced high throughput (HTP) technologies. High-performance computational resources and sophisticated analysis tools are now available to make sense of the enormous datasets generated. In spite of the difficulties and challenges, the complexity of biological systems can now finally be embraced.

The National Institute of Allergy and Infectious Diseases (NIAID), part of the National Institutes of Health (NIH), supports research to better understand, treat, and prevent infectious, immunologic, and allergic diseases. Given the potential of microbial genomic research, in the last few years NIAID has made a significant investment in genomic-related programs that provide to the scientific community comprehensive, publicly accessible resources for genome sequencing, transcriptomics, proteomics and bioinformatics, as well as rapid release of data and reagents for

basic and applied research, in support of the Institute's mission (http://www.niaid.nih.gov/topics/pathogenGenomics/research/Pages/relatedInitiatives.aspx). To leverage the availability of advanced technologies for genomics research in 2008, NIAID established the Systems Biology Program (SBP) for infectious diseases research to encourage a shift in thinking toward a more global and high-throughput approach to basic research on infectious diseases in order to gain insight into the biology of microbes, their role in pathogenesis, and their molecular interactions with the host. The SBP utilizes computational and experimental high-throughput methodologies to identify, analyze, quantify, model, and predict the structure and dynamics of molecular networks involved in host/pathogen interactions (Aderem et al. 2011). High-throughput methodologies often include next generation sequencing, transcriptomics, proteomics, metabolomics, and lipidomics. By encouraging a systems biology approach, NIAID also fosters multidisciplinary teams—including statistical modelers, computational biologists, experts in HTP genomics technologies, microbiologists, and clinicians—to tackle the complexity of infectious disease research in a more comprehensive fashion.

As in other systems biology programs, the NIAID SBP has been experiencing a number of challenges and ideal solutions are yet to be found. Comprehensive data analysis and biological interpretation of the experimental results appear to be the largest hurdles. Contemporary HTP technologies generate unprecedented amounts of experimental data. In a systems biology project, an efficient and well-designed data management infrastructure is crucial for data analysis. Experimental data must be organized and systematically maintained to allow for long-term storage, accessibility, and fast retrieval by multiple research laboratories that may be disseminated geographically while participating in joint projects.

Mathematical modeling of the behavior of biological systems in response to the tested experimental conditions is indispensable to integrate and effectively summarize the experimental data; to identify missing information (e.g. predicted essential genes that seem to be missing from the annotated genomes); to predict the systems' response to perturbations; and to suggest the next experiment. However, establishing accurate models with high confidence levels depends on enough amounts of data to set parameters, constraints, and to cross-validate, hence increasing the need for even more experimental data.

In addition, systems biology research is sometimes conducted primarily with in vitro experimentation rather than with in vivo/ex vivo systems to allow for better control of the experimental system (e.g. temperature, chemistry, infection time course) and measurements that can be replicated more easily and economically. Nevertheless, the most promising observations derived from in vitro systems are generally tested and validated in the relevant animal and human tissues. Also, whenever technically feasible, multiple HTP technologies should be applied concurrently to the same biological samples. This can present a challenge because sample preparation protocols for different genomic technologies may alter the characteristics of the sample.

The power of the systems biology approach can only be fully exploited by ensuring that—in addition to publications—the generated data, associated metadata, original experimental design and protocols, and resulting models are made easily accessible to the broad scientific community of infectious disease researchers. This is especially important given the abundance of data generated by typical systems biology projects, the limitations of the current analysis and modeling approaches, and the budgetary constraints that limit the number of validation studies that can be performed within any funded project. For that reason, NIAID is requiring that funded systems biology projects share with the broad scientific community all the generated 'omics' data and metadata, resources, and novel reagents through publicly accessible databases and reagent repositories. Still, it appears that infectious disease researchers in general need to become more familiar with the advanced technologies, analysis tools, and computational approaches—of the systems biology approach in order to take full advantage of the shared resources and further pursue much needed validation studies.

The systems biology approach is gradually being adopted by infectious disease researchers. The slow pace of adoption is not surprising, since it parallels what happened in the past with the introduction of new research tools that are generally refined over many years before they are adopted for widespread use. New methods are typically more expensive and lack evidence of reliability, accuracy, and value. However, new methods usually become more precise and economical over time and that is proving to be the case with systems biology.

HTP experimental work, data analysis, and model development are performed by technology experts, bioinformaticians, and computational scientists respectively. Biologists and infectious disease scientists more often provide interpretation of the data in the context of the biological systems being investigated and contribute to the design of important follow-up validation studies with more traditional 'reductionist' approaches (e.g., phenotype characterization of gene knock-outs) to pursue the most promising new hypotheses gleaned from the data. While the controversy in the scientific community continues as to where the balance should lie between the 'reductionist' and the 'systems' approach in biomedical research, the controversy should be settled as traditional methodologies are still needed to fill in the details of specific biological events, with systems biology acting as a hypothesis generator pointing to areas needing further investigation.

NIH supports a broad array of basic and clinical research. Systems biology ultimately seeks to improve health by approaching disease not as a pathology in individual biochemical pathways of a particular cell type in a single organ, but as the result of complex and interdependent processes. The long-term expectation is that systems biology will more quickly identify the biological processes involved in disease and facilitate the development of therapeutic strategies, vaccines, and diagnostics based on a more comprehensive and systems-wide understanding of the

mechanisms implicated in the disease processes. Several chapters in this book already demonstrate the promise and initial successes of the systems biology approach.

William Alexander
Peter A. Dudley
Valentina Di Francesco
Division of Microbiology and Infectious Diseases
National Institute of Allergy and Infectious Diseases
National Institutes of Health
Bethesda, MD, USA
E-mail: vdifrancesco@niaid.nih.gov

References

Aderem A, Adkins JN, Ansong C et al (2011) A systems biology approach to infectious diseases research: innovating the pathogen-host research paradigm. MBio 2(1): e00325-10

Noble D (2006) The music of life: biology beyond the genome. Oxford University Press, USA

Contents

Contributors

Alan Aderem Seattle Biomedical Research Institute, 307 Westlake Ave N, Suite 500, Seattle, WA 98109, USA, e-mail: alan.aderem@seattlebiomed.org

Joshua N. Adkins Pacific Northwest National Laboratory, Biological Separations and Mass Spectroscopy Group, PO Box 999, MSIN: K8-98, Richland, WA 99352, USA, e-mail: Joshua.Adkins@pnnl.gov

Charles Ansong Pacific Northwest National Laboratory, Biological Separations and Mass Spectroscopy Group, PO Box 999, MSIN: K8-98, Richland, WA 99352, USA

Carole Baskin Science Foundation Arizona, 400 East Van Buren Street, Phoenix, AZ 85004, USA, e-mail: cb2@u.washington.edu

Arndt G. Benecke Centre National de la Recherche Scientifique, Institut des Hautes Études Scientifiques, 35 route de Chartres, 91440 Bures sur Yvette, France, e-mail: arndt@ihes.fr

Kelly F. Benedict Department of Biological Engineering, Massachusetts Institute of Technology, Room: 16-343, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Roslyn N. Brown Pacific Northwest National Laboratory, Biological Separations and Mass Spectroscopy Group, PO Box 999, MSIN: K8-98, Richland, WA 99352, USA

Mary M. Cavanagh Department of Medicine, Stanford University, Stanford, CA 94305, USA

Pep Charusanti Department of Bioengineering, University of California-San Diego, La Jolla, CA, USA

Sadhana Chauhan Departments of Pathology and Microbiology and Immunology, University of Texas Medical Branch, Galveston, TX, USA

Brooke L. Deatherage Pacific Northwest National Laboratory, Biological Separations and Mass Spectroscopy Group, PO Box 999, MSIN: K8-98, Richland, WA 99352, USA

Alan Diercks Seattle Biomedical Research Institute, 307 Westlake Ave N, Suite 500, Seattle, WA 98109, USA

Sai S. Duraisingham Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, 954 Gatewood Road, Atlanta, GA 30329, USA

Deborah H. Fuller Department of Microbiology, University of Washington, Seattle, WA, USA; Washington National Primate Research Center, Seattle, WA, USA, e-mail: fullerdh@u.washington.edu

James Galagan Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA; Department of Microbiology, Boston University, Boston, MA 02215, USA; Bioinformatics Program, Boston University, Boston, MA 02215, USA; The Eli and Edythe L. Broad Institute of Harvard, MIT, Cambridge, MA 02142, USA

Michael Gale Jr. Department of Immunology, University of Washington School of Medicine, 1959 NE Pacific Street, Box 357650, Seattle, WA 98195, USA, e-mail: mgale@u.washington.edu

Antonio Gomes Bioinformatics Program, Boston University, Boston, MA 02215, USA

Jorg J. Goronzy Department of Medicine, Stanford University, Stanford, CA 94305, USA; Department of Medicine, Palo Alto Veteran Administration Health Care System, Palo Alto, CA 94304, USA

Fred Heffron Department of Molecular Microbiology and Immunology, Oregon Health and Sciences University, Portland, OR, USA

Daniel Hyduke Department of Bioengineering, University of California-San Diego, La Jolla, CA, USA

Reneé C. Ireton Department of Immunology, University of Washington School of Medicine, 1959 NE Pacific Street, Box 357650, Seattle, WA 98195, USA

Marcus B. Jones J. Craig Venter Institute, Rockville, MD, USA

Afshan Kidwai Department of Molecular Microbiology and Immunology, Oregon Health and Sciences University, Portland, OR, USA

Young-Mo Kim Pacific Northwest National Laboratory, Biological Separations and Mass Spectroscopy Group, PO Box 999, MSIN: K8-98, Richland, WA 99352, USA

Douglas A. Lauffenburger Department of Biological Engineering, Massachusetts Institute of Technology, Room: 16-343, 77 Massachusetts Avenue, Cambridge, MA 02139, USA, e-mail: lauffen@mit.edu

Jie Li Department of Molecular Microbiology and Immunology, Oregon Health and Sciences University, Portland, OR, USA

Andrey Loboda Informatics and Analysis, Merck Research Laboratory, West Point, PA 19486, USA

Anna Lyubetskaya Bioinformatics Program, Boston University, Boston, MA 02215, USA

Kathleen McAteer Biology Program, Washington State University Tri-Cities, Richland, WA, USA

Jason E. McDermott Computational Biology and Bioinformatics Group, Pacific Northwest National Laboratory, Richland, WA, USA

Thomas O. Metz Pacific Northwest National Laboratory, Biological Separations and Mass Spectroscopy Group, PO Box 999, MSIN: K8-98, Richland, WA 99352, USA

Vladimir Motin Departments of Pathology and Microbiology and Immunology, University of Texas Medical Branch, Galveston, TX, USA

Helder I. Nakaya Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, 954 Gatewood Road, Atlanta, GA 30329, USA

Ernesto S. Nakayasu Pacific Northwest National Laboratory, Biological Separations and Mass Spectroscopy Group, PO Box 999, MSIN: K8-98, Richland, WA 99352, USA

David Nickle Informatics and Analysis, Merck Research Laboratory, West Point, PA 19486, USA

George Niemann Department of Molecular Microbiology and Immunology, Oregon Health and Sciences University, Portland, OR, USA

Robert E. Palermo Washington National Primate Research Center, Seattle, WA, USA; Department of Microbiology, University of Washington, Seattle, WA, USA, e-mail: palermor@u.washington.edu

Bernhard O. Palsson Department of Bioengineering, University of California-San Diego, La Jolla, CA, USA

Scott N. Peterson J. Craig Venter Institute, Rockville, MD, USA

Oscar Puig Informatics and Analysis, Merck Research Laboratory, West Point, PA 19486, USA

Bali Pulendran Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, 954 Gatewood Road, Atlanta, GA 30329, USA, e-mail: bpulend@emory.edu

Christopher Roberts Informatics and Analysis, Merck Research Laboratory, West Point, PA 19486, USA

Nadine Rouphael Division of Infectious Diseases, Department of Medicine, Emory University School of Medicine, Atlanta, GA 30329, USA

Brian Schmidt Department of Bioengineering, University of California-San Diego, La Jolla, CA, USA

Richard D. Smith Pacific Northwest National Laboratory, Biological Separations and Mass Spectroscopy Group, PO Box 999, MSIN: K8-98, Richland, WA 99352, USA

David J. Stone Informatics and Analysis, Merck Research Laboratory, West Point, PA 19486, USA

I-Ming Wang Informatics and Analysis, Merck Research Laboratory, West Point, PA 19486, USA; Merck Corporation, PO Box 100, Whitehouse Station, NJ 08889-0100, USA, e-mail: I_ming_wang@merck.com

Systems Approaches to Dissecting Immunity

Alan Diercks and Alan Aderem

Abstract Systems biology is the comprehensive and quantitative analysis of the interactions between all of the components of biological systems over time. Cells of the innate immune system are the first line of defense against invading pathogens and orchestrate the ensuing adaptive response, which is critical to the establishment of long-term protective immunity. Innate immunity is well suited for systems analysis, because the relevant cells can be isolated in various functional states and many of their interactions can be reconstituted in a biologically meaningful manner. Application of the tools of systems biology to the innate immune system will enable comprehensive analysis of the complex interactions that maintain the fine balance between host defense and inflammatory disease. In this review, we discuss innate immunity in the context of the systems biology concepts, emergence, robustness, and modularity. We also describe recent efforts to apply these approaches to enable rational vaccine design and accelerate the pace of clinical vaccine trials.

Contents

A. Diercks · A. Aderem (✉)
Seattle Biomedical Research Institute, 307 Westlake Ave N, Suite 500, 98109 Seattle, WA, USA
e-mail: alan.aderem@seattlebiomed.org

Current Topics in Microbiology and Immunology (2013) 363: 1–19
DOI: 10.1007/82_2012_246

Published Online: 11 August 2012

1 Systems Biology

For the purpose of this review, we define systems biology as the comprehensive, quantitative, and temporal analysis of the manner in which all of the components of a biological system interact. Systems biology is a holistic rather than reductionist approach to deciphering complexity and understanding emergent properties. This approach requires the capture and integration of measurements from as many hierarchical levels of information as possible. These can include DNA sequences, RNA and protein measurements, protein–protein and protein–DNA interactions, biomodules, signaling and gene regulatory networks, cells, organs, individuals, populations, and ecologies. Raw measurements are then imported and annotated into comprehensive databases, many of which are accessible online to the scientific community. Both detailed graphical visualizations and mathematical modeling are used to integrate the vast quantities of individual data points into molecular networks that underlie the biology of the system. These models suggest specific hypotheses that are tested experimentally by selective molecular perturbations thereby tying the phenotypic features of the system directly to the behavior of protein and gene regulatory networks. Repeated cycles of iteration refine the model; ultimately, these models will explain the systems or emergent properties of the biological system of interest. Once a model is sufficiently accurate and detailed, it will allow biologists to accomplish two tasks never possible before: (1) predict the behavior of the system given any perturbation and (2) redesign or perturb the gene regulatory networks to create completely new emergent properties. This latter possibility lies at the heart of preventive medicine. Thus, systems biology is hypothesis driven, global, quantitative, iterative, integrative, and dynamic.

Maximizing the potential of systems approaches requires an interdisciplinary team of investigators that is also capable of developing the novel technologies and computational tools. In this model, biology dictates what new technology and computational tools should be developed. These tools often open new frontiers in biology that go well beyond the original question, driving an iterative cycle of development and discovery. Thus, biology drives technology and computation, and, in turn, technology and computation revolutionize biology. Biological systems, as opposed to engineered man-made systems, are not the result of a rational design process, but rather the result of a random evolutionary process that selects

only for function. For this reason, reverse-engineering approaches predicated on the assumption of a rational underlying design will often fail to unravel a biological system.

1.1 Basic Concepts Crucial in Understanding Complex Biological Systems: Emergence, Robustness, and Modularity

1.1.1 Emergence

Complex systems display 'emergent properties' that are not present in their individual parts and cannot be predicted even with a full understanding of the parts alone. The arch is an example of an emergent property that arises from simple constituents. A comprehensive analysis of the physical properties of rocks will not predict that they give rise to an arch when assembled in a specific context. Life is emergent and not inherent in the individual components of an organism. Simply mixing DNA, RNA, proteins, carbohydrates, and lipids does not generate a biological system: life is a consequence of the specific organization of these components and interactions between them. A systems approach is, therefore, necessary to understand how the emergent properties of living organisms are derived from their individual components.

1.1.2 Robustness

Biological systems tend to maintain phenotypic stability despite diverse perturbations from the environment, stochastic events, and genetic variation. Robustness often arises as an emergent property through positive and negative feedback loops and other forms of regulatory control that constrain gene outputs at the transcriptional, translational, or post-translational levels. These feedback mechanisms insulate the system from environmental fluctuations. Robustness is also achieved through redundancy of pathways that perform the same biological function.

1.1.3 Modularity

A network module can be defined as a set of nodes that interact strongly and perform common function. Modularity can contribute to robustness by confining damage to independent parts, preventing the spread of damage to the entire network. Modularity can also contribute to evolution of the system, where adaptation can be achieved by rewiring connections between modules rather than reconstituting the modules themselves.

1.2 A Systems Biology Approach to Studying Immunity

The complex interactions within the innate immune system that result in effective host defense under normal conditions and inflammatory disease when perturbed can only be dissected in a comprehensive way by systems biology approaches. Immunology is particularly well suited for such analysis, as the cells can be isolated in various functional states and many aspects of the immune response can be reconstituted in a biologically meaningful manner. In this review, we highlight by example three aspects of the immune response that are particularly well suited to analysis by systems-level approaches. First, we describe the complex interactions between the multitude of phagocytic and pattern-recognition receptors that initiate the immune response to invading microbes. Next, we discuss a transcriptional regulatory circuit that tunes inflammatory responses in macrophages and discriminates transient from persistent stimulation. Finally, we describe how the tools of systems biology can be used to gain an understanding of the molecular and cellular interactions that govern the response to vaccination. We argue that systems approaches offer a route to accelerating the pace of efficacy trials by identifying correlates of protection.

2 Innate Immune Receptors

The recognition, phagocytosis, and presentation of pathogens by macrophages represent emergent properties that arise from the concerted action of a number of receptors and signaling pathways. Specific pathogen-derived molecules are detected by chemotactic receptors on the macrophage, leading to alterations in the cytoskeleton that culminate in directed movement. The macrophage then uses pattern-recognition receptors (PRRs), which include the Toll-like receptors (TLRs), the NOD-like receptors (NLRs), and the RIG-I-like receptors (RLRs), to identify the nature of the pathogen by recognizing specific pathogen-associated molecular patterns (PAMPs). Phagocytic receptors, such as the Fc receptor, the complement receptor, and DECTIN, bind the particle and activate signaling pathways that lead to its internalization (Underhill and Ozinsky 2002). Upon internalization, the pathogen is degraded, and pathogen-derived antigens are presented to cells of the adaptive immune system; this process of antigen presentation constitutes the mechanism by which the innate immune system instructs adaptive immunity.

It is not possible to predict the complex behavior underlying chemotaxis, phagocytosis, and antigen presentation by having a complete understanding of each individual receptor and its cognate signaling pathway in isolation. Systems biology approaches will enable an understanding of how the crosstalk between these pathways results in the emergent properties that give rise to these functional responses.

2.1 Crosstalk Between Phagocytic Receptors and PRRs

It has long been known that phagocytosis can be uncoupled from the induction of an inflammatory response (Aderem et al. 1984, 1985). For example, phagocytosis of latex beads is not accompanied by the production of arachidonic acid metabolites unless the macrophages are primed with bacterial lipopolysaccharide (LPS), in which case a synergistic response is observed (Aderem et al. 1986). Similar synergy also occurs for Fc receptor and zymosan-induced phagocytosis but not for complement-induced phagocytosis, which will not induce arachidonic acid metabolite release even with LPS priming (Aderem et al. 1986). These interactions are even more subtle when considering the internalization of bacteria. When macrophages internalize Gram-negative bacteria, tumor necrosis factor (TNF) is only produced in the presence of TLR4. By contrast, TLR2 is required for TNF production during phagocytosis of Gram-positive bacteria (Underhill and Ozinsky 2002).

Phagocytosis of fungal zymosan provides an example of how phagocytic and PRR pathways can function as interlocking pieces in their regulation of the macrophage response (reviewed in Goodridge and Underhill 2008). Zymosan is recognized by both TLR2 and DECTIN. TLR2 signaling induces inflammatory cytokines through the MyD88 pathway and activation of NF-κB but does not induce reactive oxygen species (ROS), phagocytosis, and only weak arachidonic acid release. DECTIN, which recognizes β-glucan, activates Syk kinase, induces zymosan phagocytosis, ROS induction, and weak arachidonic acid release. When both TLR2 and DECTIN are activated, inflammatory cytokine induction, ROS production, and arachidonic acid metabolism are all synergistically enhanced.

Interactions between PRR signaling and phagocytic pathways extend beyond internalization and inflammation. TLR signaling has been implicated in the enhanced maturation of phagosomes (Blander and Medzhitov 2004). More importantly, the presence of TLR ligands within a dendritic cell phagosome markedly enhances the MHC class II-mediated presentation of antigens within that phagosome (Blander and Medzhitov 2006). Thus, the entire set of functional macrophage responses to pathogens are shaped and modulated by complex interactions between PRR, phagocytic, and other pathways.

2.2 Crosstalk Between PRRs

Macrophages are not confronted with purified PAMPs in nature. Rather, they interact with complete pathogens that present a cocktail of agonists to the numerous PRRs they express (Underhill and Ozinsky 2002; Trinchieri and Sher 2007). These combinations of PAMPs enable the innate immune cell to carry out 'multiparameter analysis', which permits far greater accuracy in the determination

of the threat. For example, if TLR4, TLR5 and the NLR IL-1β-converting enzyme protease-activating factor (IPAF) are simultaneously activated, the cell can compute that it has encountered a Gram-negative flagellated bacterium that contains a type III secretion system (Miao et al. 2006). Activation of TLR4 and TLR5 culminates in NF-κB-dependent inflammatory gene expression while detection of flagellin by IPAF recruits (Geddes et al. 2001; Poyet et al. 2001) and activates the caspase-1 inflammasome (Masumoto et al. 2003) which processes IL-1β and IL-18 for secretion (Dinarello 1998).

Dual sensing of flagellin by TLR5 and IPAF suggests a complex, two-step process for regulating the response to invading bacteria. When a macrophage encounters a *Salmonella* bacterium, TLR5 is initially stimulated by flagellin (in addition to activation of TLR4 by LPS). This signal induces, among others, the mRNAs encoding IL-1β and IL-18 and their precursor proteins. Once the bacterium is in the phagosome, flagellin is injected into the cytoplasm via the type III secretion system, and IPAF is subsequently activated. Conceptually, TLR signaling in the absence of NLRs may constitute a 'yellow alert', indicating that microbes have penetrated the physical barrier of the epithelial layer. The inflammasome NLRs, when activated in conjunction with the TLRs, may then trigger a 'red alert', alerting the immune system to the presence of pathogens which harbor more threatening virulence factors such as the type III secretion system (Miao et al. 2007). Signaling by TLRs alone or by NLRs alone does not initiate the red alert, and thus the red alert emerges from the convergent activation of the two pathways. IL-1β is not known to be capable of activating the inflammasome itself, and thus paracrine IL-1β signaling can propagate the yellow alert but not the red alert, which is reserved for the infected macrophage. A similar distinction between the reserved red alert status of the infected cell and the yellow alert status for neighboring cells activated by paracrine cytokine signaling has been postulated for viral nucleic acid detection (Stetson and Medzhitov 2006): cytotoxic lymphocytes and natural killer cells must be able to distinguish between virus-infected cells that should be targeted for apoptosis and cells that have been activated into an antiviral state by paracrine type I IFN signaling.

The system is even more complex than described due to crosstalk arising from simultaneous engagement of multiple TLRs and other receptor families. Viral RNA is recognized by at least five PRRs [TLR3, TLR7, TLR8, melanoma differentiation-associated gene 5 (MDA5), and RIG-I], and it is interesting to speculate on how convergent detection can lead to synergistic, virus-specific responses. Results from the Akira laboratory (Kumar et al. 2008) suggest that the adjuvant effects of the double-stranded RNA (dsRNA) analog polyinosinic–polycytidylic acid (polyI:C) arise from cooperative activation of TLR and cytoplasmic RLR pathways. Thus, pathogen recognition by the innate immune system is perhaps best considered as a process in which activation of several PRR pathways in combination gives rise to an emergent, pathogen-specific response that seeks to neutralize the threat, alert neighboring cells to the presence of microbes, and initiate an appropriate adaptive immune response.

2.3 Robustness and Modularity in Innate Immunity

While combinatorial PAMP detection by PRR pathways allows macrophages to accurately determine threat levels posed by invading pathogens, it also illustrates two additional key properties of the innate immune system: robustness and modularity.

To provide protection, the innate immune system must be robust: pathogens must be detected and the immune system alerted, even as evolution favors development of pathogen strategies to evade detection. The large number of PAMPs that may be detected by macrophage PRRs thus constitutes a robust, 'fail-safe' detection system: if a particular PRR fails to detect a pathogen, or if a pathogen evolves a strategy to evade a particular PRR, it nevertheless will be detected by all of the relevant remaining PRRs expressed by the cell. This level of robustness is revealed by gene-targeting studies, in which specific PRR knockouts or knockdowns fail to exhibit phenotypes. For example, TLR3, which detects viral dsRNA, when ablated does not result in universally enhanced susceptibility to viral infection (Edelmann et al. 2004), presumably because signaling by other viral RNA detectors (RIG-I, MDA5, and TLR7) is sufficient for protection. Similarly, we have demonstrated that inflammasome activation in response to *Listeria monocytogenes* involves detection by three or more cytoplasmic receptors: IPAF, NALP3, and at least one other NLR utilizing the adapter ASC (apoptosis-associated speck-like protein containing a C-terminal caspase recruitment domain) (Warren et al. 2008).

Modularity in the PRR pathways is typified by the modularity in the structures of the PRRs themselves. In the TLR family, for example, a less conserved N-terminalleucine-rich repeat (LRR) domain is coupled to a more highly conserved C-terminal Toll/Interleukin-1 receptor (TIR) domain (Roach 2005) by a single transmembrane domain. The LRR domains are so variable that they cannot be aligned over large evolutionary distances; alignment can only be accomplished using the TIR domains. The TIR domain couples the TLR to the restricted set of adapters [the linker adapters, MyD88 adapter-like protein (MAL) and translocating chain-associating membrane protein (TRAM), and the major signaling adapters, MyD88 and TRIF], whereas the LRR domain is responsible for PAMP recognition. Evolution of LRRs has resulted in an extraordinary diversity in ligands detected by the TLRs, giving rise to six major TLR families in vertebrates (Roach 2005). Structural studies of TLR–ligand complexes have revealed diversity in LRR ligand binding mechanisms (reviewed in (Jin and Lee 2008)). While TLR2/TLR1 heterodimer binding of Pam_3CSK_4 is achieved by hydrophobic interactions at the boundary between central and C-terminal domains (Jin et al. 2007), TLR3 dimers bind dsRNA at two regions near the N-terminal and C-terminal ends (Liu et al. 2008).

Robustness in the innate immune system emerges not only from the modularity of the PRRs and the pathways they activate but also from the feedback

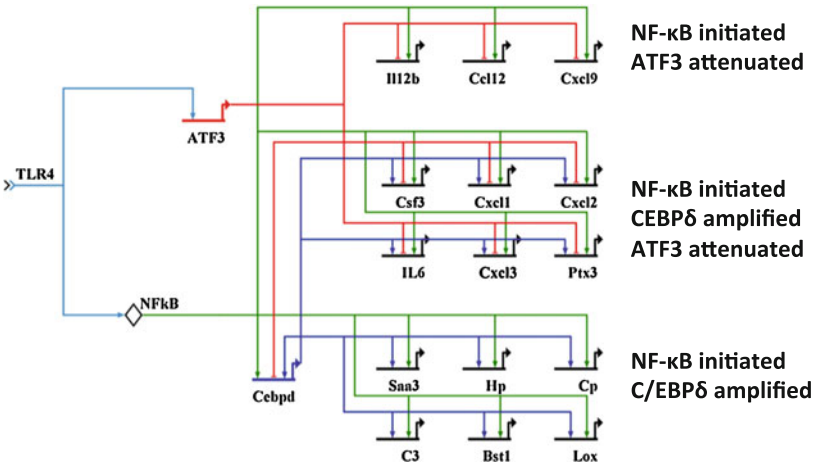


Fig. 1 Regulation of cytokine production in macrophages by the NF-κB, ATF3, and C/EBPδ circuit. Stimulation of TLR4 activates NF-κB, which initiates transcription of a number of cytokines. ATF3 is also activated and represses transcription of a subset of these cytokines. NF-κB and ATF3 also act similarly to modulate transcription of Cebpd which amplifies the transcription of a subset of cytokines as well as itself. Cytokines were classified into three categories based on genome-wide localization analysis of NF-κB, ATF3, and C/EBPδ and the classification confirmed by measuring transcriptional responses in ATF3$^{-/-}$ and Cebpd$^{-/-}$ macrophages. (Adapted from (Alon 2007))

architectures of the pathways themselves. Type I IFN induction by cytoplasmic viral sensors in fibroblasts is an example of a positive feedback loop which results in robust induction of an antiviral state (reviewed in Honda et al. 2006). Cytoplasmic detection of viral RNA by the RLRs RIG-I or MDA5 results in type I IFN induction by activated IFN regulatory factor-3 (IRF3) and IRF7 transcription factors. The type I IFN then feeds back on the cells in an autocrine manner to induce IRF7 to high levels. IRF7 then induces additional type I IFN species and increases the expression of the sensors RIG-I and MDA5 themselves, which presumably renders the cell more sensitive to viral RNA. On the other hand, precise control and robustness to intracellular noise is partly achieved in PRR pathways by negative feedback loops. For example, TLRs induce the expression of many genes that negatively regulate the TLR pathways (reviewed in Liew et al. 2005). In particular, the ubiquitin-editing protein A20 (Tnfaip3) is both induced by and is a negative regulator of TLR, RLR, and NLR pathways (Boone et al. 2004; Wang 2004; Saitoh et al. 2005; Lin et al. 2006; Hitotsumatsu et al. 2008), acting directly on key adapter molecules such as tumor necrosis factor receptor-associated factor 6 (TRAF6), TRIF, and receptor-interacting protein 2 (RIP2). The second example of this type of regulation is illustrated by the network containing the transcription factors NF-κB (Rel), ATF3, and C/EBPδ. (See Fig. 1).

3 Network Analysis of Innate Immune Responses

3.1 A Network that Enables Innate Immune Cells to Discriminate Between Transient and Persistent Activation

It is well established that transcriptional programs are propagated by sequential cascades of transcription factors (Bolouri and Davidson 2003; Smith et al. 2007). We have shown that stimulation of macrophages with LPS induced the transcription of multiple clusters of transcription factors within 3 h. We used a combination of mathematical modeling and biological experiments to predict and confirm the existence of a transcriptional network involved in TLR4 activation. The power of the approach lies in its ability to rapidly identify complex interactions between transcription factors and to define the functional emergent properties of the system, which in turn suggest the molecular underpinnings of the biological response. Analysis of the transcription factors activated immediately by LPS predicted the existence of many networks involved in the TLR4 response.

One of these networks contained the transcription factors NF-κB (Rel), ATF3, and C/EBPδ (Fig. 1). High-density temporal measurements of LPS-induced binding of these transcription factors to the Il6 promoter, combined with gene-deletion studies, enabled us to construct a model of a regulatory circuit that participates in the transcription of this cytokine-encoding gene. In this model, TLR4 stimulates translocation of NF-κB to the nucleus, where it activates weak transcription of Il6. Concomitant with that, NF-κB induces C/EBPδ, which then binds to the Il6 promoter and acts together with NF-κB to stimulate maximum transcription of Il6. At a later time point, ATF3 attenuates transcription of Cebpd and Il6. ATF3 recruits histone deacetylase 1 to the Il6 promoter in an LPS- dependent way. The ATF3-associated histone deacetylase 1 then deacetylates histones, resulting in the closure of chromatin and inhibition of Il6 transcription. It is known that C/EBPδ binds to and recruits the histone acetylase CBP to its target promoters, leading to more histone acetylation and chromatin opening. It is, therefore, likely that epigenetic chromatin remodeling contributes to this network.

The relationship between NF-κB and C/EBPδ suggests coherent feed-forward type I regulation (Alon 2007). This type of regulation has been suggested to protect biological systems from unwanted responses to fluctuating inputs (Alon 2007). The inflammatory response is like a double-edged sword, and it is therefore critical that inflammatory cells be modulate their response appropriately. The coherent feed-forward type I regulatory circuit described above could in principle enable immune cells to distinguish transient stimuli from more dangerous persistent activation. We used a combination of motif-scanning, microarray, and ChIP-on-chip analysis to identify many LPS-induced targets of C/EBPδ. These genes showed differences in transcriptional responsiveness to persistent and transient LPS-dependent stimulation of macrophages in vitro, and many have ascribed functions in host defenses against bacterial infection. Consistent with our in vitro studies, Cebpd-null mice were able to resist transient infection with a low

dose of *E. coli* H9049 but were highly susceptible to persistent infection with a higher dose.

In summary, we have used the tools of systems biology to show that TLR4-induced inflammatory responses are regulated by the integration of transcriptional 'on' and 'off' switches with 'amplifiers' and 'attenuators'. In addition, we have demonstrated a mechanism by which the macrophages are able to discriminate between transient and persistent activation. Collectively, these regulatory elements may facilitate the maintenance of effective host defense and the prevention of inflammatory disease.

3.2 Unraveling Complexity in Innate Immune Signaling

Genetic analysis of the mouse, whether through targeted gene deletions studies, chemical- or radiation-induced mutations, or spontaneous mutations has been one of the most powerful tools for unraveling immune responses. Although knockout mice are generally produced based on a hypothesized function of the targeted gene in a particular context, many genes that were originally identified by their role in other aspects of mouse biology have subsequently been shown to impact immune responses. Furthermore, large-scale phenotypic screening of mutagenized mice can uncover unpredicted components of immune regulatory pathways. In both of these cases, considerable effort is required to determine the mechanism by which these genes impact immunity.

Systems biology approaches organize information into sets of interacting networks that can serve to contextualize the role of a gene. A reference library of networks, such as those that we defined for NF-κB, ATF3, and C/EBPδ (Fig. 1), which are generated in a highly standardized manner, can be used as a comparator to identify signaling pathways that are functionally associated with mutated genes of interest. For example, the responses of macrophages carrying a mutation or gene deletion to a panel of immune stimuli can be compared with a compendium of responses from wild-type macrophages and macrophages lacking known components of TLR-induced signaling and gene regulatory networks. By identifying overlapping patterns in the responses, a testable hypothesis for the role of the gene in the immune response, and even its likely interaction partners, can be identified. These networks are generated using thousands of data points (e.g., entire transcriptomes) making it far less likely that such an overlap occurs by chance.

We have used this approach to link the *cpdm* mutation in SHARPIN to pathways known to regulate TLR responses (Zak et al. 2011). We identified SHARPIN as a potential regulator of macrophage responses in the course of a systems-level transcriptional and epigenomic analysis of combinatorial TLR pathway activation. To evaluate the role of SHARPIN in innate immunity, we measured TLR responses in macrophages derived from *cpdm* mice, which bear a null mutation in the *Sharpin* gene (Seymour et al. 2007). IL-12p40 production was markedly impaired in response to nearly all TLR ligands evaluated, including Pam_3CSK_4

(TLR2), LPS (TLR4), CpG-B (TLR9), and R848 (TLR7). The *cpdm* mutation also strongly attenuated macrophage production of IL-12p40 in response to infection with *Listeria monocytogenes*, which signals through TLR2, TLR5, and various Nod-like receptor family members (Zenewicz and Shen 2007; Warren et al. 2010; Leber et al. 2008).

Transcriptome analysis of wild-type macrophages identified 400 genes induced threefold or more by a stimulation with Pam_3CSK_4. SHARPIN deficiency arising from the *cpdm* mutation resulted in threefold impaired induction of 87 of these genes, including many pro-inflammatory cytokines. To identify the transcription factors that mediate the effect of SHARPIN on macrophage responses, we performed promoter analysis. We used PAINT (Vadigepalli et al. 2003) to scan the proximal promoter sequences of all 400 Pam_3CSK_4-regulated genes, and we then applied the gene set enrichment analysis (GSEA) algorithm (Subramanian et al. 2005) to determine which transcription factors were associated with impaired Pam_3CSK_4 responses. The only transcription factor binding sites that were overrepresented in the promoters of SHARPIN-dependent genes relative to the overall set of 400 Pam_3CSK_4-induced genes were NF-κB and AP-1. This result suggests that SHARPIN may be required for maximal NF-κB and AP-1 activation in response to TLR2 stimulation in macrophages.

We analyzed the link between SHARPIN, NF-κB, and AP-1 in greater depth by integrating the SHARPIN-dependent gene set defined above with our database of transcriptome responses in mutant macrophages. The set of 87 SHARPIN-dependent genes overlapped significantly with genes regulated by the *panr2*hypomorphic mutation in NEMO (Siggs 2010) a central node in the TLR2/NF-κB pathway. The extraordinarily strong association between the effects of these mutants suggested that SHARPIN might interact with NEMO. This interaction was confirmed by biochemical analysis and was abrogated by the *panr2* mutation.

In addition to pinpointing the location of SHARPIN in the TLR2/MyD88/NF-κB signaling cascade, this approach also revealed a previously unknown branch point in the pathway that controls a subset of the response. Although similar, the effects of SHARPIN-deficiency on macrophage responses were weaker than those of the *panr2* mutation. Some pro-inflammatory cytokine induction remain in SHARPIN-deficient macrophages that is not observed in *panr2* macrophages suggesting that the *panr2* mutation was also able to impair a SHARPIN-independent pathway. Recently, it was shown that a paralog of SHARPIN, RBCK1/HOIL-1L (Lim et al. 2001), interacts with NEMO as part of the LUBAC complex (Tokunaga et al. 2009), and therefore might mediate the SHARPIN-independent pathway. This hypothesis was reinforced by the observation that the *panr2* mutation ablates the RBCK1–NEMO interaction. Furthermore, it has recently been shown that SHARPIN and RBCK1 are present in distinct LUBAC complexes that are both capable of poly-ubiquitinating NEMO (Gerlach 2011; Ikeda 2011; Tokunaga 2011). Comparison of signaling defects induced by SHARPIN deficiency and by the *panr2* mutation suggested a model in which the MyD88 pathway bifurcates at NEMO. In this model, summarized in Fig. 2, maximal induction of many pro-inflammatory cytokines requires SHARPIN while the activation of a significant number of downstream

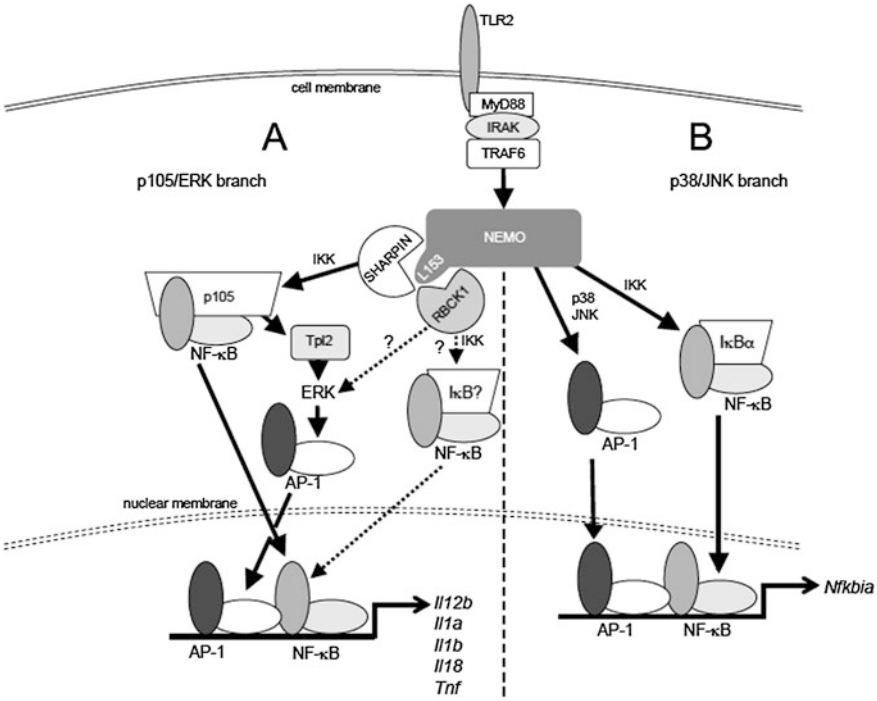


Fig. 2 SHARPIN is an essential adaptor distal to the branch point defined by the *panr2* mutation in NEMO. **a** The signaling responses most strongly impaired by SHARPIN deficiency and NEMO L153P (*panr2*) are the phosphorylation of p105 and ERK, suggesting that p105 IκB activity and TPL2 sequestration are dominant regulators of Toll-like receptor 2 (TLR2)-induced proinflammatory cytokine expression. The greater deficiency in signaling and pro-inflammatory cytokine induction observed in *panr2* compared with *cpdm* macrophages may result from SHARPIN-independent interactions between NEMO and the SHARPIN paralog and the linear ubiquitin chain assembly complex constituent RBCK1, which are also abrogated by NEMO L153P. **b** TLR2-induced IκBα degradation, phosphorylation of p38 and JNK, and Nfkbia gene induction were unimpaired in *cpdm* macrophages and *panr2* mutant macrophages, implying the existence ofa branch of NEMO-dependent I-kappa-B kinase (IKK) and MAPK activity that proceeds independently of SHARPIN and NEMO residue L153. (Adapted from Zak et al. 2011)

genes occurs independently. Combined with transcriptional network analysis, this also suggests previously unappreciated specificity in NF-κB and AP-1 activities since these molecules are effectors for both arms of the pathway.

4 Systems Vaccinology

To date, vaccines have been created by "trial and error"; vaccines are generated from related pathogens, attenuated pathogens, or pathogen components. Systems biology will enable rational vaccine design.

The response of an individual to vaccination depends on a multitude of interacting genetic, molecular, and environmental factors spanning numerous temporal and spatial scales. Systems biology provides a powerful toolset for deciphering complex biological networks and has been applied extensively to identify and contextualize novel regulators of the innate immune response (Amit et al. 2009; Zak 2011; Litvak et al. 2009; Ramsey et al. 2008; Gilchrist 2006; Suzuki 2009). The application of this approach to explore the innate-adaptive interface in the context of vaccination has already yielded new insights into the mechanisms of action of the 'gold standard' yellow fever vaccine YF-17D (Querec 2009; Gaucher 2008) and the seasonal influenza vaccine (Nakaya 2011). Furthermore, systems analysis of vaccination promises to generate useful biomarkers for protection and to identify mechanisms of immunogenicity that will guide rational vaccine design. As these topics have already been discussed in numerous reviews (Rappuoli and Aderem 2011; Pulendran et al. 2010; Zak and Aderem 2009; Shapira and Hacohen 2011; Gardy 2009), we will instead provide a high level perspective on systems vaccinology analysis that, by necessity, involves large number of model systems, each providing unique opportunities for discovery despite numerous practical constraints.

Systems vaccinology can be divided into five essential steps: measurements of the innate (1) and adaptive responses (2) to vaccination, determination of vaccine efficacy (3), systems-level data integration leading to the identification of biomarkers and mechanistic insights (4), and perturbation of the vaccine response in an appropriate experimental system (5). In the paragraphs below, we follow one cycle through the iterative systems vaccinology process, defining the inherent constraints and opportunities at each step.

4.1 An Iterative, Multistep Approach

Step I The starting point of the approach is the comprehensive analysis of the innate immune response to vaccination. A wide range of technologies is employed to make these measurements including transcriptomics, high-throughput serum analyte profiling, and proteomics. Transcriptome analysis is the most reliable and robust and is the predominant technique employed in systems vaccinology. In humans, vaccine-induced innate responses are most often measured indirectly by profiling readily accessible blood-derived cell populations (Querec 2009; Gaucher 2008; Nakaya 2011; Bosinger 2009; Palermo 2011).

Innate response measurements made by profiling whole blood or blood cell subsets, although indirect, are nevertheless highly informative. This analysis probes multiple aspects of the response, all of which occur in parallel. These include the subset of cells that respond directly to the vaccine, cells responding to inflammatory mediators induced by the vaccine, and changes in the composition and activation states in circulating cells.

Step II The next step in the approach is to measure vaccine-induced adaptive immune responses (immunogenicity). In contrast to innate responses, measurements of immunogenicity can be directly obtained from cells accessible in the blood or mucosa. These include antibody responses and antigen-specific $CD4^+$ and $CD8^+$ T cell responses (Hersperger 2011). Importantly, these measurements of adaptive immune function can be easily quantified at multiple time points to define the peak and memory responses as well as the impact of the initial vaccine 'primes' and subsequent 'boosts'.

Step III The third step of the approach is the measurement of vaccine efficacy. In some cases, such as malaria, efficacy can be measured directly in challenge studies. Alternatively for infections such as HIV where challenge studies are impossible, efficacy can be determined through measurements of vaccine-reduced acquisition rates, post-infection viral loads, transmission, or other aspects of the infection.

Step IV The full compendium of measurements are then computationally integrated in a systems-level analysis to derive mechanistic insights and biomarkers. When direct measurements of vaccine-induced innate immune responses are available, it is possible to make computationally guided predictions about the causal regulatory networks controlling the vaccine-induced responses. We, and others, have employed these approaches to derive and validate novel regulatory networks controlling Toll-like receptor activated networks in innate immune cells (Amit et al. 2009; Zak 2011; Litvak et al. 2009; Ramsey et al. 2008; Gilchrist 2006). These approaches can be readily extended to predict regulatory networks controlling responses to vaccines, which are likely to activate several innate immune pathways in parallel (Querec 2006; Lindsay 2010; Delaloye 2009). When vaccine-induced innate immune responses (direct or indirect) and immunogenicity or efficacy measurements are available from the same animal or volunteer, it becomes possible to computationally identify predictive signatures of protection (Fig. 3). Currently, the most powerful application of systems vaccinology is the identification of these immunogenicity and efficacy signatures. In the best case, robust predictive signatures illuminate novel mechanistic insights; in the worst case these signatures serve as valuable biomarkers (Querec 2009; Nakaya 2011; Brooks 2008; Zou 2005).

Step V While biomarkers achieve practical utility once they are validated in additional cohorts, the power of the mechanistic insights obtained in the first round of analysis is only realized when they are used to design and execute appropriate systems-level perturbations in an experimental model. The perturbations most directly related to molecular signatures are overexpression or knockdown (in vitro) or genetic ablation (murine in vivo) of the relevant genes. Although in vitro systems are the most easily perturbed, they are also the least appropriate for evaluating vaccine immunogenicity and efficacy. The murine genetic ablation validation strategy was recently

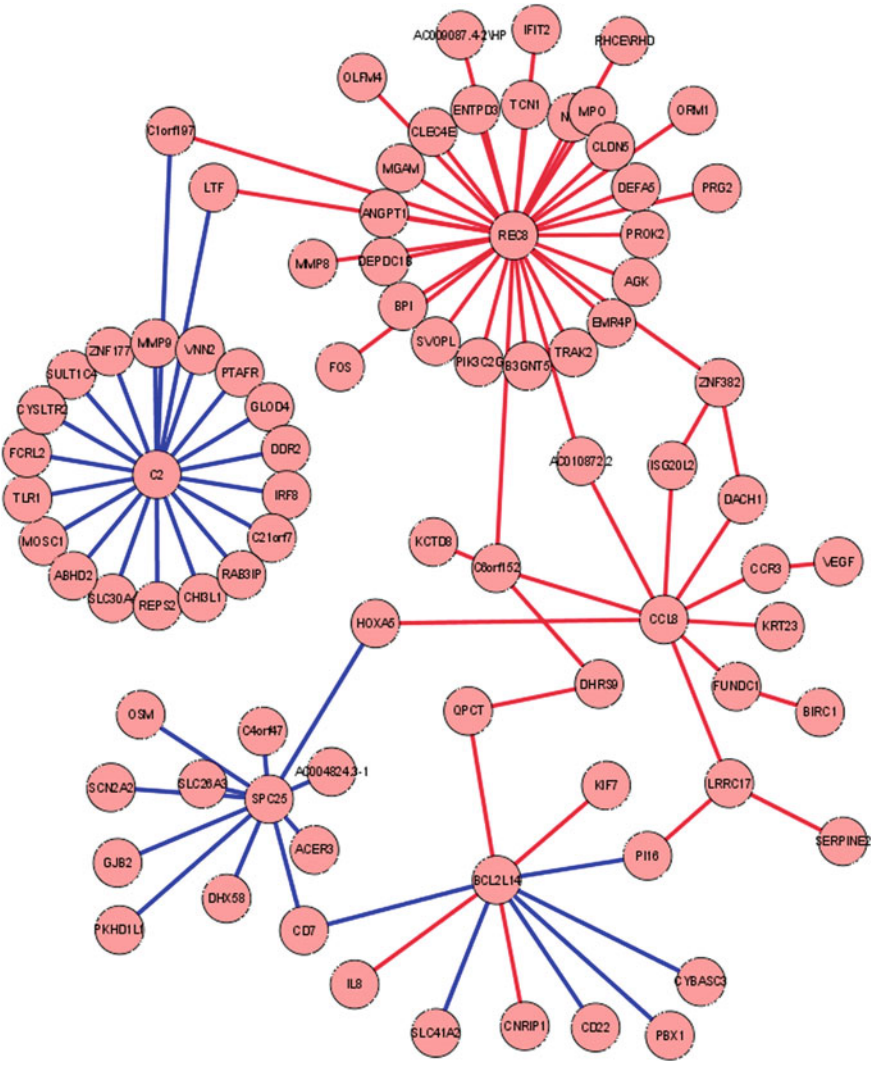


Fig. 3 Network of gene expression signatures associated with $CD4^+$ responses and SIV titers in macaques. The network represents innate immune signatures, measured days after primary vaccination, which predict enhanced SIV Gag-specific $CD4^+$ T cell responses, and reduced SIV load after challenge, measured several months later. Each circle represents a gene expressed in PBMCs of macaques 6 days after vaccination. Blue edges represent gene pairs associated with enhanced $CD4^+$ response; red edges represent gene pairs associated with immediate protection after challenge. (Adapted from Rappuoli and Aderem 2011)

applied in a study that identified immunogenicity signatures for the seasonal influenza vaccine in humans (Nakaya 2011). Small molecule agonists and inhibitors, specific for the networks implicated by the predictive signatures, can also be used as perturbations. Combinations of this class of drugs and vaccines have been explored in model systems (Araki 2009; Tan 2011) and may ultimately identify pharmacologic agents that can be combined with vaccines to improve efficacy.

4.2 Accelerating Efficacy Trials

During the 30 years since the discovery of HIV only four efficacy trials have been performed, an average of one trial every 8 years. Two of them have shown that anti-gp120 antibodies alone do not work; one has shown that T cells alone do not work; and one has shown that a prime-boost regime involving B and T cells may work. Altogether, only three hypotheses (elicitation of anti-gp120 antibodies, activation of T cells, and simultaneous elicitation of B and T cell responses) have been tested. Similarly, although challenge models are possible for malaria and numerous vaccines have been tested in phase I studies, all of these trials tested two hypotheses: peptide-based vaccines and RTS, S-based vaccines. To date, no efficacy trials have been performed for a new preventive vaccine against tuberculosis.

Accelerated clinical development can be achieved by improving the design of trials to test several hypotheses in parallel, incorporating systems biology to derive mechanistic insights and biomarkers, and employing a flexible strategy to expand the arms of the trial that are most promising (Freidlin and Simon 2005; Campbell 2009). For instance, several prime/boost strategies could be initiated concurrently in a large phase II study where subsets of the enrollees are monitored by systems biology approaches to test both safety and immune responses. This approach would identify vaccines that elicit qualitatively similar immune responses and permit rapid discrimination of different vaccine platforms and exploration of diverse approaches. Information collected during the early phases of the trial could be used to expand the most promising arms of the trial in order to gain sufficient statistical power to show the efficacy required for vaccine registration. Although this approach may require larger budgets during the initial phases, over the entire course of vaccine development, it will save money and time and will increase the probability of success. Several studies have demonstrated the ability to use early vaccine-response signatures to predict later immune responses (Querec 2009; Gaucher 2008; Nakaya 2011) and therefore, in principle, be used to make early decisions regarding the course of a clinical trial.

References

Rappuoli R, Aderem A (2011) A 2020 vision for vaccines against HIV, tuberculosis and malaria. Nature 473(7348):463–469

Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8(6): 450–461

Zak DE et al (2011) Systems analysis identifies an essential role for SHANK-associated RH domain-interacting protein (SHARPIN) in macrophage Toll-like receptor 2 (TLR2) responses. In: Proceedings of the National Academy of Sciences of the United States of America, 2011

Underhill DM, Ozinsky A (2002) Phagocytosis of microbes: complexity in action. Annu Rev Immunol 20:825–852

Aderem AA, Scott WA, Cohn ZA (1984) A selective defect in arachidonic acid release from macrophage membranes in high potassium media. J Cell Biol 99(4 Pt 1):1235–1241

Aderem AA (1985) Ligated complement receptors do not activate the arachidonic acid cascade in resident peritoneal macrophages. J Exp Med 161(3):617–622

Aderem AA (1986) Bacterial lipopolysaccharides prime macrophages for enhanced release of arachidonic acid metabolites. J Exp Med 164(1):165–179

Goodridge HS, Underhill DM (2008) Fungal Recognition by TLR2 and Dectin-1. Handb Exp Pharmacol 183:87–109

Blander JM, Medzhitov R (2004) Regulation of phagosome maturation by signals from toll-like receptors. Science 304(5673):1014–1018

Blander JM, Medzhitov R (2006) Toll-dependent selection of microbial antigens for presentation by dendritic cells. Nature 440(7085):808–812

Trinchieri G, Sher A (2007) Cooperation of Toll-like receptor signals in innate immune defence. Nat Rev Immunol 7(3):179–190

Miao EA et al (2006) Cytoplasmic flagellin activates caspase-1 and secretion of interleukin 1beta via Ipaf. Nat Immunol 7(6):569–575

Geddes BJ et al (2001) Human CARD12 is a novel CED4/Apaf-1 family member that induces apoptosis. Biochem Biophys Res Commun 284(1):77–82

Poyet JL et al (2001) Identification of Ipaf, a human caspase-1-activating protein related to Apaf-1. J Biol Chem 276(30):28309–28313

Masumoto J et al (2003) ASC is an activating adaptor for NF-kappa B and caspase-8-dependent apoptosis. Biochem Biophys Res Commun 303(1):69–73

Dinarello CA (1998) Interleukin-1 beta, interleukin-18, and the interleukin-1 beta converting enzyme. Ann N Y Acad Sci 856:1–11

Miao EA (2007) TLR5 and Ipaf: dual sensors of bacterial flagellin in the innate immune system. Semin Immunopathol 29(3):275–288

Stetson DB, Medzhitov R (2006) Type I interferons in host defense. Immunity 25(3):373–381

Kumar H et al (2008) Cutting edge: cooperation of IPS-1- and TRIF-dependent pathways in poly IC-enhanced antibody production and cytotoxic T cell responses. J Immunol 180(2):683–687

Edelmann KH (2004) Does Toll-like receptor 3 play a biological role in virus infections? Virology 322(2):231–238

Warren SE (2008) Multiple Nod-like receptors activate caspase 1 during Listeria monocytogenes infection. J Immunol 180(11):7558–7564

Roach JC (2005) The evolution of vertebrate Toll-like receptors. Proc Nat Acad Sci USA 102(27):9577–9582

Jin MS, Lee JO (2008) Structures of TLR-ligand complexes. Curr Opin Immunol 20(4):414–419

Jin MS (2007) Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. Cell 130(6):1071–1082

Liu L (2008) Structural basis of toll-like receptor 3 signaling with double-stranded RNA. Science 320(5874):379–381

Honda K, Takaoka A, Taniguchi T (2006) Type I interferon [corrected] gene induction by the interferon regulatory factor family of transcription factors. Immunity 25(3):349–360

Liew FY (2005) Negative regulation of toll-like receptor-mediated immune responses. Nat Rev Immunol 5(6):446–458
Boone DL (2004) The ubiquitin-modifying enzyme A20 is required for termination of Toll-like receptor responses. Nat Immunol 5(10):1052–1060
Wang YY (2004) A20 is a potent inhibitor of TLR3- and Sendai virus-induced activation of NF-kappaB and ISRE and IFN-beta promoter. FEBS Lett 576(1–2):86–90
Saitoh T (2005) A20 is a negative regulator of IFN regulatory factor 3 signaling. J Immunol 174(3):1507–1512
Lin R (2006) Negative regulation of the retinoic acid-inducible gene I-induced antiviral state by the ubiquitin-editing protein A20. J Biol Chem 281(4):2095–2103
Hitotsumatsu O (2008) The ubiquitin-editing enzyme A20 restricts nucleotide-binding oligomerization domain containing 2-triggered signals. Immunity 28(3):381–390
Bolouri H, Davidson EH (2003) Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. Proc Nat Acad Sci USA 100(16):9371–9376
Smith J, Theodoris C, Davidson EH (2007) A gene regulatory network subcircuit drives a dynamic pattern of gene expression. Science 318(5851):794–797
Seymour RE (2007) Spontaneous mutations in the mouse Sharpin gene result in multiorgan inflammation, immune system dysregulation and dermatitis. Genes Immun 8(5):416–421
Zenewicz LA, Shen H (2007) Innate and adaptive immune responses to Listeria monocytogenes: a short overview. Microbes Infection/Institut Pasteur 9(10):1208–1215
Warren SE (2010) Cutting edge: Cytosolic bacterial DNA activates the inflammasome via Aim2. J Immunol 185(2):818–821
Leber JH (2008) Distinct TLR- and NLR-mediated transcriptional responses to an intracellular pathogen. PLoS Pathog 4(1):e6
Vadigepalli R (2003) PAINT: a promoter analysis and interaction network generation tool for gene regulatory network identification. OMICS 7(3):235–252
Subramanian A (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Nat Acad Sci USA 102(43):15545–15550
Siggs OM (2010) A mutation of Ikbkg causes immune deficiency without impairing degradation of IkappaB alpha. Proc Nat Acad Sci USA 107(7):3046–3051
Lim S (2001) Sharpin, a novel postsynaptic density protein that directly interacts with the shank family of proteins. Mol Cell Neurosci 17(2):385–397
Tokunaga F (2009) Involvement of linear polyubiquitylation of NEMO in NF-kappaB activation. Nat Cell Biol 11(2):123–132
Gerlach B (2011) Linear ubiquitination prevents inflammation and regulates immune signalling. Nature 471(7340):591–596
Ikeda F (2011) SHARPIN forms a linear ubiquitin ligase complex regulating NF-kappaB activity and apoptosis. Nature 471(7340):637–641
Tokunaga F (2011) SHARPIN is a component of the NF-kappaB-activating linear ubiquitin chain assembly complex. Nature 471(7340):633–636
Amit I (2009) Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science 326(5950):257–263
Zak DE (2011) Systems analysis identifies an essential role for SHANK-associated RH domain-interacting protein (SHARPIN) in macrophage Toll-like receptor 2 (TLR2) responses. Proc Natl Acad Sci U S A 108(28):11536–11541
Litvak V (2009) Function of C/EBPdelta in a regulatory circuit that discriminates between transient and persistent TLR4-induced signals. Nat Immunol 10(4):437–443
Ramsey SA (2008) Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. PLoS Comput Biol 4(3):e1000021
Gilchrist M (2006) Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. Nature 441(7090):173–178
Suzuki H (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet 41(5):553–562

Querec TD (2009) Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol 10(1):116–125
Gaucher D (2008) Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. J Exp Med 205(13):3119–3131
Nakaya HI (2011) Systems biology of vaccination for seasonal influenza in humans. Nat Immunol 12(8):786–795
Pulendran B, Li S, Nakaya HI (2010) Systems vaccinology. Immunity 33(4):516–529
Zak DE, Aderem A (2009) Systems biology of innate immunity. Immunol Rev 227(1):264–282
Shapira SD, Hacohen N (2011) Systems biology approaches to dissect mammalian innate immunity. Curr Opin Immunol 23(1):71–77
Gardy JL (2009) Enabling a systems biology approach to immunology: focus on innate immunity. Trends Immunol 30(6):249–262
Bosinger SE (2009) Global genomic analysis reveals rapid control of a robust innate response in SIV-infected sooty mangabeys. J Clin Invest 119(12):3556–3572
Palermo RE (2011) Genomic analysis reveals pre- and postchallenge differences in a rhesus macaque AIDS vaccine trial: insights into mechanisms of vaccine efficacy. J Virol 85(2):1099–1116
Hersperger AR (2011) Qualitative features of the HIV-specific CD8 + T-cell response associated with immunologic control. Curr Opin HIV AIDS 6(3):169–173
Querec T (2006) Yellow fever vaccine YF-17D activates multiple dendritic cell subsets via TLR2, 7, 8, and 9 to stimulate polyvalent immunity. J Exp Med 203(2):413–424
Lindsay RW (2010) CD8 + T cell responses following replication-defective adenovirus serotype 5 immunization are dependent on CD11c + dendritic cells but show redundancy in their requirement of TLR and nucleotide-binding oligomerization domain-like receptor signaling. J Immunol 185(3):1513–1521
Delaloye J (2009) Innate immune sensing of modified vaccinia virus Ankara (MVA) is mediated by TLR2-TLR6, MDA-5 and the NALP3 inflammasome. PLoS Pathog 5(6):e1000480
Brooks JPL (2008) E.K., Analysis of the consistency of a mixed integer programming-based multi-category constrained discriminant model. Ann Oper Res 1–64:1–20
Zou HH, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc B. 67(Part 2):301–320
Araki K (2009) mTOR regulates memory CD8 T-cell differentiation. Nature 460(7251):108–112
Tan X (2011) Retinoic acid as a vaccine adjuvant enhances CD8 + T cell response and mucosal protection from viral challenge. J Virol 85(16):8316–8327
Freidlin B, Simon R (2005) Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res: An Off J Am Assoc Cancer Res 11(21):7872–7878
Campbell H (2009) Meningococcal C conjugate vaccine: the experience in England and Wales. Vaccine 27(Suppl 2):B20–B29

Studying Salmonellae and Yersiniae Host–Pathogen Interactions Using Integrated 'Omics and Modeling

Charles Ansong, Brooke L. Deatherage, Daniel Hyduke, Brian Schmidt, Jason E. McDermott, Marcus B. Jones, Sadhana Chauhan, Pep Charusanti, Young-Mo Kim, Ernesto S. Nakayasu, Jie Li, Afshan Kidwai, George Niemann, Roslyn N. Brown, Thomas O. Metz, Kathleen McAteer, Fred Heffron, Scott N. Peterson, Vladimir Motin, Bernhard O. Palsson, Richard D. Smith and Joshua N. Adkins

Abstract *Salmonella* and *Yersinia* are two distantly related genera containing species with wide host-range specificity and pathogenic capacity. The metabolic complexity of these organisms facilitates robust lifestyles both outside of and within animal hosts. Using a pathogen-centric systems biology approach, we are combining a multi-omics (transcriptomics, proteomics, metabolomics) strategy to define properties of these pathogens under a variety of conditions including those that

C. Ansong · B. L. Deatherage · Y.-M. Kim · E. S. Nakayasu · R. N. Brown · T. O. Metz · R. D. Smith · J. N. Adkins (✉)
Biological Separations and Mass Spectroscopy Group,
Pacific Northwest National Laboratory, PO Box 999,
MSIN: K8-98Richland, WA 99352, USA
e-mail: Joshua.Adkins@pnnl.gov

D. Hyduke · B. Schmidt · P. Charusanti · B. O. Palsson
Department of Bioengineering, University of California-San Diego,
La Jolla, CA, USA

J. E. McDermott
Computational Biology and Bioinformatics Group,
Pacific Northwest National Laboratory, Richland, WA, USA

M. B. Jones · S. N. Peterson
J. Craig Venter Institute, Rockville, MD, USA

J. Li · A. Kidwai · G. Niemann · F. Heffron
Department of Molecular Microbiology and Immunology,
Oregon Health and Sciences University, Portland, OR, USA

S. Chauhan · V. Motin
Departments of Pathology and Microbiology and Immunology,
University of Texas Medical Branch, Galveston, TX, USA

K. McAteer
Biology Program, Washington State University Tri-Cities,
Richland, WA, USA

Current Topics in Microbiology and Immunology (2013) 363: 21–41
DOI: 10.1007/82_2012_247

Published Online: 9 August 2012

mimic the environments encountered during pathogenesis. These high-dimensional omics datasets are being integrated in selected ways to improve genome annotations, discover novel virulence-related factors, and model growth under infectious states. We will review the evolving technological approaches toward understanding complex microbial life through multi-omic measurements and integration, while highlighting some of our most recent successes in this area.

Contents

1 Introduction

The outcome of an intracellular bacterial infection—bacterial replication and host cell death versus host cell containment of the pathogen—is a complex process that involves multiple interactions between the host cell and the attacking bacteria. The mechanisms employed by each participant in these interactions are multifaceted and often difficult to elucidate by traditional methods. This is where a "systems approach" in which detailed data covering the bacterial and/or host transcriptome, proteome, and metabolome are integrated using metabolic and regulatory models is useful.

In this chapter, we describe the systems biology approach we apply to analyze, identify, quantify, model, and predict the overall molecular processes involved in

the pathogenesis by *Salmonella* and *Yersinia* species, two relatively closely related and medically important pathogens within the family Enterobacteria. In humans, the pathogenic *Salmonella* serovars *Salmonella Typhimurium* and *Salmonella Typhi* cause a self-limiting gastroenteritis and frequently fatal typhoid fever, respectively. *Salmonella* infection is a major public health problem, causing up to 3 million cases of infection per year in the US alone (Coburn et al. 2007), and the recent emergence of untreatable, multidrug resistant strains such as phage type DT104 has further increased the threat to public health (Glynn et al. 1998). Also pathogenic to humans, *Y. pseudotuberculosis* and *Y. enterocolitica* induce gastroenteritis in the human host, and *Y. pestis* is the causative agent of plague, an acute and lethal disease responsible for at least three pandemics that killed an estimated 200 million people (Perry and Fetherston 1997).

In addition to each other, *Yersinia* and *Salmonella* are closely related to *E. coli,* one of the best studied model systems for biological research (Brenner et al. 1969; Brenner and Falkow 1971; Brenner 1978; Sharp 1991; Lerat et al. 2003). This phylogenetic relationship is advantageous in that: (1) well-characterized biochemical pathways, (2) protein–protein interaction databases, (3) well-characterized transcriptional regulatory and start sites, and (4) molecular biology tools established for *E. coli* provide baseline information that can be applied to studies of *Salmonella* and *Yersinia*. Our premise is that knowledge gained from coordinated analysis and modeling of these two genera will lead to improved control and therapeutic treatment strategies, not only for these specific pathogens, but also for related gram-negative bacteria in general.

Before delving into the application of our systems-level approach to gain insights into pathogenicity from the perspectives of both these bacteria and the host, as well as into host–pathogen interactions, we introduce the key elements underlying our systems biology approach.

2 Elements of a Systems Approach

Our systems-level approach utilizes iterative and complementary experimental and computational methodologies to obtain sample matched, high-dimensional transcriptome, proteome, and metabolome/lipidome data for developing predictive models of pathogenicity for *Salmonella* and *Yersinia* species.

2.1 Experimental Considerations for a Biological Perspective

Experimental considerations for a systems-based analysis in which the number of components simultaneously quantified is important must balance the desire for in depth measurements and broad analyte coverage with conditions representative of the biological environment. In the majority of scenarios, in vitro culture conditions

that simulate environmental conditions encountered by the pathogen during infection represent a good experimental approach, as they are capable of generating large quantities of samples to simultaneously quantify thousands of components (i.e., transcripts, proteins, and metabolites) for a systems analysis of relevant biological interactions (Coombes et al. 2005; Ansong et al. 2008b; 2009; White et al. 2010; Yoon et al. 2011). For example, when *Salmonella* is grown in an acidic minimal media (low pH, low magnesium, and nutrient-deficient) to partially mimic the host intracellular milieu, expression of many genes required for systemic infection are appropriately regulated (Deiwick et al. 1999; Miao and Miller 2000). Similarly, growth of *Yersinia* at 37 °C in calcium-deficient chemically defined best case scenario (BCS) medium induces the type 3 secretion system (T3SS) required for *Yersinia* virulence (Brubaker 1991; Straley et al. 1993; Perry and Fetherston 1997; Cornelis 1998, 2002). Thus, in vitro growth in media with specific composition can be used as a surrogate of the host environment during infection.

A second experimental approach is based on infection of cultured macrophages (specifically RAW264.7 murine macrophage cell line) in vitro. *Salmonella* remains within professional phagocytic cells during mouse infection, and previous studies have shown that replication in macrophages is directly correlated to the ability to cause systemic infection (Fields et al. 1986). *Yersinia pestis* also displays an intracellular growth phase in macrophages, and the ability of *Yersinia* strains to infect and replicate in macrophages has been correlated with virulence (Cavanaugh and Randall 1959; Straley and Harmon 1984a, b; Fukuto et al. 2010).

The third experimental approach involves whole animal models such as mice. Mouse models of infection for *Salmonella* and *Yersinia* are widely considered to be viable surrogates of pathogenicity in humans. For *Salmonella*, the two most commonly used mouse models are C57BL/6 and Balb/c. Both of these strains are susceptible to *S. Typhimurium* infection and die following either intragastric (i.g.) or intraperitoneal (i.p.) infection with the strain used in our work (14028 s; $LD_{50} \sim 10^5$ i.g., $LD_{50} < 10^1$ i.p.). For *Yersinia,* commonly used susceptible mouse models include Swiss–Webster mice and Balb/c mice, in which intranasal/aerosol challenge and subcutaneous challenges represent pneumonic and bubonic modes of plague infection, respectively. The LD_{50} doses for the subcutaneous (s.c.) and aerosol routes are $<10^1$ and 2×10^4 colony forming units respectively (Welkos et al. 1995, 1997; Worsham et al. 1995). We note that there are many other strains of mice that contain mutations in specific anti-microbiocidal components normally expressed by professional phagocytic cells that represent additional important resources in analyzing host–pathogen interactions (Vidal et al. 2008).

2.2 Foundational Omics Technologies

Omics technologies have transformed molecular biology into a data-rich discipline by enabling scientists to simultaneously measure multiple molecular components (e.g., proteins, metabolites, and nucleic acids) that operate in a network of interactions to

generate cellular functions and phenotypic states (Joyce et al. 2006; Oldiges et al. 2007; Cascante and Marin 2008; Ly et al. 2010; Zhang et al. 2010).

In the context of systems biology, transcriptomics is a critical enabling analytical method due to the high precision and relative ease of data generation. DNA microarray transcriptome analysis platforms are now a common laboratory commodity due to the availability of high quality reagents (e.g., slides, cyanine dyes, etc.), widespread adoption of short oligonucleotide probes (70-mers) and exponential reduction in costs of oligonucleotide synthesis and commercial DNA microarray instrumentation. A limitation of microarrays is that they restrict expression profiling data to specific predicted gene annotations. Overcoming this limitation are the next generation sequencing (NGS) transcriptome analysis platforms that allow biologists to determine the primary sequence and relative abundance of every expressed transcript in a cell (whole transcriptome profiling) at an unprecedented level of sensitivity and accuracy (Wang et al. 2009; Martin and Wang 2011; Ozsolak and Milos 2011). However, even this level of information is insufficient for determining whether the transcript is translated into a protein, the macromolecules that execute biochemical functions in all cellular systems.

Comprehensive knowledge regarding protein abundances in organisms, host cells, and tissues is considered essential to the study of infectious diseases and cellular response to stresses. This information provides a basis for understanding genetic variants, gene functions, and action mechanisms, which are needed to develop the means to diagnose, treat, and protect against infectious disease organisms. While some information about relative protein expression levels may be inferred from high-throughput analysis of the mRNA complement or transcriptome (Adams 1996; Velculescu et al. 1997), measured mRNA levels do not necessarily correlate with either the corresponding activity or abundances of proteins (Anderson and Seilhamer 1997; Haynes et al. 1998; Gygi et al. 2000; Schwanhausser et al. 2011). For example, $\sim$20 % at a minimum and potentially as much as 50 % of the *S. Typhimurium* genome is post-transcriptionally regulated (Sittka et al. 2008; Ansong et al. 2009). Protein functions may also be modulated by post-translational modifications (e.g., phosphorylation, acetylation, etc.) and/or by forming complexes with other biomolecules (e.g., proteins, RNA, lipids, etc.) or small molecules (metabolites, dissolved gases, etc.). This information is not even peripherally available from transcriptome analysis. As such, proteomics—the study of the entire complement of proteins expressed by a cell under a specific set of conditions at a particular time—is another key enabling technology in the emerging science of systems biology.

As proper metabolic function underlies nearly every aspect of pathogenesis, e.g., nutrient acquisition and survival within specialized compartments inside host macrophages, metabolomics plays an important role in developing systems-level understanding. Broadly defined, metabolomics is the quantitative determination of time-related or stimuli-dependent changes in the small molecular weight complement of an integrated biological system, cell, or cell types (Nicholson et al. 1999; Kueger et al. 2012). Metabolomics can be further subdivided based on biochemical class, specifically metabolomic studies selective for lipids is termed

"lipidomics". The metabolome and lipidome are the molecules meant to be directed by the transcriptome and in turn the proteome, with small molecules playing critical roles in energy balance, intercellular communication, membrane dynamics, osmoregulation, and many other life processes.

2.3 Computational Framework for Integrating Biological Information

Extracting 'knowledge' from the volumes of omics data resulting from high-throughput measurements is nontrivial (Palsson and Zengler 2010). Two major network approaches have emerged to extract biological insight from this omics ocean: one is inference based and the other, knowledge based. Both approaches use an interconnected network of biological molecules to interpret omics data; however, there are crucial differences in how the networks are constructed and in the biological questions that can be studied. Inference-based approaches employ statistical methodologies to construct network models from correlation or recurring patterns in omics data (see Refs. (Margolin and Califano 2007; Bonneau 2008; De Smet and Marchal 2010) for reviews). Knowledge-based, which is also referred to as reconstruction based, approaches are essentially two-dimensional genome annotation efforts (Palsson 2004) that construct networks from biochemical and genetic data (reviewed in (Reed et al. 2006; Feist et al. 2009; Hyduke and Palsson 2010; Thiele and Palsson 2010)). Statistical inference methods benefit from incorporation of all data in an omics set to guide hypothesis development related to unknown interactions. However, these methods are complicated by the fact that the component measurements are not independent and that they do not account for biochemical and genetic causality (Margolin and Califano 2007). A major shortcoming of inference-based methods is that they typically solve underdetermined problems, thus they are not guaranteed to provide a unique solution (De Smet and Marchal 2010). Network reconstruction employs established biochemical, genetic, and genomic data (Reed et al. 2006; Feist and Palsson 2008; Oberhardt et al. 2009; Schellenberger et al. 2010) to assemble a knowledge base of an organism's molecular components and interactions (Thiele and Palsson 2010). Because knowledge bases are constructed from biological information, whereas inference methods are based on statistical correlations or information theory, knowledge bases provide a biological context for omics analysis (Lewis et al. 2009). The major shortcoming of the knowledge base approach is that they do not, currently, account for the activities of all genes in a genome, thereby limiting the ability to discover novel relationships important to pathogenesis.

Our systems-level strategy utilizes both inference- and knowledge-based approaches to investigate the molecular mechanisms underlying virulence as both approaches have their own unique strengths that allow us to probe the regulatory influences and biochemical mechanisms associated with virulence.

3 Pathogen Perspective: *Salmonella*

Overview

Our overarching biological approach focuses on elucidating virulence mechanisms necessary for *Salmonella* to cause systemic infection. The approach employs in-silico network reconstructions that integrate omics data into a single coherent, systems-level framework. In this section, we describe key steps in this process to improve annotation of the *Salmonella Typhimurium* genome, develop a *Salmonella Typhimurium* genome-scale metabolic reconstruction, and apply the omics-data constrained *Salmonella* metabolic model for in-silico biology applications.

3.1 Proteogeonomics

Complete and accurate genome annotation is crucial as incorrectly annotated genes and/or unannotated genes confound interpretation of experimental omics analyses and result in non- or dysfunctional computational models. However, determining protein-coding genes for most new genomes is almost completely performed by inference using computational predictions that experience significant error rates (Ansong et al. 2008a; Armengaud 2009; Payne et al. 2010). Compounding this issue is a lack of experimental evidence to support predicted protein-coding regions for the overwhelming majority of annotated genomes. Even when available, experimental evidence is typically based on expressed RNA sequences, such as from microarray or NGS experiments, which do not independently and unequivocally elucidate whether a predicted protein-coding gene is translated into a protein, or provide any reliable information on post-translational processing.

Bottom-up proteomics offers the ability to directly measure peptides arising from expressed proteins representing the current best option for independently and unambiguously identifying at least an important subset of the protein-coding genes in a genome and can be used to experimentally validate and correct in-silico gene annotations (Jaffe et al. 2004; Ansong et al. 2008a; de Groot et al. 2009; Wright et al. 2009). Toward this end, we complemented the current *Salmonella Typhimurium* 14028 in-silico annotation with bottom-up proteomics data (Ansong et al. 2011). The data provide protein-level experimental validation for approximately half of the predicted protein-coding genes in *Salmonella* and suggest revisions to several genes that appear to have incorrectly assigned translational start sites.

The proteomics data also revealed 12 non-annotated genes missed by gene prediction programs and provided evidence that suggested a role for one of these genes in *Salmonella* pathogenesis. Moreover, the data-enabled characterization of post-translational features in the *Salmonella* genome that included chemical modifications and proteolytic cleavages. This information revealed a much larger and more complex repertoire of chemical modifications in bacteria than previously thought and included several novel modifications and more than 130 signal peptide

and N-terminal methionine cleavage events critical for protein function. The refined genome annotation facilitates omics analyses and is useful for developing more complete models of metabolism and regulation.

3.2 Genome-Scale Metabolic Reconstruction

Metabolism arguably has the most complete network in *Salmonella,* relative to for example gene regulatory or protein–protein interaction networks, and its proper function underlies nearly every aspect of pathogenesis, e.g., nutrient acquisition and survival within *Salmonella*-containing vacuoles of macrophages. Thus, understanding metabolism under a variety of growth conditions provides us with key insights and testable hypotheses regarding the molecular mechanisms *Salmonella* employs during host infection. Toward this end, we reconstructed and examined the metabolic network of *S. Typhimurium* (Thiele et al. 2011). A metabolic network reconstruction contains all of the possible metabolic reactions known to occur within an organism, although only a subset of these reactions is likely to be active at any time. The *S. Typhimurium* metabolic reconstruction contains 1270 genes, 1119 biochemically unique intracellular metabolites, and 2201 network reactions. Also considered in this model is the importance of localization and movement of metabolites including distinct compartments for the cytoplasm, periplasm, and inner and outer membranes. The metabolic reconstruction by itself is a useful platform for biological discovery as discussed immediately below; however, integrating global omics measurements relevant to infection constrain the model to growth representative of infection allowing detailed studies on phenotypic behavior and analysis of network properties relevant to pathogenenesis as described in the following sections.

In an initial application, we employed the reconstruction to make a number new of predictions regarding possible therapeutic targets in a synthetic gene deletion analysis (Thiele et al. 2011). A number of 56 synthetic lethal gene pairs were found to disrupt growth of *S. Typhimurium* in silico. Notably, several gene pairs are known to be essential for virulence, but not for growth, and have known inhibitors based on the enzyme database BRENDA, further underscoring the applicability of the network reconstruction for applications such as identification of candidate drug targets.

3.3 Metabolic Model-Guided Analysis of Omics Data

Genome-scale metabolic reconstructions are attractive frameworks for multiomic analysis because they represent metabolism in chemically accurate terms and relate enzyme activities to the genome. To gain insight into the changes in the functional state of *Salmonella*'s metabolic network during infection, we grew *S. Typhimurium*

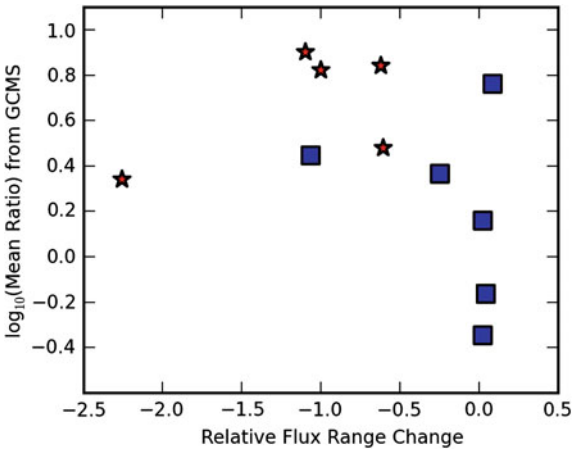


Fig. 1 *Salmonella* preferentially maintains metabolic pathways putatively associated with immunosuppression in minimal media. Omics-data tailored condition-specific models of *Salmonella* metabolism were analyzed with flux variability analysis. The *y*-axis shows the ratio for each metabolite concentration as detected by GC–MS. The *x*-axis depicts the relative change in the allowable ranges of flux through each metabolite as characterized by flux variability analysis. Metabolites capable of suppressing macrophage activation are shown as *blue* boxes. Metabolites capable of supporting macrophage activation are shown as *red* stars

14028 s in rich media and in acidic minimal media defined to mimic the intramacrophage environment after which the *Salmonella* metabolic model was used to analyze sample-matched transcriptomics, proteomics, and metabolomics data generated from these samples. While the transcript and protein data was used to inform reaction flux constraints under the conditions tested, the metabolite data informed on the turnover of intracellular metabolites. This allowed further refinement of the model by requiring that *S. Typhimurium* utilize the detected metabolites in the allowed network states. Small metabolites may play an important role in immunological processes, and we observed a number of metabolites that resulted in modulation of macrophage activation when used as substrates for cellular metabolism. Analysis of sample-matched omics data using the *Salmonella* metabolic model revealed *Salmonella* maintained the metabolic potential for high fluxes of intracellular metabolites postulated to inhibit macrophage activation, presumably allowing for adaptation to the host environment (Fig. 1).

3.4 Inference-Based Analysis of Omics Data

Our reconstruction-based (also known as knowledge-based) network approach is complemented by using an inference-based network approach that employs statistical methodologies to construct network models from correlation or recurring

patterns in omics data. This complementary approach is important as it enables hypothesis development related to unknown interactions.

As a demonstration of the utility of the inference-based modeling approach, the context likelihood of relatedness (CLR) algorithm (Faith et al. 2007), which uses mutual information to infer relationships between genes based on the coordination of their expression profiles across different conditions, was employed to predict proteins important to *Salmonella* pathogenesis (i.e., virulence factors) from sample-matched transcriptomics and proteomics data of *Salmonella* and knockout mutants of 14 regulators required for virulence (Yoon et al. 2011). This approach uncovered many of the known major virulence factors in *Salmonella* recapitulating aspects of known *Salmonella* biology that had taken decades of traditional research to arrive at as well as uncovering several novel network-predicted virulence factors a subset of which importantly were experimentally verified demonstrating the utility of the approach (Yoon et al. 2011).

3.5 Integrated Inference- and Knowledge-Based Analysis of Omics Data

As the metabolic knowledgebase is limited to only those genes associated with metabolism (1271 in the *Salmonella* metabolic reconstruction) it fails to exploit potential clues to virulence programs present in the remaining ~3000 *Salmonella* genes. To overcome this limitation and increase the knowledge extracted from proteome and transcriptome data, we developed an integrated approach that uses the CLR statistical inference method in combination with the *Salmonella* metabolic model (STM_v1.0). In this approach, CLR is utilized to infer a set of candidate 'bottleneck' genes, after which STM_v1.0 is deployed to assess the phenotypic relevance of these genes to growth. A bottleneck gene is frequently (relative to the other genes) found in the shortest path between two genes in the network and they are thought to represent important mediators of system processes (McDermott et al. 2009). The benefit of using CLR inferences with the defined metabolic network is that although CLR does not necessarily infer an actual biological network, it provides information about the influences of all genes measured.

Application of the CLR algorithm to transcriptome data identified potential bottlenecks that were analyzed in the context of the metabolic model to identify the growth conditions in which deletion of a bottleneck would reduce or abrogate growth. We performed in-silico growth simulations using flux balance analysis (FBA) (Feist and Palsson 2010; Orth et al. 2010) or the minimization of metabolic adjustments (MoMA) method (Segre et al. 2002) to assess the impact of gene deletion on growth. Comparison with experimental observations testing the predicted phenotypic effects of the metabolic model and the relevance of the select set of bottleneck genes to virulence showed the FBA method to be less accurate than

Strain	*In Silico* Growth Rate (Relative to WT) FBA	MOMA	*In Vitro* Phenotype
14028s	100%	100%	Growth
ΔatpA	76%	57%	Weak Growth
ΔtpiA	98%	66%	Weak Growth
ΔpurK	100%	100%	Growth
ΔmetN	99%	98%	Growth
ΔmetA	100%	100%	Growth
ΔfrdA	100%	100%	Growth
Δeno	74%	13%	Weak Growth
ΔcyoA	90%	72%	Growth
ΔgpsA	0%	0%	Weak Growth
ΔgapA	66%	0%	No Growth
Δpgk	66%	0%	No Growth
ΔatpA/Δpgk	35%	0%	No Growth
ΔatpA/ΔgapA	35%	0%	No Growth
ΔatpA/ΔtpiA	55%	0%	No Growth

Fig. 2 Experimental phenotypes are consistent with simulated phenotypes for genes identified to be important by coordinated inference and genome-scale metabolic analysis. FBA (quicker to compute) and MoMA (slower to compute) are different approaches for utilizing the metabolic reconstructions to predict phenotypes resulting from knocking out metabolic functions. The growth results for 14 gene deletion mutants relative to "wild type" (WT) parent 14028 s are shown. The *red* boxes indicate poor agreement between growth predictions and the experimentally observed growth result, whereas the *green* boxes indicate good agreement

the MoMA method which showed good agreement with experimental observations (Fig. 2). This finding was not surprising because FBA predicts what the metabolic network could achieve after the organism has evolved to cope with the genetic manipulations while MoMA was developed to identify the growth rate achievable immediately following a perturbation. These results demonstrate the power of leveraging the unique strengths of two different network approaches to increase the amount of knowledge extracted from omics data.

4 Pathogen Perspective: *Yersinia*

Overview

The studies and methodologies focused on Salmonellae provided a foundation from which to study the less understood pathogenic organisms of the *Yersinia* genus. In addition to elucidating virulence mechanisms necessary for *Yersinia* to cause systemic infection, an overarching goal is to understand the differences in disease manifestation among closely related species. In this section, we describe application of the system biology approach described above to gain insight into *Yersinia* biology.

4.1 Proteogeonomics

The concept of annotation refinement introduced above can be extended to include a comparative assessment of genomes across closely related species and the use of multiple omics-data sources further enhancing the value of annotation improvements. Transcriptomic and proteomic data derived from highly similar pathogenic *Yersiniae* (*Y. pestis* CO92, *Y. pestis* Pestoides F, and *Y. pseudotuberculosis* PB1/+) was used to complement the current in-silico annotation for each strain. Peptide and oligo measurements experimentally validated the expression of nearly 40 % of each strain's predicted proteome and revealed 28 novel and 68 previously incorrectly annotated protein-coding sequences (e.g., observed frameshifts, extended start sites, and translated pseudogenes) within the three current *Yersinia* genome annotations (Schrimpe-Rutledge et al. 2012). The refined genome annotations are immediately useful to facilitate omics analyses and develop more complete models of metabolism and regulation.

4.2 Genome-Scale Metabolic Reconstruction

As a framework for integrating and analyzing omics data toward a systems approach to understanding *Yersinia* pathogenesis, we completed a metabolic reconstruction for *Y. pestis* CO92, a strain that is virulent to humans (Charusanti et al. 2011). The metabolic network of *Y. pestis* possesses sufficient flexibility as to endow the organism with the ability to survive and proliferate in its two hosts: (1) the flea insect vector (growth at 26-28 °C) and (2) mammalian vectors such as rodents and humans (growth at 37 °C). The reconstruction contains 815 genes, 678 proteins, 936 unique metabolites, and 1678 reactions, considers localization as for *Salmonella* (see above), and includes two biomass objective functions that account for differences in cellular biomass composition when *Y. pestis* is grown at the two different temperatures.

We employed the reconstruction to analyze gaps in various *Y. pestis* CO92 metabolic pathways. The reconstruction identified two critical gaps in the lysine and fatty acid biosynthesis pathways that needed to be filled in order for model simulations to occur. This necessity prompted a search for alternative genes in *Y. pestis* CO92 that could catalyze the same reactions as those catalyzed by the missing genes. A search for paralogs of YPO0170, the missing gene in lysine biosynthesis, uncovered YPO1962, a potential open reading frame with 59 % nucleotide identity that might have the same catalytic ability; however, there were no apparent paralogs for *fabI*, the missing gene in fatty acid biosynthesis, within the *Y. pestis* CO92 genome.

We searched for alternative enzymes by overlaying global transcript and protein expression data onto the reconstructed metabolic pathways in *Yersinia* as illustrated for the transcript data in Fig. 3. Our reasoning was that any enzyme

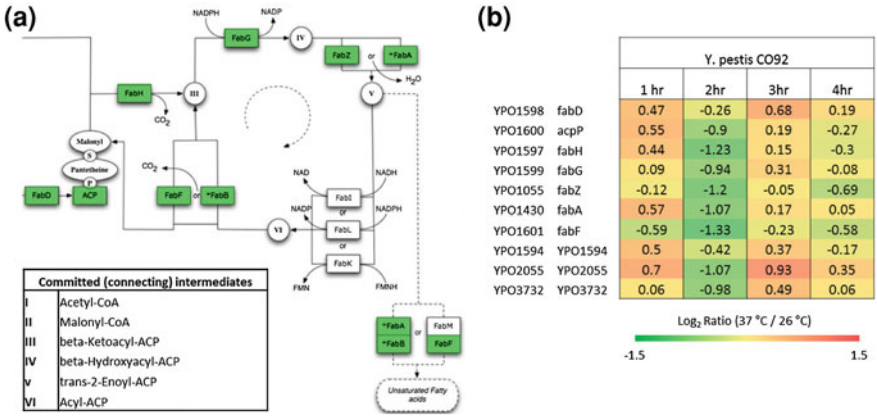


Fig. 3 Visual representation of integrated omics data and reconstruction. Temporal expression pattern of identified fab genes (panel B) in the fab pathway (panel A) are shown including putative fabI candidates YPO1954, YPO3732, and YPO2055. Each column represents ratio of 37 °C/26 °C across time (1 h, 2 h, 4 h, 8 h). The color scale ranges from *green* (total low relative abundance) to *red* (high relative abundance)

having the same catalytic function as FabI should be located near the fatty acid biosynthetic cluster (YPO1595 to YPO1601), exhibit correlated expression with genes in this cluster, and be annotated as hypothetical. The best match based on these criteria turned out to be the hypothetical gene YPO1594. Other genes that showed correlated expression, but were located farther away from the biosynthetic cluster, were YPO3732 and YPO2055.

5 Host Perspective

During infection, pathogens attempt to hijack resources from the host, while host cells attempt to limit the materials available for pathogen reproduction and virulence. The importance of metabolism in host–pathogen interactions is exemplified by the battle over free iron. While the host attempts to limit the iron available to pathogens, pathogens have evolved high-efficiency chelators to scavenge available iron from the active sites of a variety of metabolic enzymes, such as those used in amino acid biosynthesis. However, the extent to which we can accurately measure the complete molecular makeup of a pathogen during infection is limited. Therefore, a systems-level model of host–pathogen interactions possesses the potential to identify a key subset of molecular features that should be measured to unravel the pathogen's molecular decision-making processes during infection.

To better understand the metabolic features of the host during *Salmonella* infection, we completed a genome-scale metabolic reconstruction for the murine RAW 264.7 macrophage cell line (Bordbar et al. Mol Sys Biol 2012—in press). This reconstruction contains 820 genes, 574 unique metabolites, and 1067 reactions. Physiological metabolic rates of the reconciled metabolic network were evaluated for biomass growth, ATP production, and NO synthesis and compared to experimental values. Overall, our results indicate that the reconciled metabolic network is predictive of physiologically relevant experimental rates when in vitro experimental uptake rates are imposed.

In an initial application, the macrophage metabolic model was used to analyze transcriptomics and proteomics data from the time course responses of RAW 264.7 macrophages to lipopolysaccharide (LPS) stimulation. Host cell response(s) to *Salmonella* infection and to LPS treatment are similar in that they both result in expression of multiple antimicrobial factors. This analysis resulted in the identification of metabolites and enzymes associated with immunomodulation. We have also shown this using inference-based modeling of macrophages, which revealed a common response to multiple immune challenges (McDermott et al. 2011a). To determine if nutrient availability could affect macrophage activation, we performed sensitivity analysis for a set of activation phenotypes as a function of in silico medium composition. Our analysis identified a number of nutrients with the potential to modulate macrophage activation such as glutamine, urea, and threonine. This study demonstrates that the role of metabolic processes in regulating host cell activation may be greater than previously anticipated and elucidates underlying metabolic connections between activation and metabolic effectors.

6 Host–Pathogen Interaction

6.1 Integrated Host–Pathogen Model of Metabolism

Computational genome-scale metabolic models of individual pathogens or their respective hosts are undoubtedly useful for integrating omics and physiologic data for systemic, mechanistic analysis of metabolism. However, the next step toward understanding the interactions between a pathogen and its host requires integrated modeling of both host and pathogen metabolic networks. To this end, we pioneered an approach for integrative analysis of host–pathogen interactions that employs in-silico mass-balanced, genome-scale models and tested it using the closely related *Mycobacterium tuberculosis* (M. tb)-human alveolar macrophage interaction as a model system, as resources related to this system were more mature (Bordbar et al. 2010). Briefly, we constructed a cell-specific alveolar macrophage model iAB-AMØ-1410 from the global human metabolic reconstruction, Recon 1 (Duarte et al. 2007). This model was then integrated with an *M. tuberculosis* H37Rv model, iNJ661, to build an integrated host–pathogen genome-scale reconstruction, iAB-AMØ-1410-Mt-661. Importantly,

this integrated host–pathogen network enables simulation of the metabolic changes during infection.

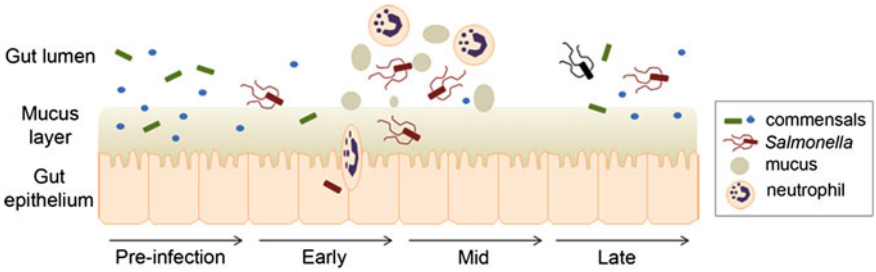
Deployment of the host–pathogen metabolic model to analyze high-throughput data from infected macrophages representing three distinct *M. tuberculosis* infectious states (latent, pulmonary, and meningeal) highlighted differences in metabolism among the three different states (Bordbar et al. 2010). This pioneering effort demonstrates integrated host–pathogen reconstructions can form a foundation upon which understanding the biology and pathophysiology of a variety of infections can be developed. Further, the foundational efforts described above have now been performed to enable this approach with *Salmonella* and *Yersinia* with a mouse macrophage cell line.

6.2 The Host–Pathogen Interface

The interplay between effector proteins secreted by the pathogen and host cells exposed to these effector proteins are relevant to infection in many Enteropathogens and as such can be useful in modeling host–pathogen interactions. An added benefit is that some of these virulence factors may be potential new drug targets.

We applied our systems approach to characterize the *Salmonella* secretome, using omics technologies, inference-based computation and biological experimentation. In this case, we experimentally identified secreted virulence factors by analyzing the extracellular medium from wild type *Salmonella*, a mutant that promotes secretion (ΔSsaL), and a mutant that inhibits secretion (ΔSsaK) (Niemann et al. 2011). Proteomics analysis of the secreted fraction identified the overwhelming majority of known secreted virulence factors and revealed more than 20 new putative secreted virulence factors. In parallel, we utilized SIEVE (SVM-based identification and evaluation of virulence effectors), a machine learning algorithm we developed (Samudrala et al. 2009; McDermott et al. 2011b), to predict novel secreted effectors.

Coupling the SIEVE algorithm with the proteomics data proved to be an efficient way to select novel proteins for characterization. We tested ten proteins based on input from the SIEVE algorithm and proteomics data for secretion into J774 macrophages using CyaA' assays and confirmed that eight of the ten were secreted into the macrophage cytosol. Additional in vivo infection studies demonstrated that deletion mutants of six of the above eight confirmed secreted proteins (Δ*spvD*, Δ*steE*, Δ*gtgE*, Δ*steD*, Δ*sssA* and Δ*ssaB*) were attenuated for virulence. Importantly, these results demonstrate the utility of a systems approach for predicting proteins relevant to understanding host–pathogen interactions.



S. Typhimurium	Absent	Primary proliferation	Secondary proliferation; fucose use	Clearance
Host	Homeostasis	Inflammatory response	Glycan release, inflammation	Clearing pathogen
Commensals	Abundant	Decreasing	Low	Increasing
Metabolites	Homeostasis	Sugar accumulation	Sugar accumulation	Regaining homeostasis

Fig. 4 Model of host–pathogen-commensal interactions during *S. Typhimurium*-induced gastroenteritis. Using a systems biology approach and the available literature, we developed a model of the interplay between the mouse, *S. Typhimurium*, and the commensal population during gastrointestinal infection. Prior to pathogen introduction, the commensal population thrives in the homeostatic gut. Early in infection, *S. Typhimurium* proliferates, stimulates an inflammatory response characterized by neutrophil activation, and disrupts this microbial community. As the commensal population profile changes, so do metabolites in the gut that are normally metabolized by the microbial community such as fucosylated glycans. *S. Typhimurium* senses and responds to fucose availability during gastrointestinal infection, as evidenced by increased expression of fucose utilization proteins. Finally, pathogen clearance from the gut occurs, allowing the gastrointestinal environment to begin to return to pre-infection conditions

6.3 Host–Pathogen Interactions in the Gut Microbiome

The commensal microbiota of the host represents a relatively unexplored contributor to the host–microbe interactions during infection. As a complete understanding of pathogenesis will undoubtedly need to consider the host microbiota, we undertook an exploratory study to investigate the interplay between host, pathogen, and commensal microbes during *S. Typhimurium*-induced gastroenteritis.

For these studies, we chose a mouse model of persistent Salmonellosis, which requires no antibiotic treatment prior to infection and allows *Salmonella* colonization of the gut, allowing us to observe activities of the commensal microbial population. Application of integrated proteomics, metabolomics, metagenomics, and glycomics measurements revealed oral *Salmonella* infection disrupts the commensal population, which allows *S. Typhimurium* to proliferate; concurrently, the host immune system (specifically neutrophil infiltration and release of various inflammatory markers) is activated (Fig. 4). Loss of commensal microbes (likely

due in part to the host inflammatory response) and their associated functions is evident mid-way through infection, when metabolites such as fucose and other sugars normally utilized by commensal bacteria accumulate in the gut. During this time, *Salmonella* thrives, sensing increased host glycan release and utilizing available fucose moieties, among other functions. Resolution of infection by later time points is observed, with a decrease in *S. Typhimurium* abundance, re-establishment of metabolite composition, and outgrowth of indigenous microbiota. Importantly, this model of interactions during *Salmonella*-induced gastroenteritis provides a framework that is both consistent with known factors and provides new insights into infection through integration of omics studies. We anticipate that future endeavors will similarly take advantage of the increased knowledge that can be gained through this systems-level approach.

7 Conclusion and Future Prospects

In this chapter, we have highlighted application of our systems biology approach to investigate interactive host–pathogen mechanisms necessary for two closely related pathogens *Salmonella* and *Yersinia* to cause systemic infection. With the increasing body of knowledge and data arising from high-throughput omics approaches, it is very important that more sophisticated computational approaches be developed to use this information. For example, the integration of inference and knowledge-based modeling approaches discussed above. Comprehensive system models of *Yersinia* and *Salmonella* pathogenesis will have applications for antibiotic development, new strategies for therapeutic treatments, and further understanding of the complex interplay between pathogen and host and the microbiota during infection. In addition to what has been discussed, the reconstruction of other networks including transcriptional regulatory networks and more recently transcription and translation processes (i.e. macromolecular synthesis) are becoming established (Herrgard et al. 2004; Thiele et al. 2009). Methods for their integration with the metabolic models discussed here are in development and should provide a more comprehensive systems-level model enabling systems-level simulations of host–pathogen interactions.

Acknowledgments Research described was supported by the National Institute of Allergy and Infectious Diseases NIH/DHHS through Interagency agreement Y1-AI-8401. Proteomics and metabolomics capabilities were developed under support from the US Department of Energy (DOE) Office of Biological and Environmental Research (BER) and the NIH grants 5P41RR018522-10 and the National Institute of General Medical Sciences grant (8 P41 GM103493-10), and work was performed in the Environmental Molecular Sciences Laboratory, a DOE-BER national scientific user facility at Pacific Northwest National Laboratory (PNNL). PNNL is a multiprogram national laboratory operated by Battelle Memorial Institute for the DOE under contract DE-AC05-76RLO 1830.

References

Adams MD (1996) Serial analysis of gene expression: ESTs get smaller. BioEssays 18(4): 261–262

Anderson L, Seilhamer J (1997) A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18(3–4):533–537

Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD (2008a) Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct Genomic Proteomic 7(1):50–62

Ansong C, Tolic N, Purvine SO, Porwollik S, Jones M, Yoon H, Payne SH, Martin JL, Burnet MC, Monroe ME et al (2011) Experimental annotation of post-translational features and translated coding regions in the pathogen *Salmonella Typhimurium*. BMC Genomics 12:433

Ansong C, Yoon H, Norbeck AD, Gustin JK, McDermott JE, Mottaz HM, Rue J, Adkins JN, Heffron F, Smith RD (2008b) Proteomics analysis of the causative agent of typhoid fever. J Proteome Res 7(2):546–557

Ansong C, Yoon H, Porwollik S, Mottaz-Brewer H, Petritis BO, Jaitly N, Adkins JN, McClelland M, Heffron F, Smith RD (2009) Global systems-level analysis of Hfq and SmpB deletion mutants in Salmonella: implications for virulence and global protein translation. PLoS One 4(3):e4809

Armengaud J (2009) A perfect genome annotation is within reach with the proteomics and genomics alliance. Curr Opin Microbiol 12(3):292–300

Bonneau R (2008) Learning biological networks: from modules to dynamics. Nat Chem Biol 4(11):658–664

Bordbar A, Lewis NE, Schellenberger J, Palsson BO, Jamshidi N (2010) Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. Mol Syst Biol 6:422

Brenner DJ (1978) Characterization and clinical identification of Enterobacteriaceae by DNA hybridization. Prog Clin Pathol 7:71–117

Brenner DJ, Falkow S (1971) Genetics of the Enterobacteriaceae. C. Molecular relationships among members of the Enterobacteriaceae. Adv Genet 16:81–118

Brenner DJ, Fanning GR, Johnson KE, Citarella RV, Falkow S (1969) Polynucleotide sequence relationships among members of Enterobacteriaceae. J Bacteriol 98(2):637–650

Brubaker RR (1991) Factors promoting acute and chronic diseases caused by yersiniae. Clin Microbiol Rev 4(3):309–324

Cascante M, Marin S (2008) Metabolomics and fluxomics approaches. Essays Biochem 45:67–81

Cavanaugh DC, Randall R (1959) The role of multiplication of *Pasteurella pestis* in mononuclear phagocytes in the pathogenesis of flea-borne plague. J Immunol 83:348–363

Charusanti P, Chauhan S, McAteer K, Lerman JA, Hyduke DR, Motin VL, Ansong C, Adkins JN, Palsson BO (2011) An experimentally-supported genome-scale metabolic network reconstruction for *Yersinia pestis* CO92. BMC Syst Biol 5:163

Coburn B, Grassl GA, Finlay BB (2007) Salmonella, the host and disease: a brief review. Immunol Cell Biol 85(2):112–118

Coombes BK, Wickham ME, Lowden MJ, Brown NF, Finlay BB (2005) Negative regulation of Salmonella pathogenicity island 2 is required for contextual control of virulence during typhoid. Proc Natl Acad Sci U S A 102(48):17460–17465

Cornelis GR (1998) The Yersinia deadly kiss. J Bacteriol 180(21):5495–5504

Cornelis GR (2002) Yersinia type III secretion: send in the effectors. J Cell Biol 158(3):401–408

de Groot A, Dulermo R, Ortet P, Blanchard L, Guerin P, Fernandez B, Vacherie B, Dossat C, Jolivet E, Siguier P et al (2009) Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium *Deinococcus deserti*. PLoS Genet 5(3):e1000434

De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. Nat Rev Microbiol 8(10):717–729

Deiwick J, Nikolaus T, Erdogan S, Hensel M (1999) Environmental regulation of Salmonella pathogenicity island 2 gene expression. Mol Microbiol 31(6):1759–1773

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 104(6):1777–1782

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. PLoS Biol 5(1):e8

Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7(2):129–143

Feist AM, Palsson BO (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol 26(6):659–667

Feist AM, Palsson BO (2010) The biomass objective function. Curr Opin Microbiol 13(3): 344–349

Fields PI, Swanson RV, Haidaris CG, Heffron F (1986) Mutants of *Salmonella typhimurium* that cannot survive within the macrophage are avirulent. Proc Natl Acad Sci U S A 83(14): 5189–5193

Fukuto HS, Svetlanov A, Palmer LE, Karzai AW, Bliska JB (2010) Global gene expression profiling of *Yersinia pestis* replicating inside macrophages reveals the roles of a putative stress-induced operon in regulating type III secretion and intracellular cell division. Infect Immun 78(9):3700–3715

Glynn MK, Bopp C, Dewitt W, Dabney P, Mokhtar M, Angulo FJ (1998) Emergence of multidrug-resistant *Salmonella enterica* serotype typhimurium DT104 infections in the United States. N Engl J Med 338(19):1333–1338

Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc Natl Acad Sci U S A 97(17): 9390–9395

Haynes PA, Gygi SP, Figeys D, Aebersold R (1998) Proteome analysis: biological assay or data archive? Electrophoresis 19(11):1862–1871

Herrgard MJ, Covert MW, Palsson BO (2004) Reconstruction of microbial transcriptional regulatory networks. Curr Opin Biotechnol 15(1):70–77

Hyduke DR, Palsson BO (2010) Towards genome-scale signalling network reconstructions. Nat Rev Genet 11(4):297–307

Jaffe JD, Berg HC, Church GM (2004) Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4(1):59–77

Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely SA, Palsson BO, Agarwalla S (2006) Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. J Bacteriol 188(23):8259–8271

Kueger S, Steinhauser D, Willmitzer L, Giavalisco P (2012) High-resolution plant metabolomics: from mass spectral features to metabolites and from whole-cell analysis to subcellular metabolite distributions. Plant J 70(1):39–50

Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol 1(1):E19

Lewis NE, Cho BK, Knight EM, Palsson BO (2009) Gene expression profiling and the use of genome-scale in silico models of *Escherichia coli* for analysis: providing context for content. J Bacteriol 191(11):3437–3444

Ly M, Laremore TN, Linhardt RJ (2010) Proteoglycomics: recent progress and future challenges. OMICS 14(4):389–399

Margolin AA, Califano A (2007) Theory and limitations of genetic network inference from microarray data. Ann N Y Acad Sci 1115:51–72

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12(10): 671–682

McDermott JE, Archuleta M, Thrall BD, Adkins JN, Waters KM (2011a) Controlling the response: predictive modeling of a highly central, pathogen-targeted core response module in macrophage activation. PLoS One 6(2):e14673

McDermott JE, Corrigan A, Peterson E, Oehmen C, Niemann G, Cambronne ED, Sharp D, Adkins JN, Samudrala R, Heffron F (2011b) Computational prediction of type III and IV secreted effectors in gram-negative bacteria. Infect Immun 79(1):23–32

McDermott JE, Taylor RC, Yoon H, Heffron F (2009) Bottlenecks and hubs in inferred networks are important for virulence in *Salmonella typhimurium*. J Comput Biol 16(2):169–180

Miao EA, Miller SI (2000) A conserved amino acid sequence directing intracellular type III secretion by *Salmonella typhimurium*. Proc Natl Acad Sci U S A 97(13):7539–7544

Nicholson JK, Lindon JC, Holmes E (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29(11):1181–1189

Niemann GS, Brown RN, Gustin JK, Stufkens A, Shaikh-Kidwai AS, Li J, McDermott JE, Brewer HM, Schepmoes A, Smith RD et al (2011) Discovery of novel secreted virulence factors from *Salmonella enterica* serovar Typhimurium by proteomic analysis of culture supernatants. Infect Immun 79(1):33–43

Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320

Oldiges M, Lutz S, Pflug S, Schroer K, Stein N, Wiendahl C (2007) Metabolomics: current state and evolving methodologies and tools. Appl Microbiol Biotechnol 76(3):495–511

Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28(3): 245–248

Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12(2):87–98

Palsson B (2004) Two-dimensional annotation of genomes. Nat Biotechnol 22(10):1218–1219

Palsson B, Zengler K (2010) The challenges of integrating multi-omic data sets. Nat Chem Biol 6(11):787–789

Payne SH, Huang ST, Pieper R (2010) A proteogenomic update to Yersinia: enhancing genome annotation. BMC Genomics 11:460

Perry RD, Fetherston JD (1997) *Yersinia pestis*–etiologic agent of plague. Clin Microbiol Rev 10(1):35–66

Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nat Rev Genet 7(2):130–141

Samudrala R, Heffron F, McDermott JE (2009) Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. PLoS Pathog 5(4):e1000375

Schellenberger J, Park JO, Conrad TM, Palsson BO (2010) BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11:213

Schrimpe-Rutledge AC, Jones MB, Chauhan S, Purvine SO, Sanford JA, Monroe ME, Brewer HM, Payne SH, Ansong C, Frank BC et al (2012) Comparative omics-driven genome annotation refinement: application across Yersiniae. PLoS One 7(3):e33903

Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M (2011) Global quantification of mammalian gene expression control. Nature 473(7347):337–342

Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A 99(23):15112–15117

Sharp PM (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J Mol Evol 33(1):23–33

Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, Hinton JC, Vogel J (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator Hfq. PLoS Genet 4(8):e1000163

Straley SC, Harmon PA (1984a) Growth in mouse peritoneal macrophages of *Yersinia pestis* lacking established virulence determinants. Infect Immun 45(3):649–654

Straley SC, Harmon PA (1984b) *Yersinia pestis* grows within phagolysosomes in mouse peritoneal macrophages. Infect Immun 45(3):655–659

Straley SC, Plano GV, Skrzypek E, Haddix PL, Fields KA (1993) Regulation by Ca2+ in the Yersinia low-Ca2+ response. Mol Microbiol 8(6):1005–1010

Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, Bazzani S, Charusanti P, Chen FC, Fleming RM, Hsiung CA et al (2011) A community effort towards a knowledge-base and mathematical model of the human pathogen *Salmonella Typhimurium* LT2. BMC Syst Biol 5:8

Thiele I, Jamshidi N, Fleming RM, Palsson BO (2009) Genome-scale reconstruction of *Escherichia coli's* transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. PLoS Comput Biol 5(3):e1000312

Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5(1):93–121

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW (1997) Characterization of the yeast transcriptome. Cell 88(2):243–251

Vidal SM, Malo D, Marquis JF, Gros P (2008) Forward genetic dissection of immunity to infection in the mouse. Annu Rev Immunol 26:81–132

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63

Welkos SL, Davis KM, Pitt LM, Worsham PL, Freidlander AM (1995) Studies on the contribution of the F1 capsule-associated plasmid pFra to the virulence of *Yersinia pestis*. Contrib Microbiol Immunol 13:299–305

Welkos SL, Friedlander AM, Davis KJ (1997) Studies on the role of plasminogen activator in systemic infection by virulent *Yersinia pestis* strain C092. Microb Pathog 23(4):211–223

White AP, Weljie AM, Apel D, Zhang P, Shaykhutdinov R, Vogel HJ, Surette MG (2010) A global metabolic shift is linked to Salmonella multicellular development. PLoS One 5(7):e11814

Worsham PL, Stein MP, Welkos SL (1995) Construction of defined F1 negative mutants of virulent *Yersinia pestis*. Contrib Microbiol Immunol 13:325–328

Wright JC, Sugden D, Francis-McIntyre S, Riba-Garcia I, Gaskell SJ, Grigoriev IV, Baker SE, Beynon RJ, Hubbard SJ (2009) Exploiting proteomic data for genome annotation and gene model validation in *Aspergillus niger*. BMC Genomics 10:61

Yoon H, Ansong C, McDermott JE, Gritsenko M, Smith RD, Heffron F, Adkins JN (2011) Systems analysis of multiple regulator perturbations allows discovery of virulence factors in Salmonella. BMC Syst Biol 5:100

Zhang W, Li F, Nie L (2010) Integrating multiple 'omics' analysis for microbial biology: application and methodologies. Microbiology 156(Pt 2):287–301

ChIP-Seq and the Complexity of Bacterial Transcriptional Regulation

James Galagan, Anna Lyubetskaya and Antonio Gomes

Abstract Transcription factors (TFs) play a central role in regulating gene expression in all bacteria. Yet, until recently, studies of TF binding were limited to a small number of factors at a few genomic locations. Chromatin immunoprecipitation followed by sequencing enables mapping of binding sites for TFs in a global and high-throughput fashion. The NIAID funded TB systems biology project http://www.broadinstitute.org/annotation/tbsysbio/home.html aims to map the binding sites for every transcription factor in the genome of *Mycobacterium tuberculosis* (MTB), the causative agent of human TB. ChIP-Seq data already released through TBDB.org have provided new insight into the mechanisms of TB pathogenesis. But in addition, data from MTB are beginning to challenge many simplifying assumptions associated with gene regulation in all bacteria. In this chapter, we review the global aspects of TF binding in MTB and discuss the implications of these data for our understanding of bacterial gene regulation. We begin by reviewing the canonical model of bacterial transcriptional regulation using the lac operon as the standard paradigm. We then review the use of ChIP-Seq to map the binding sites of DNA-binding proteins and the application of this method to mapping TF binding sites in MTB. Finally, we discuss two aspects of the binding discovered by ChIP-Seq that were unexpected given the canonical model: the substantial binding outside the proximal promoter region and the large number of weak binding sites.

J. Galagan (✉)
Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA
e-mail: jgalag@broadinstitute.org

J. Galagan
Department of Microbiology, Boston University, Boston, MA 02215, USA

J. Galagan · A. Lyubetskaya · A. Gomes
Bioinformatics Program, Boston University, Boston, MA 02215, USA

J. Galagan
The Eli and Edythe L. Broad Institute of Harvard, MIT, Cambridge, MA 02142, USA

Current Topics in Microbiology and Immunology (2013) 363: 43–68
DOI: 10.1007/82_2012_257

Published Online: 15 September 2012

Contents

1 Introduction

Transcription factors (TFs) play a central role in regulating gene expression in all bacteria. Yet, until recently, studies of TF binding were limited to a small number of factors at a few genomic locations. Although these data have provided a wealth of detailed mechanistic insight, they have also been extrapolated and simplified to form a set of widely help assumptions about the global nature of TF binding in prokaryotes. Only recently have techniques become available to map TF binding sites in an unbiased fashion. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) provides the ability to globally map binding sites for TFs, and the scalability of the technology enables the ability to map binding sites for every DNA binding protein in a prokaryotic organism.

As part of the NIAID funded TB systems biology project http://www.broadinstitute.org/annotation/tbsysbio/home.html, an effort to map the binding sites for every transcription factor in the *Mycobacterium tuberculosis* (MTB) genome is underway. MTB is the causative agent of human tuberculosis. With more than 8 million new cases of active disease and nearly 1.5 million deaths annually, TB is a global health emergency of overwhelming proportions (World Health Organization 2001). Mapping the binding sites for all MTB TFs and using these data to reconstruct the regulatory network of this organism promises to reveal mechanistic insight into TB pathogenesis that could be used to speed the development of new and more effective drugs, vaccines, and diagnostics. And the analysis of data from the mapping of 50 TFs has already led to biological insights (Galagan et al. Submitted). But beyond the insights into MTB biology, the ChIP-Seq data from MTB are also beginning to challenge the simplifying assumptions of gene regulation in bacteria in general.

In this chapter, we review the surprising diversity of MTB TF binding sites being discovered by ChIP-Seq and discuss the implications of these data for our understanding of bacterial gene regulation. We begin by reviewing the canonical model of bacterial transcriptional regulation using the lac operon as the standard paradigm. We then review the use of ChIP-Seq to map the binding sites of DNA-binding proteins and the application of this method to mapping TF binding sites in MTB.

Finally, we discuss two aspects of the binding discovered by ChIP-Seq that were unexpected given the canonical model: the substantial binding outside the proximal promoter region and the large number of weak binding sites. We review the background literature that helps place these data into a biological context, and helps extend the canonical prokaryotic regulatory model to encompass these new findings. We also describe the public release and availability of these data at Tuberculosis Database (TBDB.org).

2 The Canonical Model of Bacterial Transcriptional Regulation: The Lac Operon

In prokaryotes, the core event of transcription is the binding of a DNA-dependent RNA polymerase to the promoter region of an operon. The core polymerase is a multisubunit enzyme that is capable of transcription, but not the initiation of transcription. Nor is the core polymerase enzyme capable of regulating the selection of promoters to which it binds. Transcription initiation requires that the polymerase interacts with a sigma factor to form a holoenzyme. Sigma factors also represent the first level of transcriptional regulation in that they direct that holoenzyme to specific promoter sequences. Bacteria typically possess multiple sigma factors that enable recognition of multiple sets of promoters in response to changing cellular states. Additionally, bacteria possess many different TFs that can bind to DNA and influence the transcription rate at different promoters. TFs respond to environmental and cellular conditions and the activity of other regulatory proteins to orchestrate the genome-wide pattern of transcription. Understanding the interactions between TFs and the mechanisms by which TFs modulate transcription is thus a central challenge for understanding the functioning of cells.

The classical model for studying the mechanisms of gene regulation in prokaryotes is the lac operon in *Escherichia coli (E. coli)*. The lac operon encodes three genes required for the digestion of lactose. Through a two-part regulatory mechanism, *E. coli* only transcribes these three genes when necessary. Ground breaking experiments by Jacob and Monod (1961) and later by Gilbert (Gilbert and Muller-Hill 1966) provided the framework by which this regulation occurs. This framework, in turn, has provided the paradigm for understanding regulation in prokaryotes in general. The lac operon has been well described, and we only review the core aspects here. The essential characteristic of this framework is the binding of two different proteins to the lac operon promoter region: the lac repressor and the catabolite activator protein (CAP).

The lac repressor represses the lac operon in the absence of lactose. In this condition, the repressor protein binds tightly to an operator site in the promoter of the lac operon just upstream of the transcription start site called O_1 (Fig. 1a). This binding is the central mechanism of repression as it sterically inhibits access to the promoter by the polymerase. When lactose is present, the inducer allolactose is

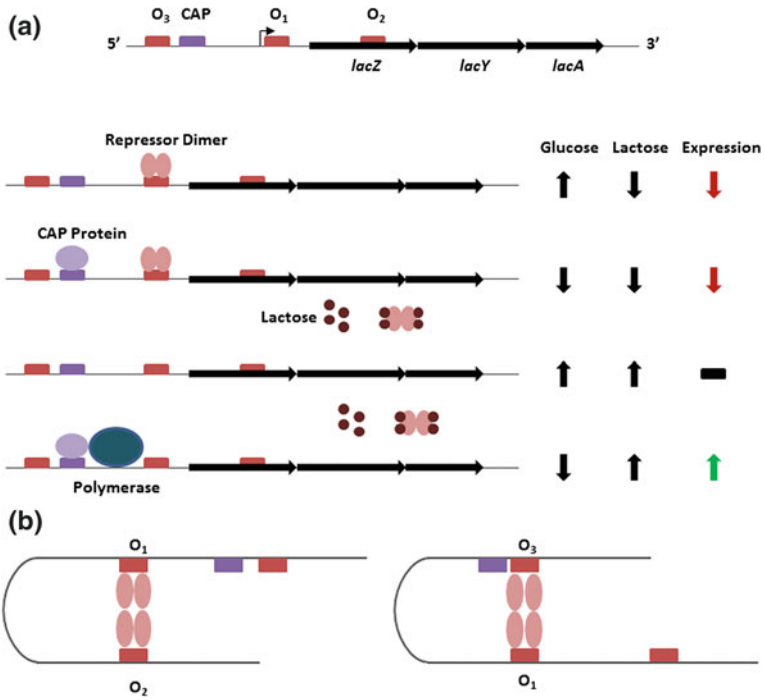


Fig. 1 The lac operon model of bacterial transcription initiation regulation. **a** Canonical model of regulation of the lac operon by the lac repressor and catabolite activator protein (CAP). **b** Model of DNA looping of the lac operon. Schematic of DNA looping is shown as a simplified schematic only. Many different topologies of the looped DNA are possible

produced and binds to the lac repressor, rendering the repressor unable to bind. CAP activates the lac operon in the absence of glucose. In this condition, cAMP is produced and binds to CAP, enabling CAP to bind to a binding site in the promoter upstream of the polymerase binding site (Fig. 1a). Interactions between CAP and the polymerase facilitate binding of the latter to the promoter region, thus activating transcription.

It is now understood that the full complexity of lac operon regulation is more complex than this simplified model. Yet, the basics of this model drive much of the interpretation of bacterial regulation. In particular, the canonical model implies that TFs should bind primarily to the proximal promoter such that direct interactions with the polymerase complex can mediate regulatory effects. As described below, however, many examples exist of more complex mechanisms. Moreover, genome-wide mapping data from chromatin immunoprecipitation (ChIP) studies for TFs are suggesting that such mechanisms may be more common than previously thought.

3 Chromatin Immunoprecipitation for Mapping DNA Binding Sites

Chromatin immunoprecipitation followed by sequencing is a method for globally mapping the binding locations of a protein on the genome sequence of an organism (Fig. 2) (Johnson et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007a). ChIP is the first step. This is performed on a population that can range from 10^4 to 10^7 cells (Park 2009). Proteins bound to the genomes of these cells are cross-linked to DNA with formaldehyde. The cells are then broken open and the DNA sheared through sonication or enzymatic digestion. DNA fragments bound by a protein of interest (a transcription factor, for example) are immunoprecipitated using an antibody to the protein. The antibody can either be selected to recognize the native protein, or to recognize an epitope (or tag) genetically engineered into the protein sequence (Kim et al. 2008; Mazzoni et al. 2011). Cross-linking is then reversed to remove the proteins, the precipitated DNA fragments are isolated, and sequencing used to generate reads from the ends of the fragments. Sequencing reads are aligned to the corresponding genome sequence, and genomic locations from which the DNA fragments are derived are identified as regions that are over-represented with aligned reads. (An older technology, called ChIP–ChiP, uses hybridization to a microarray to identify the location of DNA fragments).

Ideally, only genomic regions that were bound by the protein of interest would display read coverage. In practice, DNA fragments will be nonspecifically isolated and sequenced as well, resulting in a background coverage of reads aligning across the genome sequence. To assess this background coverage, one or more control experiments are typically used. Several different types of controls can be utilized that assess different processes giving rise to background coverage. Mock ChIP runs are frequently used; they include every step of the ChIP process with the exception of the addition of the antibody. These experiments control for the many non-antibody steps of ChIP that may lead to nonspecific DNA isolation. In cases where antibodies are used against genetically tagged proteins, the same antibodies can be used for ChIP against the strains that lack the genetic tag. Such experiments assess the degree to which the antibody recognizes nonspecific targets. A genomic DNA preparation can also be generated as a control. Sequencing of this genomic DNA, as well as the other control preparations described, help control for the differential efficiency of isolation and sequencing of different locations of the genome. For example, certain genomic regions are highly susceptible to isolation—perhaps owing to chromatin structure, base composition, or repetitive nature—that may appear as false-positive peaks in coverage relative to the rest of the genome in control lanes.

Binding sites for the protein of interest are regions along the genome characterized by greater read coverage in the protein ChIP-Seq experiment than the background coverage. These can be identified using a wide range of available computational tools (Pepke et al. 2009; Wilbanks and Facciotti 2010). The degree to which the protein ChIP-Seq shows greater coverage than background is termed

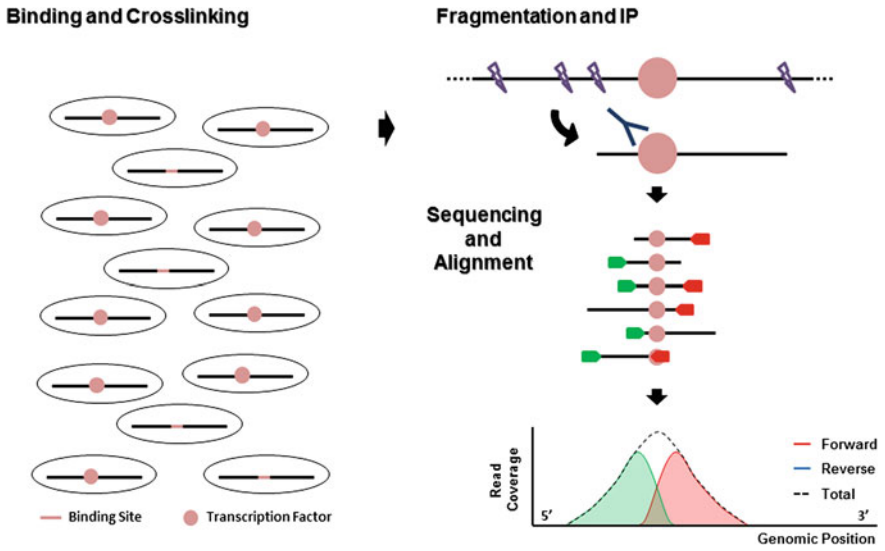


Fig. 2 Schematic overview of Chromatin Immunoprecipitation followed by Sequencing for mapping DNA Binding Sites. DNA-binding proteins are crosslinked to binding sites in a population of cells. The cells are broken open and the DNA sheared through sonication or enzymatic digestion. DNA fragments bound by a protein of interest (a transcription factor, for example) are immunoprecipitated using an antibody to the protein. DNA fragments are isolated, and sequencing used to generate reads from the ends of the fragments. Sequencing reads are aligned to the corresponding genome sequence, and genomic locations from which the DNA fragments are derived are identified as regions that are over-represented with aligned reads. Because sequencing reads are generated from one end or the other of the DNA fragments in a randomly selected fashion, reads from the forward strand align upstream of the binding site and reads from the reverse strand align downstream

enrichment, and most algorithms effectively select a threshold on enrichment. In some cases, the control data are used directly to calculate enrichment, while in other cases a statistical model is built based on the control data. At any given threshold, some regions may display enrichment due to chance fluctuations in coverage, giving rise to false discovery of peaks. Thus, most algorithms also estimate a false discover rate (FDR). For example, in the case where the background coverage is modeled by an explicit distribution, the coverage associated with protein ChIP-Seq can be assigned a p value relative to this distribution. Standard methods can then be used to select a p value threshold to control for FDR (Benjamini and Hochberg 1995; Storey 2002, 2003; Storey and Tibshirani 2003).

ChIP-Seq also produces a strand specific signature of enrichment that can be used to identify true binding peaks. When DNA fragments from ChIP are sequenced, reads are typically generated from one end of the fragment or the other. If the fragment is bound by the protein of interest, the binding site will occur between the ends of the DNA fragment. Thus sequencing reads will align to one side or the other of the binding site. Sequence reads are generated from 5′ to 3′.

Thus reads that are generated from the 5′ end of the DNA fragment, upstream of the binding site, will align to the forward strand of the genome reference. Conversely, reads from the 3′ end of the DNA fragment, downstream of the binding site, will align to the reverse strand. If coverage is visualized for the two strands separately, this process gives rise to a bimodal enrichment profile: the coverage of the forward strand will be shifted upstream with respect to the actual binding site, while the coverage on the reverse strand will be shifted downstream (Fig. 2). The distance between the forward and reverse coverage profile is determined by the size of the DNA fragments sequenced, which can be estimated during the sequencing process. Nonspecific binding, by contrast, often results in enrichment that lacks this bimodal shift. This shift can thus be used to filter for localized binding events. Specifically, the spatial cross-correlation between the forward and reverse coverage can be calculated, and regions corresponding to the desired binding will show a positive correlation at a spatial lag roughly equivalent to the average DNA fragment size.

DNA binding proteins, especially TFs, typically bind to short DNA sequences (on the order of 15 bp or less). Enriched peaks, however, typically span a region of several hundred base pairs as a consequence of the larger fragment size generated during ChIP (typically around 250 bp). Moreover, when multiple closely spaced binding sites for a protein exist in a particular location, the read coverage for these sites can merge into a single broad enriched region. Several different methods have been used for identifying the specific binding site(s) within enriched regions. The most straightforward method is to select the specific point of highest total read coverage within an enriched region. Since forward and reverse read coverage profiles are separated by the DNA fragment length, these two profiles can be shifted toward each other by half the expected DNA fragment length and then summed to ensure a single coverage peak. Due to variations in coverage or multiple binding sites, however, the highest peak in an enriched region may not be easy to determine or be an accurate estimate of the underlying binding sites.

An alternative approach is to treat binding site detection as a signal detection problem. Conceptually, binding sites can be considered a point source input that, through the process of ChIP-Seq, give rise to broader output signal of coverage. Multiple binding sites give rise, to a first approximation, to an output coverage that is the sum of the outputs of each of the individual binding sites. This process can be modeled as a linear convolution. In this model, the output signal arising from a point input is called an impulse response (or point spread function). In the case of the ChIP-Seq, the impulse response is a consequence of "transmitting" the input signal through the process of randomly sequencing the ends of large DNA fragments that overlap the point source and "receiving" this transmission in terms of coverage after aligning these reads. The impulse function essentially "blurs" the output signal arising from a point input. Input signals are modeled as the sum of multiple point sources, or impulse functions, that are each scaled to a particular magnitude. In the case of ChIP-Seq, impulse functions correspond to binding sites where the scaling associated with each site corresponds roughly to relative enrichment (and thus relative occupancy as described above). The output signal is

then the sum of the correspondingly scaled impulse functions, or a convolution in mathematical terms (specifically a discrete convolution since DNA coordinates are integer based).

The operation of recovering the binding site locations from regions of enriched coverage is then a de-convolution–or the inverse process of convolution. In the case where the impulse function is known, this process is straightforward (Oppenheim et al. 1997). For ChIP-Seq, however, this function is not known and depends in part on the details of each specific experiment. Thus, the impulse function must be estimated at the same time that the coverage signal is being de-convolved. This operation is termed a blind de-convolution (blind because the impulse function is not known a priori), and techniques have also been developed to solve this problem (Levin et al. 2011).

Blind deconvolution methods have been developed for ChIP-Seq processing (Gomes et al. In Preparation; Lun et al. 2009). One method, called CSDeconv, uses a re-estimation method to solve the blind de-convolution problem. The basic approach begins by generating an initial estimate of the impulse function. This is typically generated by selecting a set of peaks with high coverage that are then used to fit a model of the impulse function based on an initial estimate of the binding site locations in these regions. This estimated impulse function can then be used to perform blind de-convolution on all enriched regions which results in new binding site locations for all enriched regions. These new binding site locations are then used to fit an updated model of the impulse function, and the process iterates in this fashion until convergence criteria are reached.

The final output of this method is a list of binding sites with high spatial resolution and accuracy (Lun et al. 2009). Based on a test using ChIP-Seq data for the GABP transcription factor in human and the DosR transcription factor in MTB, this approach is able to identify binding locations to within an average absolute difference of less than 24 bp (Lun et al. 2009). Moreover, the method can accurately predict multiply spaced binding sites within the same ChIP-Seq enriched region. In the case of DosR, binding sites located less than 57 bp apart in the same intergenic region could be resolved, while several closely spaced binding sites for GABP were observed, two as close as 20 bp apart.

The model-based approach can also provide additional insight into the potential roles of closely spaced binding sites. As described in more detail below, closely spaced sites have been shown to mediate cooperative binding that can substantially alter the apparent affinity of individual sites. The de-convolution approach described above implicitly assumes that TFs bind individual sites independently of all other sites. The approach can be generalized, however, to explicitly model dependencies between sites for the same TF. Using this approach, it is possible to predict known cooperative interactions between closely spaced sites for the DosR transcription in MTB (Gomes et al. In Preparation).

More recently, a modification to ChIP-Seq, termed ChIP-exo (Rhee and Pugh 2011), has demonstrated the ability to experimentally resolve TF binding sites to within single nucleotide resolution, while also providing greater sensitivity for detecting binding sites. ChIP-exo utilizes the strand-specific 5′–3′ lambda (λ)

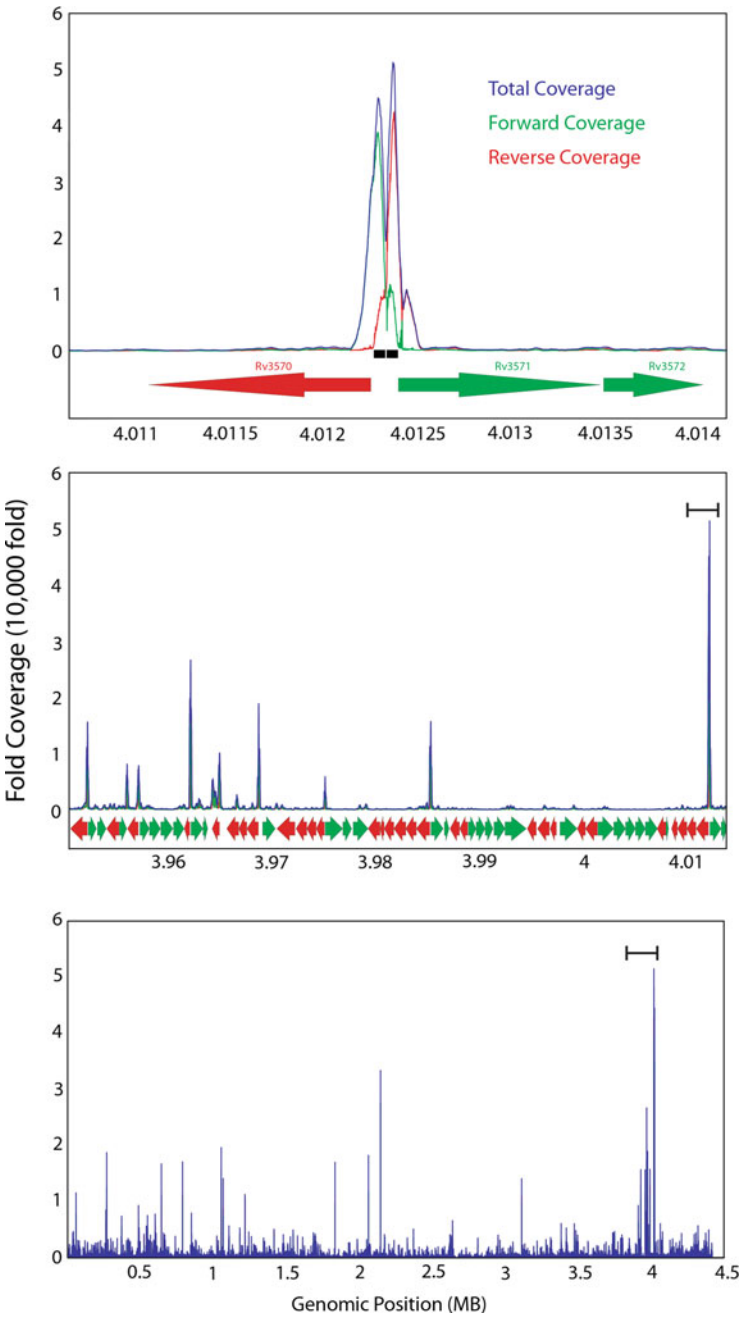
exonuclease to degrade DNA fragments isolated by ChIP. Bound proteins block the exonuclease and produce fragments in which one strand borders the protein binding site. The opposite strand remains intact and provides a template for sequencing. The result is a set of sequencing reads tightly spaced around the binding site, and thus whose coverage can be used to directly resolve the site with high spatial accuracy. Moreover, DNA fragments that are unbound by proteins but are nonspecifically isolated by ChIP are completely degraded which substantially decreases background coverage, increases signal-to-noise ratio for true binding sites, and increases peak prediction sensitivity. Although recently developed, ChIP-exo has produced binding site mapping data with sufficient resolution and accuracy to begin to confirm the complex pattern of TF binding suggested by ChIP-Seq data.

Finally, a variety of post-processing analyses are typically performed on ChIP-Seq data. First, enriched regions can be used to predict binding site motifs using a number of standard software packages (Bailey et al. 2009; Chen et al. 2008; Machanick and Bailey 2011). Although most commonly performed as a post-processing step after binding site locations have been identified, motif discovery can also be incorporated into the binding site prediction. The latter approach has the advantage that the underlying sequence motif can be used to enhance the accuracy of binding site prediction (Gomes et al. In Preparation). Second, binding sites can be used to infer potential functional relationships between TFs and target genes. The canonical model in prokaryotes suggests that binding in the proximal promoter region of a gene implies an interaction between the TF and that gene. However, even this simple association has exceptions and is complicated by the common occurrence of pairs of genes transcribed from a divergent promoter (in which case the site could modulate either or both genes, or neither). But the task of assigning binding sites to regulation has become far more challenging given the diversity of binding site locations discovered by ChIP mapping, as described next.

4 Large-Scale Mapping of MTB Transcription Factor Binding Sites

ChIP-Seq has been applied extensively to map TF binding sites in a range of eukaryotic organisms. The data from these studies have provided a wealth of information on the global binding patterns of TFs that have overturned many previously help assumptions about the nature, diversity, and possible functions of TF binding sites. Only recently, however, ChIP-Seq has been applied on a large scale to globally map TF binding sites in prokaryotes.

The most extensive publically available source of TF mapping data for a prokaryote using ChIP-Seq is the NIAID funded TB systems biology project (URL). An important aspect of the pathology of MTB is the ability to survive within macrophages of the human host for years without causing active disease. The adaptations required for this rare ability are not fully understood. The overall goal of the NIAID TB systems biology project is to comprehensively map the


Total Coverage
Forward Coverage
Reverse Coverage
Rv3570
Rv3571
Rv3572
4.011
4.0115
4.012
4.0125
4.013
4.0135
4.014
Fold Coverage (10,000 fold)
3.96
3.97
3.98
3.99
4
4.01
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Genomic Position (MB)

◀**Fig. 3** Example ChIP-Seq data set for the MTB transcription factor KstR. Each panel shows a plot of read coverage across the genome at different zoom levels. The x axis is the genome coordinate and the x axis is read fold coverage. The bars in the bottom two panels indicate the region that is displayed in the next panel up. The top panel displays a region in MTB known to contain two closely spaced binding sites; these sites are indicated by square boxes below the binding region

regulatory and metabolic programs that underlie this ability of MTB. Toward this end, the project has released ChIP-Seq data mapping the binding sites of all 50 TFs, toward the eventual goal of mapping all 200 TFs in the MTB genome. These data are all publically available at TBDB.org (see below).

The MTB genome is 4.4 Mb in length and contains approximately 4,000 genes. The small size of the MTB genome in particular and bacterial genomes in general serves as an advantage for ChIP-Seq mapping. ChIP-Seq coverage for binding sites scales with the overall number of reads generated relative to the size of the reference genome. A typical 40 bp sequencing lane on an Illumina GAIIx is sufficient to provide an average of 500-fold coverage for the MTB genome. With this degree of coverage, binding sites for individual TFs can be identified with coverage that spans several logs in magnitude (Fig. 3). The differences in coverage between different binding sites reflect the probability of occupancy of each site in the population of cells on which ChIP-Seq was performed. Occupancy, in turn, reflects a number of factors including the concentration and modification state of the TF, the affinity of the binding site for the TF, the accessibility of the binding site, and the availability of molecular co-factors. The high coverage that can be generated for ChIP-Seq in MTB provides insights into the variation of these factors on a genome-wide scale with unprecedented resolution (Fig. 3).

The resulting ChIP-Seq data from 50 TFs in MTB have confirmed several surprises that have also emerged from extensive ChIP-Seq mapping of TFs in nearly all other organisms. These surprises call into question some of the simplifying assumptions of the classical model of bacterial transcriptional regulation. In particular, the data are revealing that binding of TFs in MTB (1) occurs in many more diverse genomic locations than expected based on the canonical model of regulation, and (2) involves much more weak binding than previously known. The detailed analysis of these findings is reported in a separate manuscript (Galagan et al. Submitted). Here, we review the background literature that helps place these data into a biological context, and helps extend the canonical prokaryotic regulatory model to encompass these new findings.

5 Diverse Binding Locations

The canonical model of bacterial transcriptional initiation focusses on the role of binding in the proximal promoter region. ChIP-Seq data resulting from the mapping of 50 of the approximated 200 MTB TFs confirms that binding in upstream

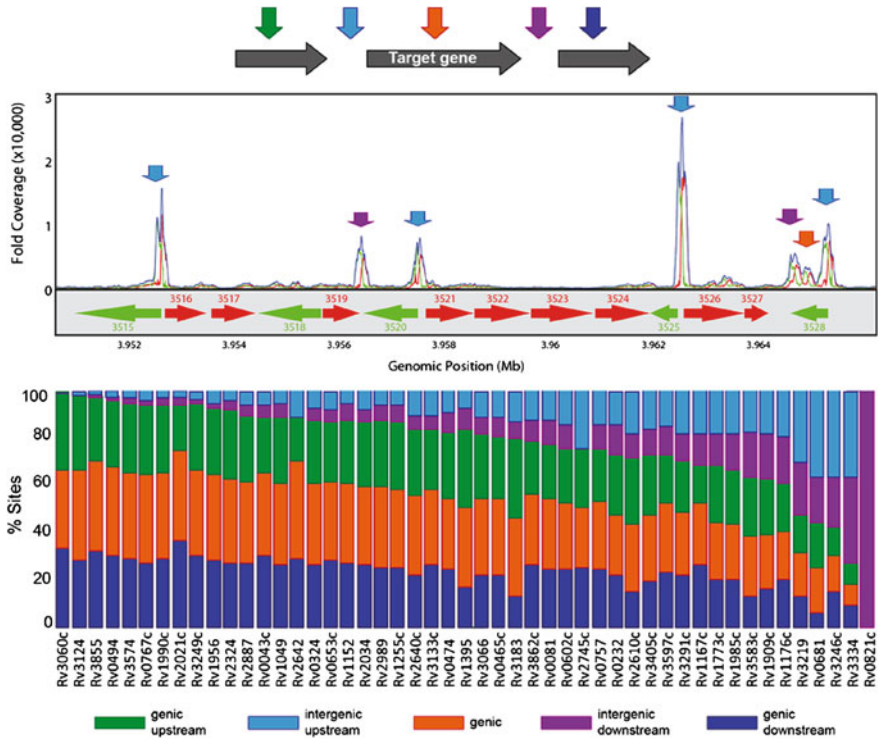


Fig. 4 Distribution of binding site locations from MTB ChIP-Seq data. The top panel displays the color coding used for categorizing binding site location, and displays an example region showing binding for the MTB TF KstR. The bottom panel shows the distribution of binding site locations for 49 MTB TFs

intergenic regions is enriched over what would be expected by chance. But surprisingly, binding in this region is the exception. As shown in Fig. 4, binding to intergenic regions represents less than 40 % of the binding events for any TF. The majority of binding events occur outside of upstream intergenic regions.

A number of explanations are possible that are consistent with the canonical model. The most straightforward explanation is the presence of errors in the annotation of coding regions. The majority of genes in all prokaryotic genomes are the result of computational predictions. Although the accuracy of computational algorithms for predicting the presence of an open reading frame is generally high (Delcher et al. 1999), the prediction of start codons remains a challenging problem. For a given coding region, with the rare exception of codon redefinition (Bekaert et al. 2010), the first stop codon is the essentially unambiguous end of the open reading frame. By contrast, multiple start codons are nearly always possible for an open reading frame, and the selection of the correct alternative is rarely obvious. Generally, computational algorithms are biased to selected start codons that produce longer reading frames. Moreover, numerous examples of alternative start

codon usage have been documented, further complicating the problem. These considerations suggest that, in some cases, binding sites that occur at the 5′ ends of annotated coding regions may in fact reflect intergenic binding relative to an actual downstream start codon.

Another explanation consistent with the canonical model is the well-known fact that promoter regions are not strictly limited to intergenic regions. Intergenic regions comprise only a small fraction of MTB genome, and all prokaryotes are similarly restricted in the amount of genomic real estate that is not occupied by genes. Although canonical promoter signals are enriched in intergenic regions and may be selected against in genic regions (Froula and Francino 2007; Huerta et al. 2006), many examples of promoter regions occurring in coding regions have been described in prokaryotes (Koide et al. 2009). Most notably, as described below, the lac operon on which the canonical model of regulation is based is known to contain a binding site for the lac repressor downstream of the proximal promoter region in the lacZ gene. Other examples include:

- In the bacterium *B. subtilis*, *ahrC* represses *argCAEBD-cpa-argF* operon binding at two sites: argC*o1* upstream (−60 to −9) and argC_{o2} within the coding region (+120) of argC. In vitro, the second binding site is bound only at high concentrations of the TF (i.e. the binding site within coding region has lower binding affinity). However, in vivo, argC_{o2} is essential for high levels of repression (Czaplewski et al. 1992).
- In the bacterium *C. crescentus*, the flagellar genes *flaN* and *flaG* are transcribed from a divergent promoter that includes a number of conserved *cis*-acting sequences. Two such *cis*-acting sequences, *ftr2* and *ftr3*, are located within the coding region of *flaN*, at positions +82 and +120 respectively. Mutations in either of these sequences dysregulate the expression of *flaN*. Moreover, mutations in *ftr3* also dysregulate the expression of *flaG* although this *cis*-acting sequence resides >300 bp from the start codon of *flaG* in the coding region of the upstream *flaN* gene (Mullin and Newton 1993).
- In the archeon *H. salinarium*, a comprehensive analysis of transcription start sites revealed extensive transcription initiation within coding sequences. A significant fraction of such initiation could be associated with the binding of known TFs inside annotated genes (Koide et al. 2009).

Although such examples are typically thought to be the exception, ChIP-Seq mapping data suggests they may be more common than previously expected. Furthermore, although the canonical view of transcriptional repression by TFs is associated with binding to the promoter region and blocking polymerase access, an analysis of *E. coli* TFs and their binding sites from the RegulonDB (Gama-Castro et al. 2011) indicates that repression occurs more often through binding downstream of the proximal promoter (Collado-Vides et al. 1991; Madan Babu and Teichmann 2003). Although downstream binding close to the promoter region may result in the repression of transcriptional initiation by blocking access to the polymerase complex, binding further downstream may also play a repressive role by blocking transcriptional elongation (Browning and Busby 2004).

These explanations based on local interactions with the proximal promoter, however, are unlikely to fully account for the binding that has been observed. In particular, the high frequency of binding sites for nearly all TFs that occur at a distance from the proximal promoter region suggests an important role for longer range interactions. Although interactions from distant binding sites are widely known to play important roles in regulating the activity of promoters in eukaryotes, such interactions are generally considered to be the exception in prokaryotes. Yet, there is substantial evidence in the literature that functional binding in prokaryotes can and does occur at much larger distances from promoters (Belitsky and Sonenshein 1999; Czaplewski et al. 1992; Dandanell et al. 1987; Dunn et al. 1984; Flashner and Gralla 1988; Narang 2007; Ninfa et al. 1987; Oehler et al. 1990; Reitzer and Magasanik 1986; Ueno-Nishio et al. 1983, 1984; Wedel et al. 1990).

Once again, the lac operon has provided a paradigm for understanding possible mechanisms. Although the primary interactions of the lac repressor and CAP provide the basis for the canonical model, a substantial body of the literature makes clear that the regulation of transcription initiation even in this system is substantially more complicated than the simplified model. Two general mechanisms play key roles in a more complicated picture of transcription regulation: DNA looping and cooperative interactions between distal and proximal binding sites.

As noted above, the *lac* repressor in *E. coli* binds to a high affinity operator site called O_1 that overlaps the transcription start site. Binding of the lac repressor to this site is sufficient for a degree of repression of the lac operon by the repressor in the absence of inducer. These data, however, only describe part of the mechanism of repression. In addition to the primary binding site O_1, two additional lower affinity binding sites for the lac repressor are also present (Reznikoff et al. 1974). One site (O_2) is located 401 bp downstream of O_1 in the lacZ gene. The other site (O_3) is located immediately 92 bp upstream of O_1 (Fig. 1a). Importantly, one of either O_2 or O_3 in combination with O_1 is required for full repression of lac by the repressor (Narang 2007). This was revealed by experiments that demonstrated that:

- Removal of either O_2 or O_3 decreases repression by the lac repressor 2 to 3-fold while removal of both decreases repression by over 50-fold (Oehler et al. 1990).
- The presence of either O_2 or O_3 in the absence of O_1 is not sufficient for repression (Oehler et al. 1990).
- Despite their low affinity, binding does occur to O_2 or O_3 although their effect on repression is abolished when both are moved further than 3,600 bp from O_1 (Oehler et al. 2006).
- Binding of the lac repressor to O_1 is strengthened threefold by O_2 (Flashner and Gralla 1988).
- Binding of the lac repressor to O_2 is strengthened 12-fold by O_1 (Flashner and Gralla 1988).

These data have led to model of lac operon regulation in which full repression requires the formation of a stable DNA loop mediated by the binding of a lac tetramer to O_1 and either O_2 or O_3 (Fig. 1b). DNA looping is mediated by the

formation of a lac repressor tetramer in which one dimer binds to the proximal site and the other dimer binds to a distal site. Binding solely at either O_2 or O_3 in the absence of binding at O_1 does not result in repression, confirming that blocking polymerase access to the proximal promoter is the primary mechanism of repression, consistent with the canonical model. But binding to O_1 is substantially potentiated by cooperative repression resulting from binding to either one of the distal operator sites and the consequent formation of an energetically favorable DNA loop. Binding of the lac repressor to O_2, in turn, provides a secondary mechanism of repression by blocking elongation of the lacZ gene (Flashner and Gralla 1988). Importantly, these cooperative interactions influence not only the overall magnitude of binding, but also the kinetics of binding with respect to inducer (Oehler et al. 2006), and thus heavily influence the degree of binding at low inducer or repressor concentrations.

The mechanisms of cooperative binding and DNA looping also play a role in the many other examples of long range interactions that have been described for prokaryotes. These examples include:

- In *E. coli*, repression of the L-arabinose operon *araBAD* by *araC* requires a binding site located 280 bp upstream of the *araBAD* operon which is located in the coding region of *araC* (Dunn et al. 1984; Schleif 2003). AraC proteins bind both sites and dimerizes to form a DNA loop (Dunn et al. 1984; Hahn et al. 1986; Lobell and Schleif 1990, 1991; Martin et al. 1986). This repressive loop formation requires that both sites be located on the correct face of the DNA double helix. Within this restriction, the two sites can be located up to 500 bp apart or directly adjacent to each other and still support looping (Lee and Schleif 1989).
- In *B. subtilis*, the RocR protein regulates *rocG* through a binding site located $\sim$100 nucleotides downstream of the 3′-end of the gene or 1.5 kb downstream of the *rocG* promoter. The binding site activates *rocG* transcription if relocated 15 kb downstream or upstream of the *rocG* promoter. The same RocR binding region is essential for regulation of the downstream *rocABC* operon; thus, the same binding area works as a canonical, upstream activation sequence of *rocABC* and novel, downstream activation sequence of *rocG* (Belitsky and Sonenshein 1999).
- In *E. coli*, the *glnALG* operon is regulated by the NR_I protein (also known as *glnG* or *ntrC*) through five binding sites located at positions −259 to −60. Binding sites 1 and 2 maintain their function if moved 700 bp upstream or 950 bp downstream of the promoter. Re-location of the regulatory region 3.1 kb upstream (high affinity sites only) or 3 kb downstream of the promoter (low affinity sites only) does not affect the activation of transcription at appropriate NR_I concentration levels (Ninfa et al. 1987; Reitzer and Magasanik 1986; Ueno-Nishio et al. 1983, 1984; Wedel et al. 1990). Consistent with DNA looping, binding site interactions require that sites be present on the same DNA molecular and sites 1 and 2 lose their function if moved closer to the promoter.
- In *E. coli*, the *deo* operon is repressed through three *deoR* binding sites. While one site is located at −8 (P2 operator), two additional sites are located at −606

(P1 operator), and −885. Binding sites P1 and P2 are essential for transcription factor function. Moreover, reporter constructs with P1 placed 1–5 kb downstream of the P2 show efficient repression (Dandanell et al. 1987). Again, consistent with a mechanism involving DNA looping, cooperative interactions between the two binding sites require that that they both be present on the same DNA molecule.

- In *K. pneumonia*, the *nifF* and *nifLA* operons are transcribed from a divergent intergenic region. Within this region, *nifA* binding to an upstream activator sequence over 200 bp from the *nifF* proximal promoter activates *nifF*. Similarly, two binding sites for NTRC are located over 140 bp from the proximal promoter of *nifLA* and binding by NTRC to these sites activate this operon (Minchin et al. 1988).
- In MTB, the gene Rv2034 is known to regulate the GroEL2 gene via a binding site 746 bp upstream of the GroEL2 start codon (Gao et al. 2011).

A common theme that emerges from such reports is that binding sites located more distant from the promoter appear to have weaker regulatory effects. This is consistent with experiments in which binding sites are experimentally moved (Dandanell et al. 1987; Ninfa et al. 1987; Reitzer and Magasanik 1986; Ueno-Nishio et al. 1983, 1984). These results likely reflect constraints on the size of DNA fragments that can be looped (Lee and Schleif 1989). These findings may also explain, in part, why regulation from more distal binding sites has not been more frequently reported. Binding with weaker regulatory effects would be more difficult to detect with standard perturbation experiments, and their effects could also be masked by stronger effects from more proximally located promoters. Although such distant TF binding sites may thus be less functionally impactful, the examples above suggest that they may remain functionally relevant nonetheless.

Finally, it is possible that TF binding sites in prokaryotes may also play roles in beyond the classical activation or repression of transcriptional regulation. In particular, in eukaryotes TF binding has been shown to modulate higher order DNA packaging and accessibility through the modulation of chromatin structure (Cao et al. 2010). Although bacteria lack histone proteins associated with eukaryotic chromatin, a wide range of proteins that perform analogous tasks have been described in prokaryotes (Browning et al. 2010; Dillon and Dorman 2010; Rimsky and Travers 2011; Wang et al. 2011). These proteins, termed nucleoid-associated proteins (NAPs), alter the degree of compaction, looping, and DNA supercoiling of bacterial chromosomes through interactions that bend, wrap, or bridge DNA (Dillon and Dorman 2010). Through such interactions, NAPs can repress or activate the transcription of a substantial number of genes. Importantly, as more NAPs have been characterized, the distinction between proteins that modulate DNA structure and proteins that regulate transcription has become blurred (Dillon and Dorman 2010).

Most recently, the EspR DNA-binding protein in MTB has been described that emphasizes the ambiguity between the concepts of a classical TF and a NAP. EspR was first described as a regulator of the espACD operon in MTB (Hunt et al. 2012),

which is a component of the ESX-1 secretion system required for virulence (Pym et al. 2003). The espACD operon has a transcription start site at −67 and is activated by EspR binding at the promoter. However, maximal transcription is achieved when EspR binds to the *espA* activating region (EAR) located between −1004 and −884. Moreover, a deletion bringing the EAR region as close as ∼200 nucleotides to the promoter abolishes its function (Hunt et al. 2012).

The EspR protein contains a helix-turn-helix DNA binding domain typical of many TFs and also a C-terminal domain that has been shown to mediate dimerization (Rosenberg et al. 2011). Structural studies have led to a model whereby EspR acts as a dimer of dimers in which each HTH domain in a dimer can bind to distantly separated binding sites (Blasco et al. 2011). This model was corroborated by atomic force microscopy which revealed DNA binding and DNA loop formation in conjunction with EspR binding. DNA loop formation, in turn, provided the likely mechanism for the ability of EspR binding at the EAR to activate the espACD operon at a distance. These data also led to the proposition that EspR acts in a manner more characteristic of an NAP.

The hypothesis that EspR acts as an NAP was further boosted by the results of ChIP-Seq mapping (Blasco et al. 2012). The data from this experiment revealed the EspR binds to over 165 loci in the MTB genome, and was distributed nearly equally between intergenic and genic regions. Moreover, re-analysis of the EspR binding data reveals a substantial overlap with the binding sites for Lsr2, a known nucleoid associate protein in MTB (Colangeli et al. 2007, 2009; Gordon et al. 2010). Lsr2 is known to also play a role the regulation of the espACD operon. Considered in the context of the canonical model of transcriptional regulation, these data led to the hypothesis that EspR does not behave as a traditional TF but instead regulates transcription globally as an NAP through long range interactions and DNA structure modifications.

A number of considerations, however, suggest that EspR may not be as unusual as suggested. As noted above, binding to genic and intergenic regions has emerged as the rule rather than the exception for TFs. Indeed, the binding profile of EspR in this regard is indistinguishable from the 50 other TFs that have been mapped in MTB by ChIP-Seq (Fig. 4). Moreover, as described above, long range interactions mediated by DNA looping has substantial precedent for other prokaryotic TFs. The proposed mechanism of a dimer of dimers, in particular, is similar into the mechanism described for the operation of the lac repressor. Finally, ChIP-Seq mapping revealed a specific DNA binding motif more consistent with traditional TFs rather than NAPs which tend to bind more nonspecific DNA sequences (Lsr2 binds general AT rich sequences in MTB, for example).

These considerations do not contradict the finding that EspR can mediate long range interactions through DNA looping mechanisms, and thus may play a role in global DNA structure. Rather, the data for EspR, considered in the context of the surprisingly diverse binding for many other TFs in both prokaryotes and eukaryotes, suggest that such behavior may not be unusual. If this were the case, much of the surprising diversity in binding location emerging from ChIP-Seq studies in

prokaryotes may be explained by a more general role of TFs in the modulation of DNA structure.

6 Extensive Weak Binding

A nearly universal finding of TF ChIP-Seq studies is the unexpectedly large number of weak binding sites that are found. Extensive weak binding has been observed for TFs in every eukaryotic organism in which ChIP-Seq mapping of TFs has been performed (Cao et al. 2010; Farnham 2009; Li et al. 2008; MacQuarrie et al. 2011; Rhee and Pugh 2011; Robertson et al. 2007b; Tanay 2006; Zeitlinger et al. 2007; Zhong et al. 2010). With the mapping of 50 MTB TFs, this observation has been extended to prokaryotic organisms (Galagan et al. Submitted), and indicates that extensive weak binding may be a general property of TFs. Owing in part to its ubiquity, the physiological significance of this weak binding is extensively debated.

The obvious concern associated with weak binding is that they may reflect artifacts of the ChIP-Seq procedure. There are several aspects to this concern. One is that the weak binding reflects random association of TFs with DNA locations that have no natural affinity for the TF. The second is that the experimental procedure is revealing DNA locations that do have an affinity for the TF, but that would not normally be bound under natural circumstances. Owing to such issues, it has been estimated that up to 30 % of binding sites identified by ChIP-Seq in eukaryotes may be false positives (Rhee and Pugh 2011).

Conversely, a number of considerations suggest that many, if not most, weak sites cannot be excluded as simple random binding. First, in cases where experimental procedures have been optimized, the reproducibility of weak binding is very high. In particular, ChIP-exo experiments on several human TFs have revealed even more weak binding than ChIP-Seq, but with high reproducibility. In MTB, comparisons of replicates of ChIP-Seq mapping for eight different TFs revealed very high reproducibility in both the positions and amplitudes of binding sites (Galagan et al. Submitted); the majority of binding sites were found with 50 bp of one another in replicates and the correlation coefficient for comparisons for binding site heights was typically greater than 0.9, even weak binding sites can be associated with an underlying motif that bears some, albeit degraded, relationship with the motif associated with the strongest binding sites for each TF (Rhee and Pugh 2011). And in studies in yeast using ChIP–ChIP, the strength of weak binding sites was correlated with the strength of the underlying motif (Tanay 2006). Third, the stereotypical shift between forward and reverse read coverage is identical in weak and strong peaks, indicating that weak peaks correspond to localized binding. Finally, in human studies using ChIP-exo where high spatial resolution was possible, the locations of weak peaks were found to be nonrandom and instead concentrated at fixed distances to genomic features (Rhee and Pugh 2011).

Assuming that at least some weak peaks represent sequence-specific binding sites with true affinity for the corresponding TF, the more contentious question is whether or not such binding has physiological significance. The common criticism is that much of the binding observed is too weak to be physiologically significant. And evidence for this is the finding that only a fraction of binding sites found in nearly all experiments can be assigned a function when the corresponding TF is perturbed. But this criticism also raises subtle biological and operational issues.

Biologically, the issue is how to define a threshold for biological relevance and whether such a threshold exists. The notion of a threshold for relevance implies a digital behavior for biological systems such that a factor or binding site does or does not have a regulatory effect (Rhee and Pugh 2011; Tanay 2006). Yet, cells are analog systems in which behavior and functional impact can exist along a continuum. The relevance of sites that exist at the high end of the impact continuum is easy to acknowledge, but sites with weaker impact may play a role in fine-tuning aspects of a regulatory response that are less obvious but, nonetheless, real. A continuum of effect renders the issue of a threshold for weak sites into an operational question.

Operationally, selecting a threshold requires defining a level of functional effect that is visible with the current methods used for perturbing and observing regulation. The most common method has involved genetically knocking out a TF of interest and then comparing the expression of genes in the knockout versus their expression in wild-type cells using microarrays. But several issues are associated with this approach. First, determining which gene is regulated by a given binding site is a difficult problem. This is certainly true for eukaryotes where long range interactions are common and the regulated gene may not be the closest gene to the binding site. But as described above, this is also an issue for prokaryotes. Second, the signal-to-noise ratio of microarrays is not unlimited. Thresholds for microarrays are set by statistical considerations that in part reflect the sensitivity of the array technology. A common criterion is to select only genes that display a greater than 2-fold change in expression, yet, there is no reason to believe that all biologically meaningful changes must be greater than 2-fold. More generally, it is unlikely that any instrumentation-based threshold would happen to match the biological thresholds for physiological relevance for all TFs, even if such biological thresholds existed. Third, TFs do not always act in isolation, but may operate in combination with other factors to regulate a particular gene. In such instances, perturbing any individual TF may not be sufficient to effect the expression of a jointly regulated gene. Finally, the standard knockout experiment only assays for regulatory effects long after the perturbation. Ignoring the possibility for artificial compensation with the regulatory network, this approach is also limited in that it only assays for effects that impact equilibrium expression levels. This ignores regulatory effects that may play a role in expression dynamics, and in particular the well-known role of regulatory network motifs (Alon 2006, 2007; Kalir et al. 2001, 2005; Kashtan et al. 2004; Mangan and Alon 2003; Mangan et al. 2003; Milo et al. 2002; Rosenfeld et al. 2002; Roy et al. 2010; Setty et al. 2003; Shen-Orr et al. 2002). Given these considerations, the inability to find a regulatory

effect of a particular binding site must be conditioned on the limitations of the method used to look for an effect.

In addition, growing positive evidence suggests that weak binding sites can mediate physiological effects if assayed and analyzed in an appropriate manner. One of the first global studies of this phenomenon was performed on ChIP–ChIP data from yeast (Tanay 2006). Consistent with ChIP-Seq data in other organisms, ChIP–ChIP in yeast revealed that weak binding sites likely represented the majority of binding events for most TFs. The substantial noise associated with ChIP–ChIP as compared to ChIP-Seq necessitates caution in the analyses of these data. Yet despite this, a clear relationship was observed between the predicted binding energy of promoters for different TFs and the regulatory effect of perturbing that TF. Most notably, this trend was observed even for promoters whose effect fell below standard significance thresholds. In other words, a clear trend existed relating the strength of promoters to the strength of a regulatory effect, and the statistical threshold applied to each site individually masked this trend and the weaker effects.

In addition, substantial evidence exists that weak sites may play a significant role in modulating the effects of other binding sites through cooperative interactions. As described above, numerous examples exist in prokaryotes that support the role of weak binding sites in mediating long range cooperative interactions, and the role of long range interactions in eukaryotes is well known. But cooperative effects can also modulate the impact of weak binding sites in canonical promoter regions as well. One elegant demonstration of this was based on the analysis of artificial promoter sequences in yeast (Gertz et al. 2009). As part of this study, synthetic promoter sequences coupled to fluorescent promoters were used to test the impact of different combinations of binding sites for different TFs. The interactions between a weak and strong binding site for Mig1, a TF not previously known to participate in cooperative binding, were studied in particular detail. The weak site selected was known to have low affinity for Mig1 and was shown to have a weak regulatory effect when present alone in a promoter sequence. But promoters containing the weak and strong binding site had an equivalent regulatory effect as promoters containing two strong sites, which was stronger than the impact of one strong site by itself. Thus, in this system, the weak binding site displayed strong cooperativity with a strong binding site in the same promoter region.

These and other data suggest that one role of weak binding may be to modulate the overall affinity of a promoter sequence for a transcription factor. Through cooperative interactions, the magnitude of the occupancy of the promoter as a whole can be increased beyond that which would be seen with any individual site. And, as noted above for the lac operon, cooperative interactions can also sculpt the kinetics of binding with more flexibility than is possible with a single site. Moreover, if cooperative interactions are a significant functional consequence of weak binding, then clustering of weak and strong binding sites would be expected, and this is indeed a common observation in both eukaryotes (Rhee and Pugh 2011) and MTB.

Fig. 5 MTB TF binding data are available at Tuberculosis Database (TBDB). Binding data for 50 TFs generated by the NIAID funded TB systems biology project have been integrated the genome sequence and annotation of MTB and released at TBDB.org. Selected screen shots show online tools available for searching, browsing, and downloading these data

7 MTB Binding Data Available at TBDB

The diversity of binding sites observed for MTB TFs suggests that bacterial regulation is likely more complicated than expected from the simplified canonical model. Yet, much work remains to validate the potential functional implications of these data. Toward the end of enabling research into these questions, ChIP-Seq data for 50 MTB TFs generated by the NIAID TB systems biology project has been released through the TBDB.org (Fig. 5).

TBDB is an online database providing integrated access to genome sequence, expression data, literature curation and systems biology data for MTB and related genomes (Galagan et al. 2010). TBDB currently houses genome assemblies for numerous strains of MTB as well assemblies for over 20 strains related to MTB and useful for comparative analysis. It also houses re-sequencing data for over 31 different MTB strains selected as part of the *M. tuberculosis* Phylogeographic Diversity Sequencing Project. These data provide a global view of the genomic diversity of MTB at the level of SNPs and indels. TBDB stores pre- and post-publication gene-expression data from *M. tuberculosis* and its close relatives, including over 3000 MTB microarrays, 95 RT-PCR datasets, 2700 microarrays for human and mouse TB-related experiments, and 260 arrays for Streptomyces coelicolor. In addition, metabolic reconstructions have been performed on all organisms in the site and these models are hosted as Biocyc Pathway/Genome databases (http://biocyc.org/) in TBDB. To enable wide use of these data, TBDB provides a suite of tools for searching, browsing, analyzing, and downloading the data.

TBDB also provides a growing set of tools for utilizing the ChIP-Seq data generated by the NIAID TB systems biology project (Fig. 5). Through TBDB, users can search for regulatory binding sites by regulator, by target, or by genomic

coordinate. Users can also browse a regulatory network constructed from these data. From the results of any of these searches, users may view the regulatory network for the gene of interest, select, and view raw ChIP-Seq data in the dynamic real-time genome browser GenomeView (Abeel et al. 2012), view the summary page for each experiment, or view static images of the ChIP-Seq peak data. Users may browse experiments directly and view the entire genome for each experiment. Users can also download all raw data (Fig. 5).

The data currently available through TBDB.org represent the first release of ChIP-Seq mapping data for the NIAID funded TB systems biology project. The ultimate goal of the project is to map all $\sim$200 predicted DNA binding proteins in the MTB genome. As these additional proteins are mapped, data will continue to be released through this site.

References

Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y (2012) Genomeview: a next-generation genome browser. Nucleic Acids Res 40:e12

Alon U (2006) An introduction to systems biology: design principles of biological circuits (Chapman and Hall/CRC)

Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev 8:450–461

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37: W202–W208

Bekaert M, Firth AE, Zhang Y, Gladyshev VN, Atkins JF, Baranov PV (2010) Recode-2: new design, new search tools, and many more genes. Nucleic Acids Res 38:D69–D74

Belitsky BR, Sonenshein AL (1999) An enhancer element located downstream of the major glutamate dehydrogenase gene of Bacillus subtilis. Proc Nat Acad Sci U S A 96:10290–10295

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc : Ser. B (Methodol) 57:289

Blasco B, Stenta M, Alonso-Sarduy L, Dietler G, Peraro MD, Cole ST, Pojer F (2011) Atypical DNA recognition mechanism used by the EspR virulence regulator of Mycobacterium tuberculosis. Mol Microbiol 82:251–264

Blasco B, Chen JM, Hartkoorn R, Sala C, Uplekar S, Rougemont J, Pojer F, Cole ST (2012) Virulence regulator EspR of mycobacterium tuberculosis is a nucleoid-associated protein. PLoS Pathog 8:e1002621

Browning DF, Busby SJ (2004) The regulation of bacterial transcription initiation. Nat Rev Microbiol 2:57–65

Browning DF, Grainger DC, Busby SJ (2010) Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. Curr Opin Microbiol 13:773–780

Cao Y, Yao Z, Sarkar D, Lawrence M, Sanchez GJ, Parker MH, MacQuarrie KL, Davison J, Morgan MT, Ruzzo WL et al (2010) Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev Cell 18:662–674

Chen X, Guo L, Fan Z, Jiang T (2008) W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ ChIP–chip data. Bioinformatics 24:1121–1128

Colangeli R, Helb D, Vilcheze C, Hazbon MH, Lee CG, Safi H, Sayers B, Sardone I, Jones MB, Fleischmann RD et al (2007) Transcriptional regulation of multi-drug tolerance and antibiotic-induced responses by the histone-like protein Lsr2 in M. tuberculosis. PLoS Pathog 3:e87

Colangeli R, Haq A, Arcus VL, Summers E, Magliozzo RS, McBride A, Mitra AK, Radjainia M, Khajo A, Jacobs WR Jr et al (2009) The multifunctional histone-like protein Lsr2 protects mycobacteria against reactive oxygen intermediates. Proc Natl Acad Sci U S A 106:4414–4418

Collado-Vides J, Magasanik B, Gralla JD (1991) Control site location and transcriptional regulation in Escherichia coli. Microbiol Rev 55:371–394

Czaplewski LG, North AK, Smith MC, Baumberg S, Stockley PG (1992) Purification and initial characterization of AhrC: the regulator of arginine metabolism genes in Bacillus subtilis. Mol Microbiol 6:267–275

Dandanell G, Valentin-Hansen P, Larsen JE, Hammer K (1987) Long-range cooperativity between gene regulatory sequences in a prokaryote. Nature 325:823–826

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27:4636–4641

Dillon SC, Dorman CJ (2010) Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Microbiol 8:185–195

Dunn TM, Hahn S, Ogden S, Schleif RF (1984) An operator at -280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. Proc Nat Acad Sci U S A 81:5017–5020

Farnham PJ (2009) Insights from genomic profiling of transcription factors. Nat Rev 10:605–616

Flashner Y, Gralla JD (1988) Dual mechanism of repression at a distance in the lac operon. Proc Nat Acad Sci U S A 85:8968–8972

Froula JL, Francino MP (2007) Selection against spurious promoter motifs correlates with translational efficiency across bacteria. PLoS ONE 2:e745

Galagan JE, Sisk P, Stolte C, Weiner B, Koehrsen M, Wymore F, Reddy TB, Zucker JD, Engels R, Gellesch M et al (2010) TB database 2010: overview and update. Tuberculosis (Edinb) 90: 225–235

Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muniz-Rascado L, Solano-Lira H, Jimenez-Jacinto V, Weiss V, Garcia-Sotelo JS, Lopez-Fuentes A et al (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (gensor units). Nucleic Acids Res 39:D98–105

Gao CH, Yang M, He ZG (2011) An ArsR-like transcriptional factor recognizes a conserved sequence motif and positively regulates the expression of phoP in mycobacteria. Biochem Biophys Res Commun 411:726–731

Gertz J, Siggia ED, Cohen BA (2009) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature 457:215–218

Gilbert W, Muller-Hill B (1966) Isolation of the lac repressor. Proc Nat Acad Sci U S A 56: 1891–1898

Gomes A et al Decoding ChIPseq with multiple binding events provides site detection with high-resolution and allows estimation of cooperative binding (In Preparation)

Gordon BR, Li Y, Wang L, Sintsova A, van Bakel H, Tian S, Navarre WW, Xia B, Liu J (2010) Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 107:5154–5159

Hahn S, Hendrickson W, Schleif R (1986) Transcription of escherichia coli ara in vitro. The cyclic AMP receptor protein requirement for PBAD induction that depends on the presence and orientation of the araO2 site. J Mol Biol 188:355–367

Huerta AM, Francino MP, Morett E, Collado-Vides J (2006) Selection for unequal densities of sigma70 promoter-like signals in different regions of large bacterial genomes. PLoS Genet 2:e185

Hunt DM, Sweeney NP, Mori L, Whalan RH, Comas I, Norman L, Cortes T, Arnvig KB, Davis EO, Stapleton MR et al (2012) Long-range transcriptional control of an operon necessary for virulence-critical ESX-1 secretion in Mycobacterium tuberculosis. J Bacteriol 194:2307–2320

Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316:1497–1502

Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, Leibler S, Surette MG, Alon U (2001) Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. Science 292:2080–2083

Kalir S, Mangan S, Alon U (2005) A coherent feed-forward loop with a sum input function prolongs flagella expression in Escherichia coli. Mol Syst Biol 1(2005):0006

Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Topological generalizations of network motifs. Phys Rev E: Stat, Nonlin, Soft Matter Phys 70:031909

Kim J, Chu J, Shen X, Wang J, Orkin SH (2008) An extended transcriptional network for pluripotency of embryonic stem cells. Cell 132:1049–1061

Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, Schmid AK, Pan M, Marzolf B, Van PT, Lo FY et al (2009) Prevalence of transcription promoters within archaeal operons and coding sequences. Mol Syst Biol 5:285

Lee DH, Schleif RF (1989) In vivo DNA loops in araCBAD: size limits and helical repeat. Proc Nat Acad Sci U S A 86:476–480

Levin A et al (2011) Understanding Blind Deconvolution Algorithms. IEEE transactions on pattern analysis and machine intelligence

Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL et al (2008) Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. PLoS Biol 6:e27

Lobell RB, Schleif RF (1990) DNA looping and unlooping by AraC protein. Science 250:528–532

Lobell RB, Schleif RF (1991) AraC-DNA looping: orientation and distance-dependent loop breaking by the cyclic AMP receptor protein. J Mol Biol 218:45–54

Lun DS, Sherrid A, Weiner B, Sherman DR, Galagan JE (2009) A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. Genome Biol 10:R142

Machanick P, Bailey TL (2011) MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics 27:1696–1697

MacQuarrie KL, Fong AP, Morse RH, Tapscott SJ (2011) Genome-wide transcription factor binding: beyond direct target regulation. Trends Genet 27:141–148

Madan Babu M, Teichmann SA (2003) Functional determinants of transcription factors in escherichia coli: protein families and binding sites. Trends Genet 19:75–79

Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. Proc Nat Acad Sci U S A 100:11980–11985

Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol 334:197–204

Martin K, Huo L, Schleif RF (1986) The DNA loop model for ara repression: AraC protein occupies the proposed loop sites in vivo and repression-negative mutations lie in these same sites. Proc Nat Acad Sci U S A 83:3654–3658

Mazzoni EO, Mahony S, Iacovino M, Morrison CA, Mountoufaris G, Closser M, Whyte WA, Young RA, Kyba M, Gifford DK et al (2011) Embryonic stem cell-based mapping of developmental transcriptional programs. Nat Methods 8:1056–1058

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP et al (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827

Minchin SD, Austin S, Dixon RA (1988) The role of activator binding sites in transcriptional control of the divergently transcribed nifF and nifLA promoters from Klebsiella pneumoniae. Mol Microbiol 2:433–442

Mullin DA, Newton A (1993) A sigma 54 promoter and downstream sequence elements ftr2 and ftr3 are required for regulated expression of divergent transcription units flaN and flbG in Caulobacter crescentus. J Bacteriol 175:2067–2076

Narang A (2007) Effect of DNA looping on the induction kinetics of the lac operon. J Theor Biol 247:695–712
Ninfa AJ, Reitzer LJ, Magasanik B (1987) Initiation of transcription at the bacterial glnAp2 promoter by purified E. coli components is facilitated by enhancers. Cell 50:1039–1046
Oehler S, Eismann ER, Kramer H, Muller-Hill B (1990) The three operators of the lac operon cooperate in repression. EMBO J 9:973–979
Oehler S, Alberti S, Muller-Hill B (2006) Induction of the lac promoter in the absence of DNA loops and the stoichiometry of induction. Nucleic Acids Res 34:606–612
Oppenheim AV, Willsky AS, Nawab SH (1997) Signals & systems, 2nd edn, Upper Saddle River, NJ: Prentice Hall
Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680
Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6:S22–S32
Pym AS, Brodin P, Majlessi L, Brosch R, Demangel C, Williams A, Griffiths KE, Marchal G, Leclerc C, Cole ST (2003) Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. Nat Med 9:533–539
Reitzer LJ, Magasanik B (1986) Transcription of glnA in E. coli is stimulated by activator bound to sites far from the promoter. Cell 45:785–792
Reznikoff WS, Winter RB, Hurley CK (1974) The location of the repressor binding sites in the lac operon. Proc Natl Acad Sci U S A 71:2314–2318
Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147:1408–1419
Rimsky S, Travers A (2011) Pervasive regulation of nucleoid structure and function by nucleoid-associated proteins. Curr Opin Microbiol 14:136–141
Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A et al (2007a) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Meth 4:651–657
Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A et al (2007b) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657
Rosenberg OS, Dovey C, Tempesta M, Robbins RA, Finer-Moore JS, Stroud RM, Cox JS (2011) EspR, a key regulator of Mycobacterium tuberculosis virulence, adopts a unique dimeric structure among helix-turn-helix proteins. Proc Natl Acad Sci U S A 108:13450–13455
Rosenfeld N, Elowitz MB, Alon U (2002) Negative autoregulation speeds the response times of transcription networks. J Mol Biol 323:785–793
Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF et al (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330:1787–1797
Schleif R (2003) AraC protein: a love-hate relationship. BioEssays : news and reviews in molecular, cellular and developmental biology 25:274–282
Setty Y, Mayo AE, Surette MG, Alon U (2003) Detailed map of a cis-regulatory input function. Proc Nat Acad Sci U S A 100:7702–7707
Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 31:64–68
Storey JD (2002) A direct approach to false discovery rates. J Roy Stat Soc B 64:479–498
Storey JD (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Stat 31:2013–2035
Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Nat Acad Sci U S A 100:9440–9445
Tanay A (2006) Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res 16:962–972
Ueno-Nishio S, Backman KC, Magasanik B (1983) Regulation at the glnL-operator-promoter of the complex glnALG operon of Escherichia coli. J Bacteriol 153:1247–1251

Ueno-Nishio S, Mango S, Reitzer LJ, Magasanik B (1984) Identification and regulation of the glnL operator-promoter of the complex glnALG operon of Escherichia coli. J Bacteriol 160:379–384

Wang W, Li GW, Chen C, Xie XS, Zhuang X (2011) Chromosome organization by a nucleoid-associated protein in live bacteria. Science 333:1445–1449

Wedel A, Weiss DS, Popham D, Droge P, Kustu S (1990) A bacterial enhancer functions to tether a transcriptional activator near a promoter. Science 248:486–490

Wilbanks EG, Facciotti MT (2010) Evaluation of algorithm performance in ChIP-seq peak detection. PLoS ONE 5:e11471

World Health Organization (2001). Global Tuberculosis Control

Galagan J et al Reconstruction of the *Mycobacterium tuberculosis* regulatory network and deconstruction of the hypoxic response. Nature (Submitted)

Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, Young RA, Levine M (2007) Whole-genome ChIP–chip analysis of dorsal, twist, and snail suggests integration of diverse patterning processes in the drosophila embryo. Genes Dev 21:385–390

Zhong M, Niu W, Lu ZJ, Sarov M, Murray JI, Janette J, Raha D, Sheaffer KL, Lam HY, Preston E et al (2010) Genome-wide identification of binding sites defines distinct functions for caenorhabditis elegans PHA-4/FOXA in development and environmental response. PLoS Genet 6:e1000848

The Role and Contributions of Systems Biology to the Non-Human Primate Model of Influenza Pathogenesis and Vaccinology

Carole Baskin

Abstract Nonhuman primates have proven to be valuable models in the study of seasonal and highly pathogenic influenza virus infections, prophylaxis, and therapy. Due to their close genetic relationship to humans, these animals share anatomic, postural, physiological, and immune features with us of key importance when it comes to progression and mitigation of respiratory infections. Their lower susceptibility to natural influenza infection even presents an advantage in the laboratory setting because of the need for immunologically na animals, and since nonhuman primates are relatively genetically diverse within one species, their study provides an essential complement to the body of knowledge acquired with inbred animal models. However, ethical and cost considerations typically result in smaller experiments and a need to look at additional levels of biological information in order to maximize insights gained from these studies. Systems biology is a powerful tool for this purpose, because it provides a much needed wide angle view of complex interactions taking places in organisms which are more than the sum of their parts. This chapter will describe the extent to which functional genomics and proteomics have successfully integrated with other, more traditional tools in the areas of clinical presentation, pathology, and immunology during influenza infections in nonhuman primates. It will also describe the unique contributions systems biology has made to our understanding of host–virus interactions, as well as response to vaccination and antiviral therapy.

C. Baskin (✉)
Science Foundation Arizona, 400 East Van Buren Street,
Phoenix, AZ 85004, USA
e-mail: cb2@u.washington.edu

Current Topics in Microbiology and Immunology (2013) 363: 69–85
DOI: 10.1007/82_2012_248

Published Online: 14 August 2012

Contents

1 Introduction: How Systems Biology Helps us See the Covert but Far-Reaching Effects of One Infected Cell

Medicine is undergoing a revolution that will take shape over the next several decades: an individual's medical record will no longer consist of a few data points mostly collected during disease events, but of millions of pieces of information accumulated over long periods of time while the person is healthy. So not only will the focus of medicine shift from populations to individuals, a substantial portion of the new information gained will be about health rather than disease. This will enable us to understand the delicate balance of ongoing activities that collectively result in the absence of disease for any one individual and avoid rampant mis- and overdiagnosis. We will also learn to detect the subtle pre-symptomatic signs of impending disease against the complex background noise of healthy cellular processes earlier than ever before, heralding a new era in preventative and predictive medicine, and widening the window of opportunity for therapeutic intervention. Indeed, I believe that a single event, such as the infection of upper respiratory cells by influenza virions, has a detectable and unique signature, locally and systemically, well before detection by conventional diagnostics and before a critical mass of lung tissue is injured. Since neither the effects of seasonal strains nor those of more virulent strains, such as the 1918 pandemic and H5N1 viruses can be solely explained by the damage that occurs in cells where these viruses replicate, it is critical to look at early host–wide effects of infection to understand and successfully mitigate the course of the disease.

Our efforts, over the past few years, have shown that the information is there for the taking and that we can start making sense of it. In studying infection of nonhuman primates with seasonal and highly pathogenic influenza viruses, we have gained clinically relevant insights about interactions between host and pathogens. Some of these interactions were common to different viruses, while others were unique or indicative—either by their nature or timing—of virulence,

and thus helped refined our understanding of pathogenicity. We observed that even mild infections had far-reaching effects pre-symptomatically, which could be detected in noninfected tissues, including peripheral blood cells (Baas et al. 2006; Tolnay et al. 2010). When we used proteomic in addition to genomic analyses, we found a high level of concordance early in infection between gene expression and translation, and less so as disease progressed. This suggests complex interplays between these two processes that may provide valuable insights on how cells and organisms attempt to maintain homeostasis during stressful events by using subtle, subclinical protective mechanisms (Baskin and Katze 2008).

Mice, ferrets, guinea pigs, cotton rats, hamsters, and macaques have all been used to study influenza viruses as well as the tools to prevent or combat infections. Each model presents unique advantages and disadvantages (Bouvier and Lowen 2010). We chose to conduct experiments on nonhuman primates, in addition to mice, for a variety of reasons: since we studied gene expression in response to infection, we wanted a model as close as possible to humans genetically and at the same time exhibiting a large degree of genetic diversity within one species as humans do, something that commonly used laboratory mice would not provide. We wanted a model that could be infected with unmodified low pathogenicity influenza viruses, unlike mice, but not so easily that it would be difficult to find seronegative animals, as is the case with ferrets. While signs of influenza infection are outwardly more severe in humans, they are essentially identical, save for the lack of sneezing in nonhuman primates. Finally, cell tropism of influenza viruses during in vivo infections is similar in humans and macaques, including for highly pathogenic avian influenza viruses (HPAIV) (Baskin et al. 2009; Chen et al. 2009; Shinya et al. 2012). Other common features, such as the size and orientation of the respiratory tract during waking activity, were judged likely to affect the speed and nature of infectious spread, and therefore tipped the scale in favor of studying the nonhuman primate model.

2 How ‘Omics’ Can Assist as Early Prognostic Tools

In humans, uncomplicated influenza is characterized by an acute onset of symptoms including fever from 100 °F to as high as 106 °F, chills, head and muscle aches, lethargy, anorexia, coughing, and rhinorrhea. While systemic involvement typically subsides a few days before respiratory symptoms improve, it is the former that clinically differentiates influenza from other viral upper respiratory tract infections, such as the common cold. Pulmonary complications of seasonal influenza include primary viral pneumonia, with potential appearance of dyspnea and hypoxemia, which can be followed by a secondary bacterial pneumonia, characterized by recrudescence and worsening of systemic manifestations (Bouvier and Lowen 2010). Viral shedding typically starts the day of infection, peaks 2 days later, yet symptoms may not appear for another day or so. Consequently, a window of time exists during which a person is contagious without

being aware s/he is infected and occasionally, infections are completely asymptomatic, although viral shedding still takes place. This characteristic of influenza infections justifies the need for sensitive and accurate pre-symptomatic diagnostics, so that individuals who may have been exposed to an emergent influenza virus can be quarantined and the appropriate treatment implemented without delays.

Nonhuman primates are susceptible to infection with a number of unadapted human influenza A isolates, including viruses of the H1N1 subtype (Bouvier and Lowen 2010). This has been demonstrated by sampling sera of macaques and other species in areas where they are kept as pets or where human and nonhuman primate populations cohabit. Results of the study showed that the animals had been infected with a number of common human infectious illnesses. Additionally, epidemics of these diseases, including influenza, occur commonly in captivity when protective measures are lacking (Jones-Engel et al. 2001).

It was, therefore, no surprise when macaques experimentally infected with a seasonal strain of influenza developed signs consistent with the human disease, including listlessness, anorexia, and nasal drip (Baas et al. 2006; Baskin et al. 2004, 2007). Most interestingly, we observed that there was co-upregulation of a number of interferon-induced or interferon signaling genes in lung tissue and peripheral immune blood cells as early as 2 days after infection (Fig. 1). We found that many of these genes were also translated to proteins when we performed a proteomic analysis on lung tissue. These results were remarkable in view of the low pathogenicity of the virus, the absence of viremia, and the relatively minor nature of the lesions present in the lungs. Yet, the transcriptional signal in peripheral blood was clearly discernable amongst the more 'routine' traffic of immune cells. When macaques were infected with highly pathogenic viruses, such as the H5N1 HPIA, 1918 pandemic virus, or recombinant versions of the latter (Baskin et al. 2009; Cillóniz et al. 2009; Kobasa et al. 2007; Rimmelzwaan et al. 2001), the animals quickly developed the rapidly progressing lung disease that characterizes human infections with these viruses. Examples of signs observed included dramatically increased respiratory rates and a decrease by as much as 36 % in lung oxygenation, as measured by pulse oximetry, all indications of severely impaired lung function. As with the seasonal virus, the presence of several key cytokines and chemokines in serum coincided early on with high expression in lung tissue.

In a different study looking at the effect of vaccination with a novel, replication-deficient live influenza vaccine, we collected bronchial epithelial cells through a bronchoscope and performed gene expression analysis. Two days after vaccination with the modified live vaccine, we observed strong up-regulation of interferon-induced genes in these cells (Fig. 2). By the time the animals were challenged with a homologous virus (day 21), gene expression had returned to baseline. However, 2 days after the challenge (day 23), the same set of genes was induced again, although less so than in unvaccinated animals or animals which had received a killed vaccine. The vaccination and inoculation process only briefly and minimally exposed bronchial epithelial cells to the replication-deficient virus (the killed vaccine was administered intra-muscularly) or the challenge virus, as evidenced by

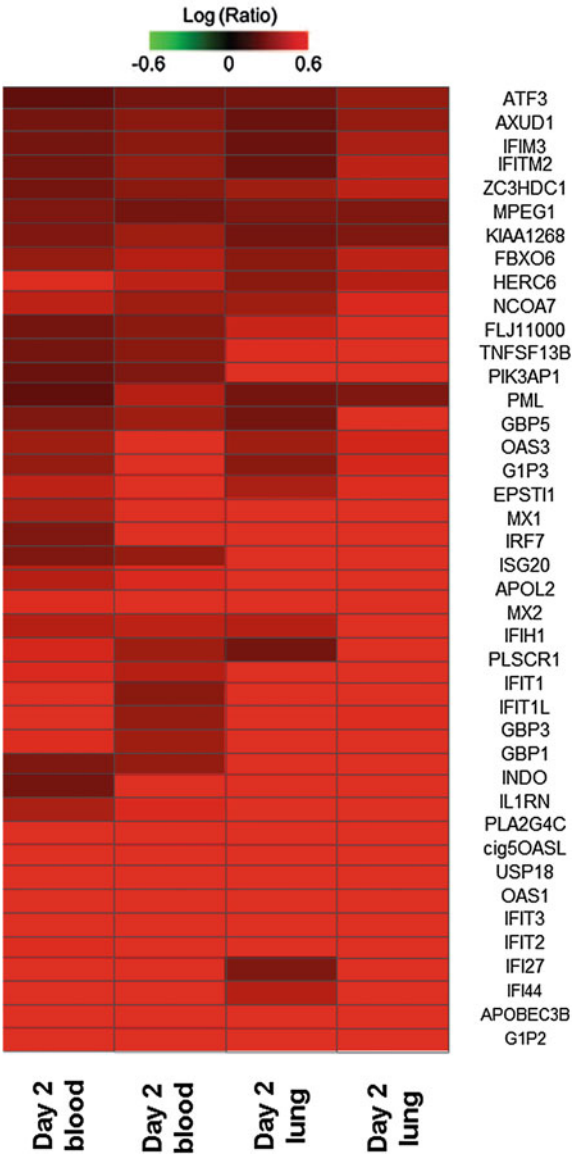


Fig. 1 Two days after infection with seasonal influenza, there is co-upregulation of interferon-regulated and antiviral genes in peripheral blood cells and affected lung tissue of two macaques used in this experiment. Modified from Baas et al. (2006)

the lack of viral mRNA detected in these cells after inoculation and/or infection. Yet, the cells exhibited striking transcriptional induction soon after these events. This response was prescient of the strong protection elicited by the modified live vaccine: post-infection pathology in the lungs was minor in nature and extent and this finding was reflected the differences observed in gene expression and serum antibody levels (Baskin et al. 2007).

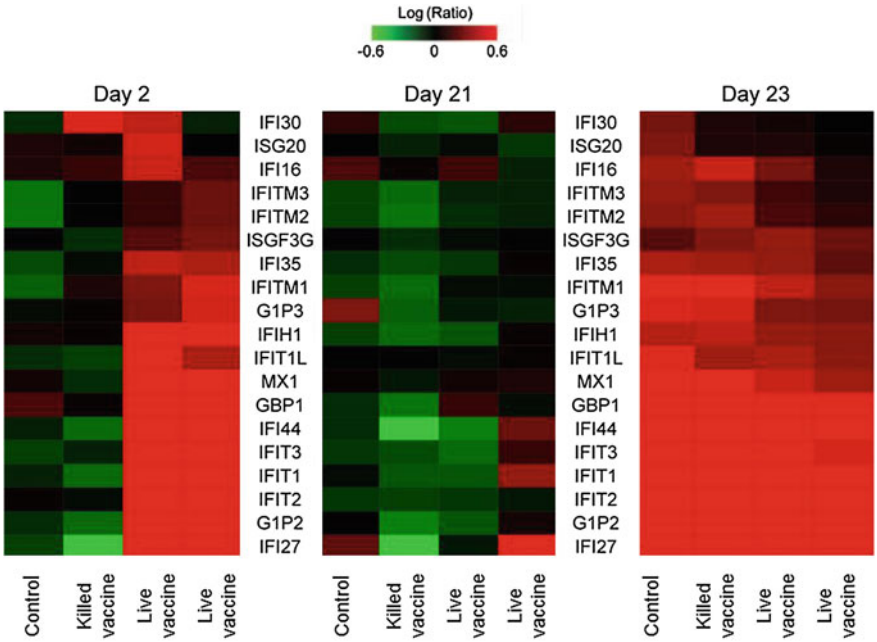


Fig. 2 Gene expression analysis on bronchial epithelial cells shows up-regulation of interferon-induced genes 2 days after vaccinating with a replication-deficient live influenza vaccine and 2 days (day 23 of protocol) after challenge with a homologous influenza virus. Adapted from Baskin et al. (2007)

This experiment raised again the possibility of looking to gene expression in clinical samples for predicting the course of infectious diseases. Although a common gene expression response to different causes of acute lung inflammation was identified in rodents and nonhuman primates (Pennings et al. 2008), dedicated responses were found as well, which were specific to pathogen types. These results were consistent with another study comparing gene expression in peripheral mononuclear cells of human patients, which found gene expression profiles to have diagnostic value (Ramilo et al. 2007). One should remember, however, that patient responses to an emergent influenza virus vary from asymptomatic to deadly, seemingly in an unpredictable manner, supporting the argument that we need to focus less on viral identity or load to determine prognosis and more at individual host responses to understand the pathogenesis and predict the course of these infections. The best example was the 1918 pandemic influenza virus, which killed over 50 millions individuals, with previously healthy young adults exhibiting the worst outcomes (Loo and Gale 2007). Since a main feature of the 1918 strain infections was dysregulation of the immune response, individuals with the most robust immune systems were, therefore, the most vulnerable. This type of divergent pattern cannot be anticipated from the epidemiology of seasonal influenza viruses, and it is unknown whether it will repeat itself in the event of an H5N1

pandemic because the population affected is still too small currently to make this type of prediction, despite similarities in the pathology and clinical course with the 1918 virus. It is now possible to use real-time PCR at points of care to diagnose influenza infections, at least by subtype, so it seems that the true value of gene expression analysis in blood or other clinical samples is prognostic, rather than diagnostic. It may allow medical personnel to provide care tailored to the developing host response almost immediately after exposure, for an improved outcome and better management of existing resources.

3 The Immune Response to Low and High-Path Influenza Viruses: Role of Systems Biology in the Study of Very Early yet Critical Differences in Host Response

Respiratory epithelial cells and leukocytes infected with influenza virus A respond to the insult in part by secreting interferon, whose production starts within a couple of hours and activates antiviral defenses in neighboring cells, limiting viral spread, and eliminating a large proportion of the original viral load. Concurrent release of cytokines and chemokines serve to attract innate immune cells to infected tissues. Consistently, genes with functions relevant to neutrophils and macrophages were significantly up regulated in the lungs of macaques 4 days after infection with seasonal influenza (Baskin et al. 2004). Dendritic cells play a central role in the subsequent development of an adaptive immune response, because they are directly stimulated by interferon and migrate from tissue to local lymph nodes to present antigens to T cells. In the same study, genes with functions relevant to dendritic cells were highly activated at day 4, which is the general time frame for initiation of the adaptive response. In macaques exposed to an NS1 truncated influenza virus used as a modified live vaccine, there was strong evidence of early dendritic cell maturation, which could not be explained solely by the live nature of the vaccine since it replicated poorly. Instead, it is likely that the truncation of NS1 resulted in poor suppression of interferon gene transcription by the virus (Kochs et al. 2007).

In support of this hypothesis, we observed the strong transcriptional induction of interferon induced and related genes shortly after inoculation with the experimental vaccine and this was quickly followed by a 7-fold increase in percentage of virus-specific CD4+ T cell circulating in peripheral blood, immunoglobulin G production, and transcriptional induction of T- and B cell pathways in lung tissue (Baskin et al. 2007). After challenge with a homologous virus, these animals exhibited lower transcriptional induction of interferon and inflammatory pathways in bronchial cells and lungs, compared to animals which had not been vaccinated or received a killed vaccine, suggesting that the ability of a host to respond to infection with robust virus-specific immunity directly helps decrease interferon production and inflammation. During H5N1 HPAI infections in macaques,

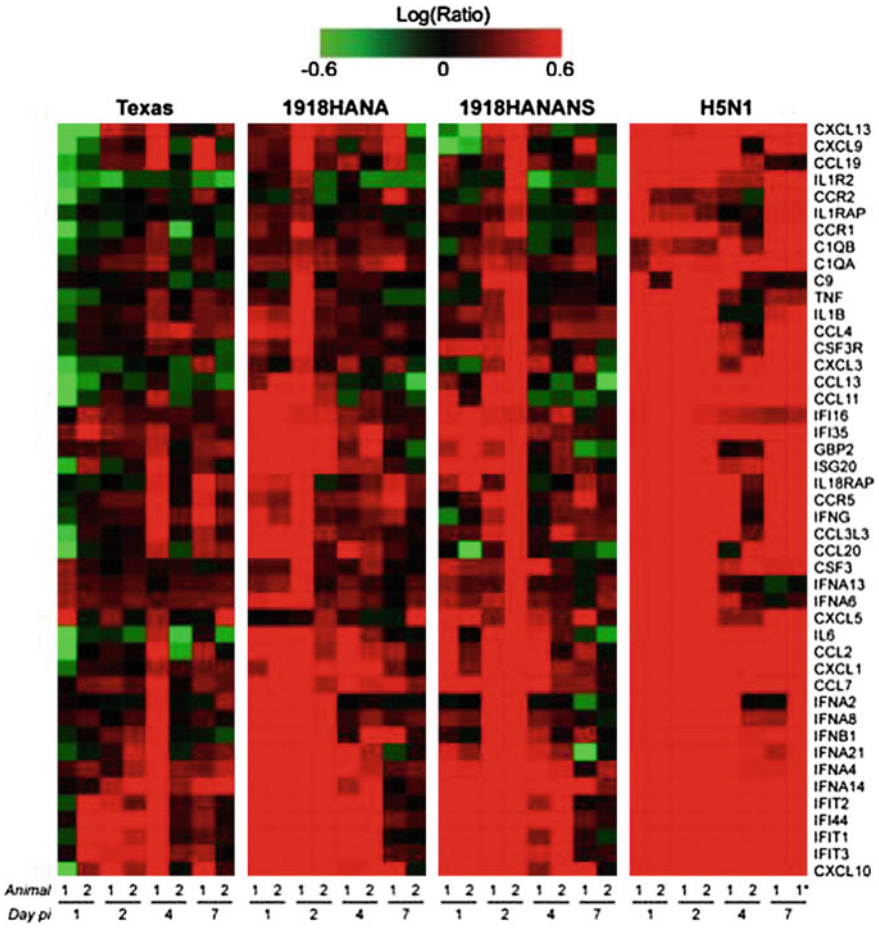


Fig. 3 Gene expression analysis in lungs of macaques infected with either H5N1 HPAI, seasonal influenza recombined with either two or three genes from the 1918 pandemic virus, or the seasonal virus as a control, demonstrates strong and protracted transcriptional induction of interferon-induced and other inflammatory genes in the H5N1 infected animals. Adapted from Baskin et al. (2009)

prolonged induction of the innate immune response was apt to be primarily caused by vigorous viral replication taking place throughout the 7 days of the study, targeting type II pneumocytes which are prone to secreting large amount of cytokines, in addition to interferons. However, the en masse apoptosis of dendritic cells, discovered by pathology examination of lungs and tracheobronchial lymph nodes, may also have contributed to this phenomenon by impairing antigen presentation, and therefore optimal development of an adaptive response (Fig. 3).

Another macaque study, which looked at gene expression in lungs after infection with the fully reconstructed 1918 pandemic virus, also showed a protracted innate

immune response (Kobasa et al. 2007), and a follow up experiment comparing this virus with H5N1 HPAI determined that even within hours of infection, highly pathogenic viruses induced interferon, chemokines, and cytokine pathways in a similar manner (Cillóniz et al. 2009). An extensive proteomic analysis performed on lungs of macaques infected with H5N1 confirmed that many of these proteins were not only transcribed, but also translated (Brown et al. 2010). When considering that innate immunity is supposed to be self-limiting to the extent that induction of inflammatory responses by an infectious event is to be followed by dampening of this response (Brown et al. 2007), the knowledge we gained from these genomic and proteomic studies about infections with highly pathogenic influenza viruses is suggestive of one or several of the following scenarios: despite the high induction of the innate immune response early on, this response is unable to control replication of these viruses, leading to the 'runaway' inflammatory response seen on gene arrays and confirmed by proteomic analysis and pathology; something specific to these viruses impeded the negative feedback that normally prevents an excessive inflammatory response; perhaps through interference with antigen presentation or interference with effective circulation of lymphocytes due to interferon-induced margination, the hosts are unable to mount or implement an effective virus-specific adaptive response in time to control viral replication and inflammation before extensive tissue damage takes place.

Immunity is comprised of a complex set of integrated responses arising from a dynamic network of thousands of molecules subject to multiple influences (Gardy et al. 2009). Therefore, an immune response cannot be understood by only focusing on discrete biochemical events. It should be thought of as a complex 'color by numbers' design: important interactions are not easily distinguishable from those less decisive until we view each in the context of the others. Systems biology highlights where isolated events converge to drive the clinical course of a disease and provides information on genes previously unknown to be involved in a process and transcriptional or translational occurrences, which may deserve further study through a reductionist approach. It also helps quantify these events and to unravel the impact of natural variations in a population (Shapira and Hacohen 2011). In conclusion, study of the gene categories and pathways transcriptionally or translationally activated by an event can reveal much about the biological processes underlying a phenomenon of interest. Gene expression studies, in particular, hold the potential to provide early and accurate predictions of developing pathologies, mounting immune responses, and clinical course of a disease.

4 Pathology of Influenza Infections: Systems Biology Enhances Findings of Conventional Analyses

Genomic and proteomic studies in tissues directly or indirectly affected by an infection can provide a snapshot of all ongoing events and thereby complement conventional pathology. Macroscopic pathology provides information on the extent

and possible etiologies for observed lesions, but often little else. Microscopic pathology is informative but typically requires time consuming stains to be diagnostic and a preconception of what we are looking for, which may ultimately limit the insights we obtain. In this regard, systems pathology can provide means to shorten the path to a diagnosis and provide clues of ongoing events that lead to a new understanding on ongoing processes. In influenza-infected tissues, several events take place concurrently: viral replication and consequent cytopathic effect on infected cells; innate and adaptive immune responses; consequences of these responses on infected and noninfected cells; and tissue repair or permanent damage. For instance, in a study of seasonal influenza in macaques, gene expression profiling in lungs revealed the upregulation of genes related to neutrophil and monocyte or macrophage function in the same time frame as an influx of neutrophils and macrophages was observed in lungs by microscopic pathology (Baskin et al. 2004). Likewise, transcriptional activation of inflammatory cells and apoptotic pathways coincided with gross and histopathological signs of inflammation, with tissue damage and concurrent signs of repair. Examination of local lymph nodes showed that endothelial cells of most blood vessels were hypertrophic and had large vesiculated nuclei, both evidence of high synthetic activity. Consistently, gene expression showed upregulation of genes relevant to antigen presentation, and T and B cell proliferation. Finally, in this study, expression profiles in lung tissue suggested early presence and activation of cytotoxic T cells, believed to be of benefit to viral clearance early on and potentially harmful later and of natural killer cells, which also play an important cytotoxic role early in infection. Genes related to the functions of these cells were highly induced early on and much less so by day 7, a feature of this uncomplicated influenza infection that nonspecialized pathology may not have easily detected.

In a subsequent study which looked at earlier time points during infection of macaques with seasonal influenza, genomic and proteomic studies allowed us to make new observations which supplemented conventional pathology. For instance, we observed that the presence of influenza virus in a tissue section was more predictive of gene expression and translation than the presence of pathology (Baas et al. 2006). Furthermore, T- and B cell pathways were activated early in infected tissues, both with and without pathology, before significant infiltration by lymphocytes took place. This induction may have been indicative of a role of B cells in antigen presentation in addition to antibody production.

As previously discussed, we could not have predicted the response of bronchial cells to an NS1-truncated influenza virus based on early evidence of infection by the modified virus or lack thereof, or based on any pathology, yet we observed a strong induction of interferon-dependent pathways on gene arrays, which was predictive of the highly protective immune response produced by the experimental vaccine. On the other hand, the lack of transcriptional activity related to an innate response in lung tissue after challenge correlated well with the lack of pathology in animals in this vaccine group (Baskin et al. 2007).

When comparing infections between HPAI and 1918 pandemic recombinant influenza viruses (Baskin et al. 2009), histopathology was an essential tool in

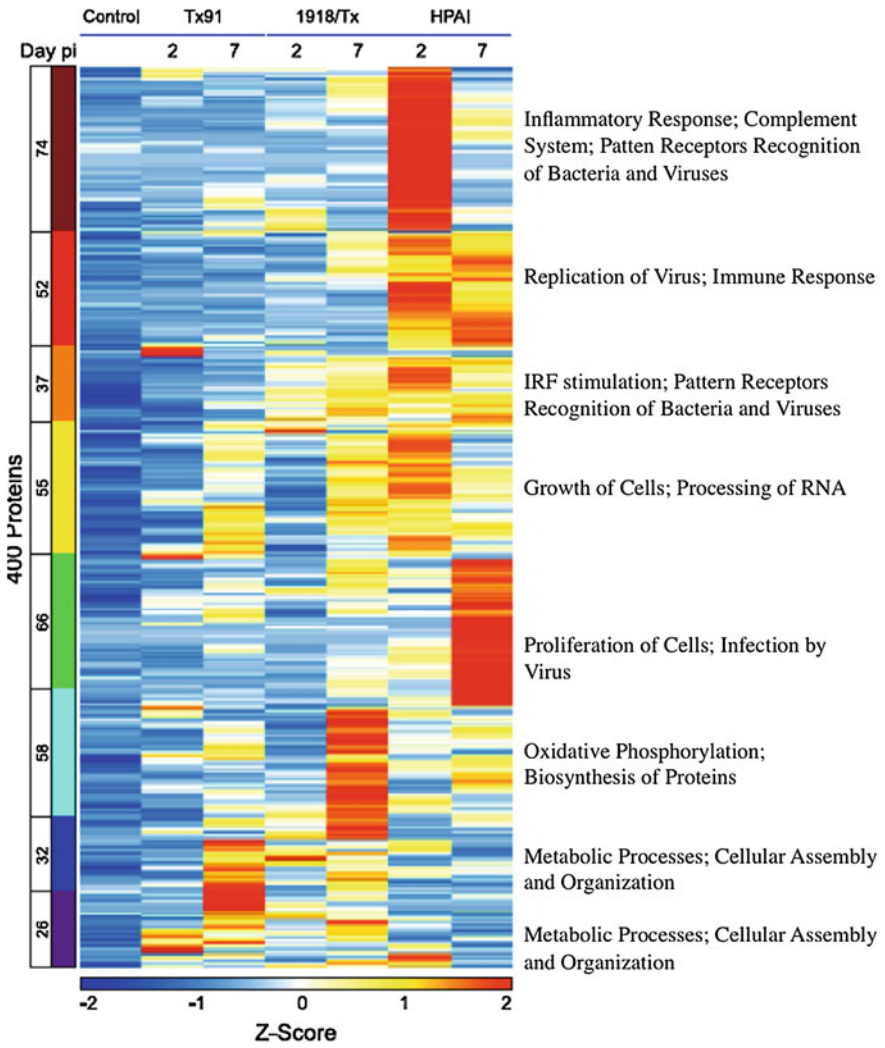


Fig. 4 Heat map of Z-scores of spectral counts for 400 increased proteins in macaques infected with H5N1 HPAI, a 1918 pandemic virus recombinant, and organized into eight clusters by using the K-means algorithm. The eight clusters are color coded relative to the condition (s) in which proteins were most highly expressed. The number of proteins within a cluster is located to the left of the cluster bar. Adapted from Brown et al. (2010)

identifying the lung cells primarily targeted by each virus, and in detecting apoptosis in dendritic cells, which were steadily disappearing from lung tissue in the H5N1 group. While inflammatory gene induction was highly consistent with ongoing lung pathology, gene expression arrays captured the strength of that induction, which was often above the upper limit of detection of the array analysis software. It also provided information on the unique kinetics of the transcriptional

induction brought about by the H5N1 infection: in this group, a slight relative decline in innate and inflammatory gene induction took place 4 days after infection, only to resurge to even higher levels 3 days later. When we looked at protein translation in these same animals, we observed that the protein response correlated well with disease progression and pathology. Interestingly, differences in innate and inflammatory protein translation between the H5N1 group and others 2 days after infection were even more striking than gene expression had suggested (Brown et al. 2010) (Fig. 4). Proteomics also provided evidence of the strong spike in immune cell proliferation process 7 days after infection with the H5N1 virus, something that pathology alone would not have easily discerned amidst the severe ongoing tissue damage. At the same time, it highlighted signs of tissue repair in the group infected with seasonal influenza virus and even in the group infected with the 1918 pandemic virus recombinant, albeit through different pathways.

Macaques infected with a reconstructed 1918 pandemic virus exhibited high and fairly constant interferon induced and innate immune gene expression from days 3 to 8 after infection, again suggesting the unrelenting quality of the response this virus causes (Kobasa et al. 2007). Another study comparing infection between the 1918 and H5N1 HPAI viruses at early time points showed that both viruses elicited roughly similar induction of interferon pathways, although activation was slightly higher at 12 h with the reconstructed 1918 virus but lower by 24–48 h. At 24 h, inflammatory and apoptotic pathways were also more activated in the 1918-infected animals. After that point, cell death induction decreased in that group, whereas it increased in the group infected with the H5N1 HPAI virus. Since this latter caused relatively less pathology, this pattern suggests that apoptosis ultimately may serve to limit viral replication and pathology, a finding confirmed through the use of tunnel assays, viral titration, and conventional pathology (Cillóniz et al. 2009). This study is a perfect example of the use of systems biology as a tool to provide direction for conventional analyses.

5 From Diagnosis to Prevention: Is There a Role for Systems Biology?

Taken together, our experiments in macaques suggest a prognostic value of gene expression and translation analyses early in infection with influenza viruses, even when using fairly un invasive samples such as peripheral blood and bronchial epithelium. Systems biology also enabled us to predict the protective response to a modified live vaccine a mere 2 days after administration. Indeed, the unique advantage of genomic studies is their ability to detect trends in host response almost immediately after an event, whether pathologic, therapeutic, or preventative in nature. Even if every over-expressed gene does not ultimately become translated into a functional protein, the reality is that the data, thanks to its size and granularity, accurately reveals the overall direction of a host response. Integration of several data

sets across diverse resources to focus on genes behaving similarly further reduces uncertainty. Several such studies have been undertaken, which combined data from human (Jenner and Young 2005) or animal (Pennings et al. 2008) experiments and allowed identification of genes co-activated in during infections, and specifically lung inflammation caused by a variety of pathogens, including viruses.

Recently, there has been increased interest in manipulating the innate immune response to control infectious or neoplastic diseases at the earliest stage. As the branch of the immune system that is the most conserved across species, the innate response is critical, powerful, and exhibits a complex system of redundant checks and balances, which when they fail may result in acute pathology of extreme severity or in some of the most debilitating chronic conditions (Brown et al. 2007). The Institute for Systems Biology (Seattle, Washington) has described an in-house interaction database comprising 5,200 bio-molecules and 17,600 interactions directly relevant to the innate immune response. Considering this complexity, trends in response to a stimulus are more easily identified through the use of systems biology than by studying specific reactions, at least initially. This is all the more so that the environment surrounding a cell, such as contact with immune cells or factors and exposure to hormones or extrinsic compounds, can have a profound effect on response to a pathogen. Further, many pathogen-encoded proteins directly interact with molecules of the innate immune response and some of these interactions are species specific in nature (Gardy et al. 2009). Consequently, successful understanding and modulation of the innate immune response in acute or chronic inflammatory events without impairment of its protective role is considered one of the next big frontiers of medicine (Brown et al. 2007). With this goal in mind, methods are being developed for systematic downstream analysis of high-throughput data sets, heralding a new era of integration of systems biology and reductionist approaches.

In regard to drug discovery, systems biology is viewed as a generation tool for new hypotheses. Specifically, it is believed to have the potential to change the current target-based drug development paradigm back to proving a drug's efficacy and safety in live biological systems directly, thereby greatly increasing the efficiency of the pharmaceutical pipeline (Butcher 2005). For instance, systems biology enables simultaneous screening of many different pathways and targets potentially affected by a drug in hosts in which it is effective, and provides a platform for downstream identification of mechanisms of action and side effects. Essentially, systems biology can help pharmaceutical companies get to the answers of: 'Does it work, does it hurt, and why?' faster than by focusing on putative targets first and trying to validate their relevance in live biological systems later, which is still a required step prior to approval for release.

The promise of systems biology for vaccine development is also significant. The search for vaccines against a number of incurable diseases, such as AIDS, has largely failed, which highlights the need for new vaccine strategies, beyond the 'isolate, inactivate, inject' paradigm. One reason for the lack of success is the failure to do comprehensive immune profiling after vaccination that lead to accurate identification of correlates of protection in the immune response to

potential vaccines. Systems biology tools can accelerate vaccine development by identifying predictors of immunogenicity and new mechanisms that underlie protective immune responses. It would help characterize good and poor responders well beyond the few biomarkers that are currently used and even allow for customization of vaccination regimens by appropriate selection of adjuvant, antigen dose, and even route of administration in order to elicit optimal immunity (Trautmann and Sekaly 2011). In fact, this concept has led to the new field of systems vaccinology.

6 Conclusion: The Role of Systems Biology in the Trend Toward Predictive and Personalized Medicine

Systems biology is still in its infancy as a means of scientific discovery, and even more so as a clinical tool of prevention, diagnosis, and therapy. However, all signs point to the fact that it stands to play a role in science and medicine that is becoming increasingly difficult to deny, despite the challenge of parsing, interpreting, and storing large data sets. Our limited studies in a model that is of high relevance to humans have helped show the potential of systems biology and its applications in human medicine, through understanding of the systemic effects of local subcellular events that can be used to predict the course of a disease such as influenza and provide an actionable head start for triage and therapy. These experiments also lead to unique insights on pathways affected by disease, specifically providing information on immune or cellular processes predicting outcomes of preventative efforts and influencing disease course and pathology, none of which were easily and quickly discernible with conventional tools.

Lee Hood described his vision of personalized medicine, what he calls 'P4' medicine: predictive, preventive, personalized, and participatory (Hood 2011). The predictive part has to do with identifying the first network perturbed by disease, which may not be the one producing the biomarkers currently considered pathognomic of a condition, but perhaps another whose induction may be completely pre-symptomatic. Biological systems are more than simple collections of proteins which interact in a linear manner. Instead, they are complex, intricately interacting sets of functional and at times redundant pathways that collectively produce coherent behaviors, which may be unique to an individual. There lies the most important contribution of systems biology to modern medicine so far: the grasp of the highly individual nature of responses to environmental changes, whether the variations are due to different genetic makeup, age, or prior events, whose effects may be cumulative. Even despite these variations, new unifying models can be developed with the essential help of carefully designed animal experiments and of clinical studies. These models will aid in acquiring an enhanced and more comprehensive understanding of disease mechanisms at the subcellular level, as long as the following conditions are met: we embrace but do not become paralyzed

by the increased complexity and size of available information on any one patient; and the need for simplified decision making in clinical settings does not cause us to ignore aspects of this information which should produce personalized, rather than generic therapeutic decisions.

Systems biology as a clinical tool will require unification of sample and data analysis methods and platforms. The lack of standardization in accepted biomarkers for certain conditions is already a problem during diagnosis and drug safety testing (FDA 2004). The volume of data generated with 'omics' tools is likely to heighten challenges in this regard and necessitate an enhanced effort toward adoption of common standards. The use of systems biology will also require a new level of teamwork and communication between basic scientists, pre-clinical scientists, and clinicians, and between clinicians of different disciplines. Such collaborative efforts are already commonplace in state-of-the-art hospitals and are not unusual for private practitioners, but a stable network of complementary expertise will be required to function as a clinician, at a level never before experienced. This collaboration will also need to include the patient in ways that it never has before, if data collection is to take place at well times in addition to during disease, which is a critical component of predictive medicine. In conclusion, systems biology, provided the right tools and approaches, will almost undoubtedly change the way science is done and medicine is practiced in the next few decades.

References

Baas T, Baskin CR, Diamond DL, García-Sastre A, Bielefeldt-Ohmann H, Tumpey TM, Thomas MJ et al (2006) Integrated molecular signature of disease: analysis of influenza virus-infected macaques through functional genomics and proteomics. J Virol 80(21):10813–10828. doi:10.1128/JVI.00851-06

Baskin Carole R, Bielefeldt-Ohmann H, Tumpey TM, Sabourin PJ, Long JP, García-Sastre A, Tolnay A-E, et al (2009) Early and sustained innate immune response defines pathology and death in nonhuman primates infected by highly pathogenic influenza virus. Proc Natl Acad Sci U S A 106(9):3455–3460 National Academy of Sciences http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2642661&tool=pmcentrez&rendertype=abstract

Baskin Carole R, Katze MG (2008) Systems biology could help us understand protect against pandemics. Microbe 3(5):227–233

Baskin CR, García-Sastre A, Tumpey TM, Bielefeldt-Ohmann H, Carter VS, Nistal-Villán E, Katze MG (2004) Integration of clinical data, pathology, and cDNA microarrays in influenza virus-infected pigtailed macaques (Macaca nemestrina). J Virol 78(19):10420–10432 American Society for Microbiology http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15367608

Bouvier NM, Lowen AC (2010) Animal models for influenza virus pathogenesis and transmission. Viruses 2(8):1530–1563. doi:10.3390/v20801530

Brown KL, Cosseau C, Gardy JL, Hancock REW (2007) Complexities of targeting innate immunity to treat infection. Trend Immunol 28(6):260–266. http://www.ncbi.nlm.nih.gov/pubmed/17468048

Brown JN, Palermo RE, Baskin CR, Gritsenko M, Sabourin PJ, Long JP, Sabourin CL et al (2010) Macaque proteome response to highly pathogenic avian influenza and 1918 reassortant

influenza virus infections. J Virol 84(22):12058–12068, American Society for Microbiology (ASM). http://www.ncbi.nlm.nih.gov/pubmed/20844032

Butcher EC (2005) Can cell systems biology rescue drug discovery? Nat Rev Drug Discov 4(6):461–467. http://www.ncbi.nlm.nih.gov/pubmed/17249501

Chen Y, Deng W, Jia C, Dai X, Zhu H, Kong Q, Huang L et al (2009) Pathological lesions and viral localization of influenza A (H5N1) virus in experimentally infected Chinese rhesus macaques: implications for pathogenesis and viral transmission. Arch Virol 154(2):227–233. http://www.ncbi.nlm.nih.gov/pubmed/19130169

Cillóniz C, Shinya K, Peng X, Korth MJ, Proll SC, Aicher LD, Carter VS et al (2009) Lethal influenza virus infection in macaques is associated with early dysregulation of inflammatory related genes. (MS. Diamond, Ed.)PLoS Pathog 5(10):12, Public Library of Science http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid2745659&tool=pmcentrez&rendertype=abstract

Food and Drug Administration (2012) Challenge and opportunity on the critical path to new medical products. www.fda.gov_downloads_ScienceResearch_SpecialTopics_CriticalPathInitiative_CriticalPathOpportunitiesReports_ucm113411. Accessed May 31 2012

Gardy JL, Lynn DJ, Brinkman FSL, Hancock REW (2009) Enabling a systems biology approach to immunology: focus on innate immunity. Trend Immunol 30(6):249–262 http://www.ncbi.nlm.nih.gov/pubmed/19428301

Hood L (2011) Lee Hood. Nat Biotechnol 29(3):191 http://www.ncbi.nlm.nih.gov/pubmed/21390010

Jenner R, Young R (2005) Insights into host responses against pathogens from transcriptional profiling. Nat Rev Microbiol 3(4):281–294 http://discovery.ucl.ac.uk/93597/

Jones-Engel L, Engel GA, Schillaci MA, Babo R, Froehlich J (2001) Detection of antibodies to selected human pathogens among wild and pet macaques (Macaca tonkeana) in Sulawesi Indonesia. Am J Primatol 54(3):171–178. doi:10.1002/ajp.1021

Kobasa D, Jones SM, Shinya K, Kash JC, Copps J, Ebihara H, Hatta Y et al (2007) Aberrant innate immune response in lethal infection of macaques with the 1918 influenza virus. Nature 445(7125):319–323. doi:10.1038/nature05495

Kochs G, García-Sastre A, Martínez-Sobrido L (2007) Multiple anti-interferon actions of the influenza a virus NS1 protein. J Virol 81(13):7011–7021 American Society for Microbiology http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1933316&tool=pmcentrez&rendertype=abstract

Loo YM, Gale M (2007) Influenza: fatal immunity and the 1918 virus. Nature 445(7125):267–268

Pennings JLA, Kimman TG, Janssen R (2008) Identification of a common gene expression response in different lung inflammatory diseases in rodents and macaques. (N. Papavasiliou, Ed.) PLoS ONE 3(7):7 Public Library of Science http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2442866&tool=pmcentrez&rendertype=abstract

Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, Wittkowski KM et al (2007) Gene expression patterns in blood leukocytes discriminate patients with acute infections. Blood 109(5):2066–2077 American Society of Hematology http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1801073&tool=pmcentrez&rendertype=abstract

Rimmelzwaan GF, Kuiken T, Van Amerongen G, Bestebroer TM, Fouchier RAM, Osterhaus ADME (2001). Pathogenesis of influenza A (H5N1) virus infection in a primate model. J Virol 75(14):6687–6691 American Society for Microbiology http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11413336

Shapira SD, Hacohen N (2011) Systems biology approaches to dissect mammalian innate immunity. Curr Opin Immunol 23(1):71–77 Elsevier Ltd. http://www.ncbi.nlm.nih.gov/pubmed/21111589

Shinya K, Gao Y, Cilloniz C, Suzuki Y, Fujie M, Deng G, Zhu Q et al (2012) Integrated clinical, pathologic, virologic, and transcriptomic analysis of H5N1 influenza virus-induced viral pneumonia in the rhesus macaque. J Virol. doi:10.1128/JVI.00365-12

Tolnay, A-E, Baskin CR, Tumpey TM, Sabourin PJ, Sabourin CL, Long JP, Pyles JA et al (2010) Extrapulmonary tissue responses in cynomolgus macaques (Macaca fascicularis) infected

with highly pathogenic avian influenza A (H5N1) virus. Arch Virol 155(6):905-914 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20372944

Trautmann L, Sekaly R-P (2011) Solving vaccine mysteries: a systems biology perspective. Nat Immunol. doi:10.1038/ni.2078 Nature Publishing Group

Baskin CR, Bielefeldt-Ohmann H, García-Sastre A, Tumpey, TM, Van Hoeven N, Carter VS, Thomas MJ, et al (2007) Functional genomic and serological analysis of the protective immune response resulting from Vaccination of macaques with an NS1-truncated influenza virus. J Virol 81(21):11817–11827 Am Soc Microbiol (ASM) http://www.ncbi.nlm.nih.gov/pubmed/17715226

‘Omics Investigations of HIV and SIV Pathogenesis and Innate Immunity

Robert E. Palermo and Deborah H. Fuller

Abstract In the 30 years since the advent of the AIDS epidemic, the biomedical community has put forward a battery of molecular therapies that are based on the accumulated knowledge of a limited number of viral targets. Despite these accomplishments, the community still confronts unanswered foundational questions about HIV infection. What are the cellular or biomolecular processes behind HIV pathogenesis? Can we elucidate the characteristics that distinguish those individuals who are naturally resistant to either infection or disease progression? The discovery of simian immunodeficiency viruses (SIVs) and the ensuing development of in vivo, nonhuman primate (NHP) infection models was a tremendous advance, especially in abetting the exploration of vaccine strategies. And while there have been numerous NHP infection models and vaccine trials performed, fundamental questions remain regarding host–virus interactions and immune correlates of protection. These issues are, perhaps, most starkly illustrated with the appreciation that many species of African nonhuman primates are naturally infected with strains of SIV that do not cause any appreciable disease while replicating to viral loads that match or exceed those seen with pathogenic SIV infections in Asian species of nonhuman primates. The last decade has seen the establishment of high-throughput molecular profiling tools, such as microarrays for transcriptomics, SNP arrays for genome features, and LC–MS techniques for proteins or metabolites. These provide the capacity to interrogate a biological model at a comprehensive, systems level, in contrast to historical approaches that

R. E. Palermo · D. H. Fuller
Department of Microbiology, University of Washington, Seattle, WA, USA

R. E. Palermo (✉) · D. H. Fuller
Washington National Primate Research Center, Seattle, WA, USA
e-mail: palermor@u.washington.edu

D. H. Fuller
e-mail: fullerdh@u.washington.edu

Current Topics in Microbiology and Immunology (2013) 363: 87–116
DOI: 10.1007/82_2012_255

Published Online: 25 August 2012

characterized a few genes or proteins in an experiment. These methods have already had revolutionary impacts in understanding human diseases originating within the host genome such as genetic disorders and cancer, and the methods are finding increasing application in the context of infectious disease. We will provide a review of the use of such 'omics investigations as applied to understanding of HIV pathogenesis and innate immunity, drawing from our own research as well as the literature examples that utilized in vitro cell-based models or studies in non-human primates. We will also discuss the potential for systems biology to help guide strategies for HIV vaccines that offer significant protection by either preventing acquisition or strongly suppressing viral replication levels post-infection.

Contents

1 Introduction

The ability of HIV/SIV to evade host responses, rapidly destroy key immune functions, hijack the immune response for its own benefit, and establish latency presents an enormous challenge that, to date, has not been adequately addressed through traditional approaches that focus on studying specific immune responses as independent correlates of viral control. From both a pathogenetic and preventive perspective, the first molecular and cellular events that occur upon mucosal exposure to HIV/SIV or vaccination are likely critical. HIV/SIV will cross the mucosal barrier in a matter of hours, disseminate locally within the first few days, and then progress systemically into the blood within the first few weeks of exposure. Once the virus enters the very first cells in the mucosa, however, the window for adaptive immunity to respond in time to abort or prevent viral dissemination and establishment of latency may be too short. In contrast, the innate immune system comprises the network of cells and factors

that respond more immediately to HIV exposure and provide the first line of defense. Dendritic cells are one of the first cell types to encounter HIV in the genital tract and play central role in sensing the virus, inducing antiviral defenses, recruiting cells, and stimulating adaptive immunity (Wilkinson and Cunningham 2006). Innate responses likely provide some protection against infection but importantly, they have considerable impact on the nature of adaptive responses that develop and function to establish viral set point and set the course of disease. Although early innate and adaptive responses are critical for viral control, these mechanisms are, paradoxically, also the sources for inflammation and the activation and recruitment of target cells susceptible to HIV infection that can alternatively enhance viral replication and dissemination (Staprans et al. 2004). For this reason, traditional methods that focus on studying one specific antiviral response as an immune correlate of viral control may fail to consistently predict protection, especially if the response stimulates inflammation.

Studies in humans and nonhuman primates (NHP) indicate that it should be possible to develop a vaccine that prevents sexual transmission of HIV. The HIV ALVAC/AIDSVAX vaccine trial in Thailand (RV144), employing a viral vector prime and recombinant protein boost, afforded 31 % efficacy (Rerks-Ngarm et al. 2009). Although this level was not sufficient to support vaccination of the general public, the results provided the first evidence that an HIV vaccine can prevent infection in humans. Studies in nonhuman primates also suggest a more efficacious vaccine is feasible. As in the Thai trial, some vaccines that have been tested for protection in SIV or SHIV nonhuman primate models for AIDS have afforded sterile protection against infection or profound and durable control of viral replication (Barnett et al. 2010; Barouch et al. 2012; Belyakov et al. 2001; Daniel et al. 1992; Fuller et al. 2002; Lai et al. 2011; Manrique et al. 2011; Patterson et al. 2004). Designing a vaccine that can achieve better protection, however, will require a more complete understanding of the HIV/SIV-host interactions, factors contributing to pathogenesis, effective mechanisms of antiviral immunity and vaccine action, and correlates of protection. Indeed, an understanding of early details in HIV–host interaction may illuminate vaccine strategies that more potently engage the innate immune system, thereby yielding a superior adaptive response.

Elucidating the broader network of responses that influence pathogenesis and prevention of HIV infection is critical for designing novel therapeutic and prevention strategies. Because systems biology enables the integration of large datasets from multiple analyses, it represents an exciting new frontier to elucidate how innate and adaptive immune responses to HIV are induced by infection or vaccination, coordinately regulated, and influence protection and pathogenesis. Here, we discuss how systems biology is being employed to define these immune interactions and how this information may lead to new insights into HIV/SIV prevention and disease.

1.1 Systems Level Investigations

The term Systems Biology has progressively acquired a breadth of meanings. Perhaps at its most complex and challenging, it can be termed the characterization of a biological model by multiple high-throughput molecular profiling techniques, and the ensuing use of mathematical techniques to find relationships/associations between the measured molecular entities (Aderem et al. 2011; Tisoncik et al. 2009). These molecular entities are typically measured by class-specific techniques (e.g. mRNA levels; protein abundances, possibly including post-translational modifications; lipids or other metabolites; histone modifications, etc.), and the mathematical approaches may determine relationships within the same class of analyte (e.g. mRNA—mRNA) or between classes (e.g. mRNA—protein; protein—metabolite). Inasmuch as the biological model generally tracks the system in response to a perturbation or stimulus, the modeling may ultimately yield a framework that reveals previously unappreciated associations between the measured components, and furthermore may be predictive of how the system would respond under as yet untested conditions ("hypothesis-generation"). *For a translational impact in biomedical research, this assessment of the system must ultimately integrate (predict, explain) the phenotype of the biological entity under study.* When dealing with the simplest systems such as uniform cell populations, the measured molecular entities can be the phenotype; for example a cytokine produced, or a response with a well-characterized molecular etiology such as apoptosis.

In considering the systems biology as applied to whole organisms, the complexity in the endeavor increases. High-throughput measurements are generally performed on samples composed of multiple cell types, and may contain components not endogenous to the tissue being examined. The responses of the tissue at hand may arise from bioactive materials produced or regulated elsewhere in the organism, and exhaustive characterization of all the tissues/compartments in the organism is impossible. Rigorous uniformity of experimental conditions is difficult to achieve, and this includes variation arising from genetic diversity in outbred species (e.g. nonhuman primates, human subjects). Likewise, the phenotypes to be integrated take on a much greater complexity and are likely elicited by many contributing processes. Nonetheless, the methods of systems biology can still find application in producing statistically significant correlations of molecular features to complex phenotypes. Again this serves the remit of hypothesis generation, and can possibly identify a subset of molecular features with potential diagnostic or prognostic applications. This latter form of systems biology is well exemplified by recent studies relating gene expression changes in blood following vaccination to the resulting immunogen-specific cellular or humoral immune responses (Gaucher et al. 2008; Nakaya et al. 2011; Querec et al. 2009).

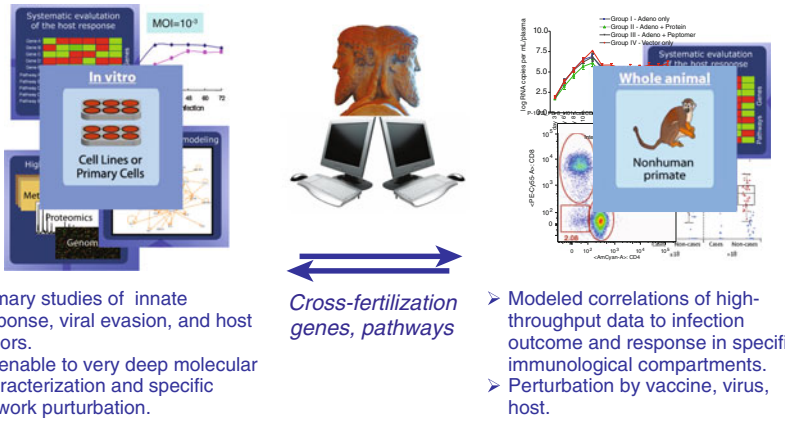


Fig. 1 Current systems biology approaches can span extremes of biological scale and complexity. Cell-based, in vitro experimental models offer the greatest level of experimental control and the potential for very comprehensive high-throughput measurements; these may then be utilized in mathematical models that can identify critical molecular pathways and crucial system components. For in vivo models, many investigators employ statistical methods to obtain correlations of molecular features to observed phenotypes, or to refine prognostic classifiers to apply in related in vivo settings. Ideally, these approaches have synergy: detailed cell-based systems biology investigations should predict the responses of such cells in an in vivo context, and thereby help parameterize more complex mathematical models. Should the cell-based predictions not be consistent with the in vivo observations, further elaboration to the in vitro systems may be necessary

As we have described previously, our approach in systems biology investigations encompasses both extremes of model complexity (cf. Fig. 1) (Aderem et al. 2011; Tisoncik et al. 2009). Simple, in vitro cell-line models offer the best experimental consistency along with ease of execution under differing conditions (perturbations). For understanding the host response to viral infection, it can also provide the means to assess the impact of the virus on a cell type that is the primary target for infection and replication. After characterizing the response in this constrained setting, one can then assess the portability of the characteristics to the in vivo context where the target cells will be just one cell type within the sample. Moreover, signaling is occurring between cells, and this can include ensuing immune surveillance of the infected cells. This approach has been the basis of our Center for the systems virology of highly pathogenic respiratory viruses (Aderem et al. 2011). Results from this endeavor have included our publications on the conserved elements of the host response to highly pathogenic avian influenza, where we discuss linking transcriptional networks determined in cell culture infections to the immune response and regulatory programs that are observed in the respiratory tissues of mice and macaques when infected with the same virus (McDermott et al. 2011).

1.2 Distinctions in Innate Immunity

The innate antiviral response is that sensing and control capacity that does not require prior "schooling" in the distinction of self versus nonself. The sensing molecules have evolved to recognize *pathogen-associated molecular patterns (PAMPs)* and this capacity is not altered by diversification or affinity maturation as occurs in the adaptive arm of the immune system and which require the somatic genome alterations that occur in B and T cells.

However, even within the broader category of innate immunity, there is an important distinction in these innate sensing capabilities. In absence of a deleterious mutation, all cells possess a capacity for antiviral sensing by RIG-I-like RNA helicases, including RIG-I and MDA5 (Yoneyama and Fujita 2009). These proteins recognize structural elements contained in viral RNA with downstream signaling events that result in the transcription of interferon-β as well as other antiviral effector molecules. Secreted IFNβ then triggers the Type I interferon response in the infected cell as well as being a paracrine regulator to nearby cells. The Type I response results in the expression of other *interferon-stimulated genes (ISG's)*, of which many are additional antiviral effectors. Signaling in the RIG-I-like receptor pathway also results in the increased expression of inflammatory genes due to the nuclear translocation of NFκB, and leads to the activation of the inflammasome. The Type I interferon response and the initial inflammatory response are important in bringing the virally infected cell under the scrutiny of other components of the immune system, through either the secretion of chemokines and cytokines that recruit immune cells to the infection locus, or by changing cell surface features that help immune cells target the infected cell. Prominent in this later category is up-regulation of components in the antigen-presentation machinery, including major histocompatibility complex (MHC) molecules; this then increases the display of antigenic peptides derived from viral proteins. It is also worth remarking that the RIG-I-like receptors are also implicated in the antiviral response to DNA viruses (and other pathogens with DNA genomes). This is an indirect sensing mechanism that occurs after RNA polymerase III transcribes portions of the pathogen's DNA, yielding RNA transcripts that engage the RIG-I like receptors, thereby broadening the role of this pathway in the general response to intracellular pathogens. (Chiu et al. 2009)

In contrast to these generally manifest pathways, the toll-like receptors (TLRs) are a category of PAMP sensors generally expressed in a subset of immune cells, and their role is most clearly understood in specialized the context of antigen-presenting cells (APCs) such as dendritic cells (DCs) and macrophages (MØs) (Kawai and Akira 2010; Moresco et al. 2011). As transmembrane receptors, TLRs enable activated DCs and MØs to sense pathogens in their environment rather than within their cytosol. Some TLRs reside on the exterior cell surface and detect components on the exterior of bacteria or viruses. The TLRs critical for antiviral sensing are TLR3 (detecting double stranded RNA; dsRNA), TLR7 (for single stranded RNA; ssRNA), and TLR9 (unmethylated CpG DNA). To detect the

genomes of pathogens these receptors are localized to the endosomes, where they can sense the RNA and DNA patterns released from endocytosed virions, bacteria, or such components from engulfed cells and cellular debris. Downstream signaling from these TLRs involves the components of either the MyD88 or TRIF pathways, and results in gene expression driven by the transcription factors IRF3, IRF7, and NFκB. Activation of the IRF3/IRF7 heterodimeric transcription factor can drive the expression of interferon-β (Ifnβ), analogous to the downstream effects from the RIG-I like cytosolic receptors. But in the case of *plasmacytoid dendritic cells (pDCs),* there is also extensive amounts of interferon-α (Ifnα) produced, driven by the IRF7 homodimer transcription factor. Following viral infection of an organism, Ifnα levels are typically far greater than Ifnβ; therefore, this Ifnα production by pDCs is the primary driver of the disseminated Type I interferon response and serves to induce a broad state of antiviral preparedness.[1] Within these specialized antigen-presenting cells, the consequences of NFκB-driven transcription and inflammasome activation are more impactful, with the production of Ifnγ, Tnfα, IL6, IL12, and other inflammatory cytokines. This results in the homing and activation of neutrophils, NK cells, macrophages, and lymphocytes, orchestrating an initial immune response that either clears the infection or keeps it in check until the adaptive response has developed sufficiently to eliminate the pathogen.

2 Investigations of Innate Immunity to HIV and SIV

The Katze laboratory has a long history of investigating virus–host interactions, innate antiviral signaling, and the Type I interferon response (Katze et al. 2002; Korth et al. 2005). Much of this work has centered on examining these processes at stages in advance of the involvement of specialized cells of the immune system. To make a pugilist metaphor, how much of a fight can an infected cell put up on its own, and how would it encourage assistance from other cells better at detection and containment?

2.1 In vitro Investigations

One important antiviral mechanism the Katze group has studied extensively is translational control, as occurs by the interferon-induced protein kinase PKR (gene symbol EIF2AK2) (Gale and Katze 1998; Katze 1995). Constitutively expressed at low levels, it is strongly up-regulated in the Type I response, and it affects

[1] We have attempted to follow a convention where genes and their transcripts are denoted by their HUGO gene symbols, which are typically all upper case. If referring to a protein, conventional abbreviations are used. E.g. IFNB versus Ifnβ.

translational shut-off upon binding dsRNA as would be present in viral genomes or replicative intermediates. Numerous viruses have developed strategies to both overcome the translational blockade that would be imposed by PKR, as well as eliminating host mRNAs that would compete with viral transcripts for the protein synthesis resources. Our early studies with HIV showed how this lentivirus may subvert translational arrest by inducing the more rapid turnover of host mRNAs, and by down-regulating the abundance of PKR (Agy et al. 1990; Katze and Agy 1990).

The advent of microarray technology in the 1990s provided the means to forego the "gene-by-gene" investigation of such experimental systems, offering instead platforms to simultaneously interrogate thousands of mRNA transcripts (Katze 2002). After demonstrating the initial application of cDNA microarrays to observe expression changes in an HIV-1 infected CD4 + T cell line, we published one of the first more comprehensive expression analyses of this infection model using an in-house array design that assessed expression of 4,600 cellular RNA transcripts (Geiss et al. 2000; van 't Wout et al. 2003). An important attribute of this in vitro cell-line model was generating a homogenous cell population with >90 % of the cells initially infected and synchronously transiting the viral life cycle. Despite the limitations as compared to contemporary technologies, this early array study yielded a number of salient observations. First, despite clear evidence of a uniform infection of the cells and the production of protein-coding viral transcripts as early as 8–12 h post-infection (hpi), there were no significant changes in host gene expression until 24 hpi. Within the 409 host differentially expressed genes (DEGs), there was up-regulation of negative cell cycle regulators and down-regulation of positive regulators, reflecting the G2 arrest of infected cells. Though most transcription factors (TFs) were down-regulated, a small group of T cell associated TFs were up-regulated (ELF4, GATA3, SP140, TAL1) along with others previously associated with HIV infection (Jun, RELB).

Moreover, despite the general decrease in transcripts for many metabolic enzymes, numerous genes in the cholesterol biosynthetic pathway were up-regulated. This impact on the cholesterol biosynthetic pathway was ascribed to the action of the Nef viral accessory gene (van 't Wout et al. 2005). It was later demonstrated that a functional Nef gene was required to result in this induction of multiple cholesterol synthesis genes, which are generally regulated by the sterol-responsive element-binding factor 2 (Taylor et al. 2011). The importance of Nef in tuning the lipid composition of host cell membranes, including the enrichment of lipid rafts, is now well evidenced from multiple investigations and appears to be an important mechanism for facilitating virus propagation (Brugger et al. 2007).

In addition to developing functional genomics capabilities with arrays based on human genome sequences, the Katze laboratory was also an important contributor to genomics tools for nonhuman primate studies (Gibbs et al. 2007; Magness et al. 2005). The importance of nonhuman models for AIDS was a very significant driver in this latter endeavor, and in collaboration with Agilent Technologies, the laboratory designed the first oligonucleotide microarrays based on nucleic acid sequences for the Indian-origin rhesus macaque. The first design derived from an

EST sequencing effort by the laboratory covered only a modest number of transcripts; however, the second design based on the draft rhesus genome was more comprehensive and could simultaneously profile over 18,000 macaque transcripts (Wallace et al. 2007). In addition, the array was annotated with the corresponding human gene symbols, allowing the ready interpretation of results based on the much richer documentation for human genes. These arrays have found extensive application in characterizing NHP biomedical models for influenza and Ebola viruses, and as will be discussed below, for NHP models of AIDS (Baskin et al. 2007; Cilloniz et al. 2011; Cilloniz et al. 2009; de Lang et al. 2007; Kobasa et al. 2007; Safronetz et al. 2011).

2.2 *In Vitro Infections in Primary Cells*

These rhesus macaque arrays have also been employed in characterizing in vitro infections in primary cells from rhesus and pigtail macaques. In the first array study, we performed a synchronous infection with SIV_{mac239} in rhesus macaque PBMCs, and the study illustrates the challenges in transitioning the systems approach from a cell line system to primary cells (Thomas et al. 2006). For example, for undetermined reasons the infection was only effective in two of three different donor animals; furthermore, in situ staining showed that at maximum Gag production, only 3–5 % of the cells were infected, despite cytometric characterization that CD4 + T lymphocytes constituted ~30 % of the cells. Many functional genomics studies are performed on samples that are mixtures of cell types, and must be interpreted as the amalgamated response of all the constituents. Expression changes are interpreted as the transcriptional response of cell types that are logically expected to be in the sample, or as might be the case with infiltrating immune cells, the DEGs may reflect the change in the proportions of cell types in the samples.

This example of in vitro infection of PBMCs can be interpreted as the response of the lymphocytes and myeloid cells within the context of low levels of infection. Despite the low percentage of infection, the comparison of infected to mock-treated cells for each animal showed hundreds of differentially regulated genes, with the greater number of changes actually occurring well in advance of peak of viral replication. For the 184 DEGs identified at this early time point, many of them were associated T and B cell signaling, antigen presentation, and integrin signaling. These results contrast with the observations from the in vitro cell-line experiments both in the earlier kinetics of the response, and the clear abundance of genes in the adaptive immune response. We cannot rule out that some of these distinctions may reflect the differences between the cell-line versus primary CD 4 + T cells, as well as the differing virus types (HIV-1 vs. SIV_{mac239}). But another simple explanation is that the response observed in the in vitro infection of the rhesus PBMCs does not originate in initially infected CD4 + T cells but rather chiefly represents the transcriptional response of other cells, such as monocytes,

macrophages, and dendritic cells that have become activated/infected after encountering the virus. In addition to their own transcriptional programs, these cells may then secrete factors that activate uninfected T cells and B cells in the sample. However, one common attribute in this study with primary cells as compared to the cell line results is the absence of any clear transcriptional signature for early innate sensing by the RIG-I-like receptors.

A similar set of observations attended another study performed in primary macaque PBMCs, on this occasion utilizing cells from pigtail macaques (Macaca nemestrina)—a species that generally experiences more severe sequelae to lentiviral infection, and in vivo can even support limited levels of HIV-1 replication (Agy et al. 1992; Batten et al. 2006; Frumkin et al. 1993). The study was a comparison of expression profiles for pigtail PBMCs following infection with either SIV_{mac239} or $HIV\text{-}1_{LAI}$ (Li et al. 2007). As in the prior study with in rhesus PBMCs, only a small percentage of cells were infected, and yet comparison of expression levels for the infected versus mock samples revealed a large number of DEGs for each virus. Likewise, the timing of the expression changes relative to the viral replication kinetics suggests that most of the gene expression changes originate from cells other than infected CD4 + T cells. Also as in the prior examples, there are no expression changes to indicate any innate immune responses originating from RIG-I-like signaling from initially infected cells. Interestingly, though the HIV-1 replication levels based on Gag production were significantly lower than for the SIV strain, the expression changes in the HIV-1 infected cells showed stronger, early up-regulation of antigen-presentation pathways and NK surface markers. This suggests that the *M. nemestrina* host factors may be more sensitive and responsive to determinants presented by the HIV strain than to those on the macaque-adapted SIV strain. This is in general keeping with the understanding that establishing a primate lentivirus infection is a complex interplay between the virus and species-specific restriction factors, as well as the capabilities of the virus to modulate the specialized cells associated with antigen presentation and adaptive immunity.

2.3 Characterization of In Vitro Systems by Proteomics

To add to the systems view of the virus/host interaction during HIV-1 infection, the Katze Laboratory has also undertaken proteomics investigations of in vitro infection models. Although the initial production of a protein is subordinate to the production of its encoding mRNA, comparative analyses often reveal many instances where protein abundance is poorly correlated to the corresponding mRNA levels (Gygi et al. 1999; Tian et al. 2004). In addition, many viruses have evolved mechanisms to advantageously target specific host proteins for destruction, such as the Vif-mediated degradation of the lentivirus restriction factor APOBEC3G (Henriet et al. 2009; Sheehy et al. 2002). Using the lymphoid cell line CEMX174, uniform infections were performed with $HIV\text{-}1_{LAI}$, and using LC/MS

proteomics techniques the protein abundance changes were determined at the peak of viral protein production 36 hpi, with > 94 % of the cells actively producing virus (Chan et al. 2007). Of ~3200 identified proteins, 687 were determined to show statistically significant abundance changes as compared to mock-infected samples; within the differentially regulated set, 83 proteins had been previously identified with known interactions with HIV viral proteins Integrase and Vpu. Functional enrichment analysis identified changes in a large number of proteasome constituents and ubiquitination enzymes that would potentiate the G1/S cell cycle arrest, and alterations of the RAN pathway for nuclear membrane trafficking with net changes that would favor nuclear export, potentially facilitating the export of viral mRNAs.

An analogous proteomics investigation was subsequently performed using isolated primary human CD4 + T cells obtained from five individuals, with infection by HIV-1_{LAI} performed at high MOI, with sampling at 8 and 24 hpi to study a single replication cycle within the synchronously infected cells (Chan et al. 2009). While there were only 25 differentially regulated proteins at 8 hpi, notable among them were up-regulated proteins in energy production and protein synthesis, with several ribosomal protein in the latter category. These observations of early increases in ribosomal and energy production proteins were also recapitulated by the laboratory in a proteomics study using a lymphoblastoid cell line at 4 hpi with HIV-1_{LAI}, suggesting that very early after HIV infection, a T cell undergoes a shift to increased protein synthesis (Navare et al. 2012). In contrast at 24 hpi, from a total of 138 differentially modulated proteins, 24 are down-regulated components in the ribosome and components in the protein synthetic pathway. This extensive negative regulation of protein biosynthetic machinery was not as evident in the aforementioned functional genomics studies, and highlights the added information garnered from applying a different high-throughput analytical technology. Moreover, this alteration may in part explain prior observations concerning the shut-off of host protein synthesis in later stages of infection (Agy et al. 1990). At 24 hpi in these primary cells, there were also signs of dysregulation in mitochondrial components and anti-apoptotic 14-3-3 proteins, observations that coincide with the cells being poised for apoptosis. This may contrast with experiments in cell lines, where genetic alterations may reduce the responsiveness of the cells to pro-apoptotic signals.

2.4 Transcriptional Analysis By Next-Generation Sequencing

Transcriptional profiling by next-generation sequencing (NGS) enables the highly sensitive measurement of transcript levels and by adjusting the cDNA library preparation, the method can interrogate just polyadenylated transcripts (typically mRNAs; mRNA-seq) or it can measure the entire compliment of RNA species, both protein coding and noncoding (total RNA-seq) (Mortazavi et al. 2008; Sultan et al. 2008). RNA-seq offers advantages of increased dynamic range, greater

Table 1 mRNA-seq analysis of differentially expressed host genes in HIV-infected SUP-T1 cells[a]

	No. of DE genes[b]	
	12 hpi	24 hpi
Up-regulated genes		
$1 < FC \leq 1.5$	37	1386
$1.5 < FC \leq 2.0$	5	1040
$FC > 2$	1	246
Total up-regulated	43	2672
Down-regulated genes		
$1 < FC \leq 1.5$	36	1079
$1.5 < FC \leq 2.0$	24	851
$FC > 2$	3	404
Total down-regulated	63	2334

[a] Triplicate samples for each condition and time point were analyzed by mRNA-seq, with data collected as 75 bp single-end reads, with ~30 million reads per sample. Reads were mapped to the human genome and gene expression levels were derived using RefSeq transcript annotations
[b] Normalized transcript abundance level were analyzed in limma for 9992 total gene loci detected in at least one biological condition. All genes listed have Benjamini-Hochberg-corrected *p*-values of less than 0.05

sensitivity, and better precision compared to array methodology; in particular, the latter features make it a superior technique for quantitative comparison of low-abundance transcripts. In addition, unlike arrays the data collection does not depend on prior knowledge of the sequences to be detected, and therefore can lead to transcript discovery (Trapnell et al. 2010), an aspect that is leading to the annotation of many new noncoding RNAs (ncRNAs) along with the deepening appreciation that these ncRNAs change in abundance under many biological conditions including viral infection and the immune response (Guttman et al. 2009; Mercer et al. 2009; Pang et al. 2009, 2010).

For these reasons we utilized mRNA-seq to examine the transcriptional changes in a CD4 + T cell line infected with HIV-1_{LAI}. As in prior studies, we strove for sample homogeneity by generating a uniform (> 90 %) synchronously infected population of cells, and sampled them during the first round of viral replication at 12 and 24 hpi. We then performed mRNA-seq to compare the transcriptional changes in infected cells versus time matched mock-infected samples. Table 1 provides a summary of the experiment. At 12 hpi, production of the Gag protein is just becoming measurable, and yet viral RNA already accounts for ~18 % of total mapped reads; at 24 hpi approaching the peak of viral production, the proportion of viral reads had increased to ~38 %. While the level of viral RNA at 12 hpi would appear to be a tremendous perturbation to the cells, the number of differentially expressed genes is still remarkably small with only 43 up- and 63 down-regulated DEGs (Table 1). Over 90 % of the DEGs from 12 hpi also pass the statistical threshold at 24 hpi, and as illustrated in Fig. 2, these exhibit the same directionality relative to mock, with the majority having increased differentials. The functional annotations showed enrichment in T cell differentiation, with six of

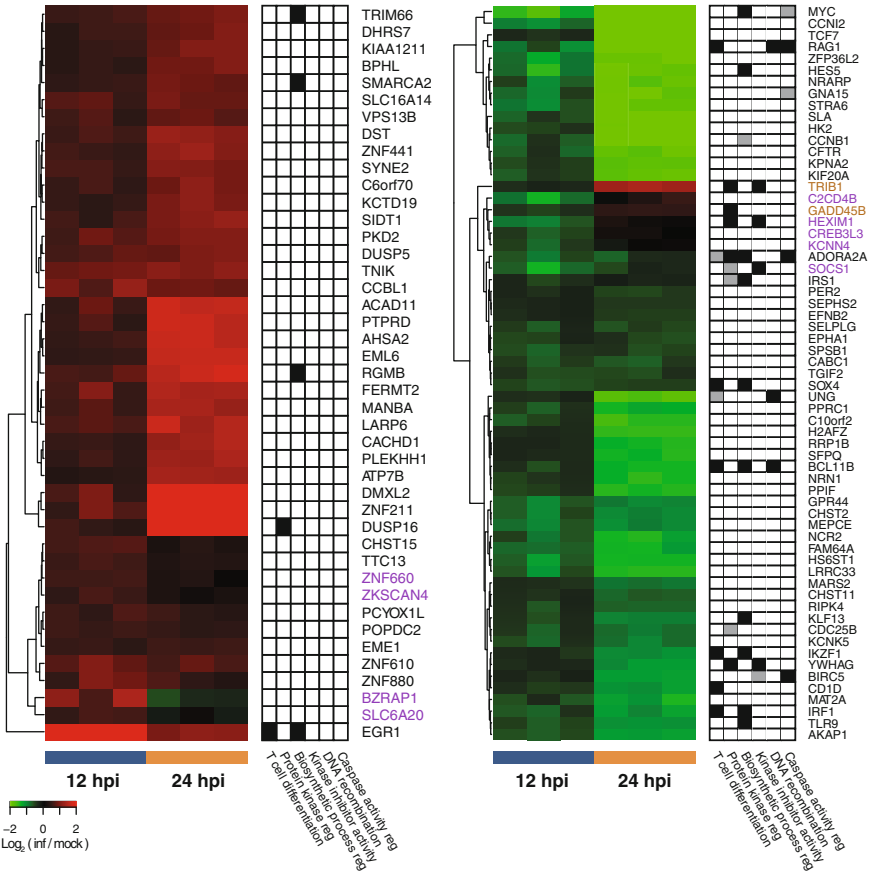


Fig. 2 mRNAseq results for differentially regulated genes at 12 and 24 hpi in HIV-1_{LAI} infected SupT1 CD4 + lymphoblastoid cells. Values shown are $\log_2$(ratios) for each individual infected replicate relative to averaged mock-infected samples at each time point. Genes were segregated by direction of change relative to mock infection at 12 hpi. Hierarchical clustering was done within each directional group. Purple font indicates genes that were not also DE at 24 hpi, while gold font indicates genes that were also DE at 24 hpi with changed directionality at 24 hpi. Annotations indicate over-represented categories in DAVID. Black squares indicate matches to top-scoring categories in each DAVID annotation cluster, while gray squares indicate matches to related categories in the same DAVID cluster as the top-scoring category. Reproduced with permission from Chang et al. 2011

seven genes down-regulated. This negative impact on T cell functionality expands within the large number of DEGs at 24 hpi with down-regulation of both central signaling nodes such as LCK as well as surface receptors including CD2, CD3, CD4, CD8, and CD28.

The number of DEGs at 24 hpi represents a massive reprogramming of the cell, even if one limits consideration to just transcripts with $>$ 1.5-fold changes. Many of the down-regulated genes are involved in RNA processing including numerous

components within the RNA splicing machinery or involved in RNA transport. These alterations may abet the production of unspliced viral RNA, and may be corrupting the RNA transport mechanisms to enable the full-length unspliced viral genome to exit the nucleus for both protein production and virion packaging. Remarkably, the up-regulated genes from 24 hpi show little enrichment in functional characteristics, and are so broadly distributed across pathways and functions as to fail to achieve statistically significance. Finally, even with the sensitivity and precision afforded by mRNA-seq, there is no evidence in the transcriptional profiling for an innate response. At the 12-hpi time point, there are no up-regulated transcripts that would have arisen downstream from engagement of the RIG-I-like signaling pathway. Even at 24 hpi, there is still no indication of an innate immune response or the typically associated inflammatory processes.

2.5 Other Examples from In Vitro Infection Models

2.5.1 CD4 + Cell Lines and Primary CD4 + T Cells

A review of the literature on microarray studies to analyze HIV-1 infection in CD4 + T cell lines and primary cells does not reveal any inconsistencies with our contention that HIV-1 infection does not trigger an early antiviral response in these cells. Unfortunately, in multiple instances the low levels of infection reported in the models qualify the interpretation of many studies. Corbeil and colleagues published an early example in 2001, where effort was made to perform a synchronous, high MOI infection using $HIV1_{LAI}$ in a CEM cell line that was modified to express green fluorescent protein driven by the HIV-1 LTR (CEM-GFP) (Corbeil et al. 2001). Despite the high MOI and the use of polybrene to enhance infection, at 24 hpi only ~30 % of the cells were positive for viral replication, rising to ca. 90 % by 48 h; therefore, the model represents multiple rounds of infection at later time points. The CEM CD4 + lymphoblastoid cell line is the same background employed in array study described as part of the Katze efforts, and both studies used CXCR4-tropic viruses. Some aspects of the studies are quite similar, such as the observation that most expression changes occur late in the course of infection and are dominated by down-regulated genes in the infected cells. However, the Corbeil study, using an early generation Affymetrix array, did report observing hundreds of differentially regulated genes at times between 0.5 and 16 hpi; this differs from the Katze results where very few DEGs were observed before 24 hpi ,and the disparity may arise from the greater number of replicates and stricter statistical criteria and used in the Katze study. The earlier publication makes passing comment that these early DEGs included the antiviral genes interferon alpha 12 (IFNA21) and the interferon inducible gene MXB. Unfortunately, there is no further articulation of this result, no direction is specified for these expression changes, and it is sparse representation for an antiviral response. Therefore, in absence of a highly efficient, synchronous infection and

better statistical criteria, there seems little evidence for early PAMP-triggered signaling events. A similar qualification attends the paper by Yin et al. on analysis of CEM-SS cells infected with HIV-1_{IIIB}, where aspects of gene expression were examined at 7 and 18 days post-infection (dpi) (Yin et al. 2004). While the proportion of infected cells was ca. 90 % at both time points they represent quite different stages in the course of the infection as most cells undergo apoptosis at the earlier lime point, whereas at 18 dpi the culture is primarily an outgrowth of viable, chronically infected cells. Nonetheless, at neither time point were there representatives of strongly up-regulated genes that typify early transcriptional events downstream of innate antiviral sensing) such as ISG15 and ISG54.

For study of primary CD4 + leukocytes, the study by Imbeault et al. used bead-based negative selection on PBMCs to obtain a population of cells enriched in the CD4 + marker, and then infected the stimulated cells with HIV-1_{NL4-3} or with an isogenic variant produced under conditions where the virion would incorporate ICAM-1 into the viral envelope (Imbeault et al. 2009). Here again, the efficiency of infection appeared quite low with only ~10 % of the cells positive for p24 production 5 days post-infection with the parental HIV-1_{NL4-3}, even with continuous exposure to the virus. For array measurements, samples were taken at 8 and 24 hpi; infections were done using cells from five donors; however, the samples were pooled for the microarrays, thereby giving only one expression measurement for each condition. The authors reported 404 DEGs impacted by HIV infection, with 80 % of the expression changes occurring at the later time point. The investigators placed particular emphasis on the up-regulation of p53 that was observed with both the 8- and 24-h data. They attributed the increased expression of p53 to Type I interferon signaling, and biochemical assays of Type I interferon levels did show increased levels (albeit still quite low) that peaked ca. 6 h after infection. However, the assertion that Type I interferon signaling has occurred in the model is not strongly supported as there was no up-regulation of other genes more archetypical for this pathway. Nonetheless, the biochemical and bioassay data on interferon production is an interesting finding that merits reexamination. A complication to the interpretation is the presence of ~10 % CD14 + myeloid cells in the CD4 + enriched cells used for the measurements. These CD14 + cells are likely monocytes, macrophages, or myeloid dendritic cells, and the authors acknowledge the possibility that the features they attribute to Type I interferon signaling could originate from these adventitious myeloid cells or even from co-purified CD4 + plasmacytoid dendritic cells.

To contrast with the work by Imbeault, two other publications examined expression changes in human PBMCs. The first study by Vahey and colleagues was a carefully executed, synchronous infection with a high MOI of HIV-1_{RF}, using PBMCs from three different donors (Vahey et al. 2002). Array comparisons of infected to mock samples were performed at 1, 12, 24, 48, and 72 hpi, using an Affymetrix GeneChip that measured levels of 12,627 transcripts. As with the previously described experiments using macaque PBMCs, the extent of virus infection appears to have been limited to a small percentage of the cells. The report does note the regulation of 57 immune response genes from 1 to 48 hpi; at 12 hpi

these include granulysin (GNLY), and macrophage colony-stimulating factor 1 (CSF1), suggesting that some of the most prominent expression changes are not from CD4 + T cells. The only antiviral gene noted is MX1, observed as up-regulated at 48 hpi, which is anomalously late for an early antiviral response. Another study with primary PBMCs was reported by Gupta et al., evaluating gene expression at 7 days post-infection at MOI 0.01, where only ~7 % of the CD4 + cells were infected (Gupta et al. 2011; Venkatachari et al. 2008). The statistical analysis of infected versus mock for six donors gave 444 DEGs, spanning functional categories such as apoptosis/cell cycle, MAP kinase pathways, and SRC kinases, but no indication of RIG-I-like signaling or a Type I antiviral response.

2.5.2 In Vitro Macrophages

In vivo, macrophages are regarded as likely reservoirs for HIV-1 inasmuch as infected macrophages appear to be much longer lived and do not succumb to virally induced apoptosis or cytopathic effects. The ease of isolating primary monocytes has made it more common for expression studies of HIV-1 infection to be performed with monocyte-derived macrophages (MDMs), ideally from multiple donors. One of the first such studies, using the CCR5-tropic BaL strain made the intriguing observation that a large number of interferon-stimulated genes are up regulated from 2 to 16 h after synchronous infection, as well as other genes that are plausibly downstream of PAMP sensing by either RIG-I-like receptors or TLR receptors (Woelk et al. 2004). This was based on array studies with cells from a single donor, and then corroborated by qRT-PCR with cells from two additional donors. Up-regulated genes included IRF7 (twofold), ISG15 (threefold), IFIT1 (fourfold); and while the percentage infection was not reported, the rapid kinetics clearly link the increased expression to the primary encounter between virus and macrophage.

Unfortunately, other studies have generally used low multiplicities of infections and have examined the expression changes after several days and multiple rounds of viral replication, and even at these later time points only a small percentage of cells appear to be infected. Despite these drawbacks, Vazquez et al. performed expression analysis 6 h postinfection of MDMs with HIV-l_{BAL} (Vazquez et al. 2005). This study was limited by use of a filter-based array with narrow coverage, but DEGs were stringently required to be > twofold regulated in six donors. They noted early increases in transcripts for inflammatory cytokines such as TNFA, IL8, and IL12—all indicative of an early response from pattern-recognition receptors.

2.5.3 In Vitro Dendritic Cells

Dendritic cells play a crucial role in the early antiviral response with the plasmacytoid subset being potent producers of Type I interferon-α following engagement of their TLR receptors (Szabo and Dolganiuc 2008). In addition, dendritic cells have been determined to play a role in the dissemination of HIV or

SIV to CD4 + T cells via cell-to-cell transfer. The interaction of HIV-1 with dendritic cells has been characterized as having two phases: in the early phase (< 24 h), the virus is endocytosed by its capture on C-type lectin receptors and gets trafficked to vacuoles and/or lysosomal compartments where it is degraded, whereupon the exposed viral genome can trigger a response via TLR signaling. The second phase appears to correspond to the actual infection of the DC, which requires fusion of the viral membrane and delivery of the capsid into the cytoplasm at which stage the viral features would come under the possible surveillance of the cytosolic RIG-I-like signaling pathway (Turville et al. 2004).

This biphasic process is evident in microarray expression profiles reported by two different research groups. Both studies utilized monocyte-derived dendritic cells (MDDCs) that were infected at an immature stage, as opposed to a mature DC that has already experienced pathogen stimulation, and thereupon changed its phagocytic- and antigen-presentation properties. In the study by Solis et al., MDDCs were infected with primary isolates representing clades A/E, B, and C, and expression changes assessed at 2, 6, 24, and 72 h after synchronous infection. Compared to time-matched mocks, the larger number of expression changes occurred at 2 and 6 hpi, involving genes in transcription, signal transduction, cell proliferation, and the immune response. The expression changes implicate increased NFκB activity along with up-regulation of inflammatory genes such as IL1A, IL1B, IL6, and INDO. The number of DEGs declined at 24 h and then showed resurgence at 72 hpi at which time there was still indication of up-regulated NFκB actively as well as increased expression of pro-apoptotic factors. Direct comparison to other studies is difficult due to the use of a custom array design. While there does appear to be some PAMP-induced signaling at early time, the gene expression features do not appear indicative of robust TLR signaling as might be typified by LPS-treatment (Napolitani et al. 2005). Interestingly, these investigators reported the only appreciable differences between the different cIades occurred at 72 hpi, when viral gene products are beginning to reprogram the infected cell.

A similar study design was employed by Harman et al. who characterized the synchronous infection of MDDCs with the R5-tropic, lab-adapted strain HIV-1_{BaL} and characterized the differential gene expression at 6, 24, and 48 hpi. In addition, they included aldrithiol-inactivated virus as a comparator, providing a means to distinguish viral sensing events versus changes attributable to HIV-1 infection. This model system also showed greater expression changes at early and later time points and an apparent lull at 24 h. With the inactivated virus, the temporal pattern reflected just the initial 6 hpi episode of transcriptional changes. The high degree of overlap seen in the 6 hpi DEGs for live and inactivated virus makes it clear that this early time point represents responses prior to viral membrane fusion and entry of the capsid to the cytosol. Among the processes up-regulated at 6 hpi are genes in endosomal pathways and associated GTPases. Up-regulated immune response genes included ISG15, IRF1, MX1, OAS1, and STAT3. MHC class II genes also increased expression levels. All these features indicate some manner of antiviral activating signal has been transduced during at this early time point, but these

effects have disappeared by 24 hpi. Cells infected with the live virus increase the number of DEGS at 48 hpi, a stage when the cells are beginning genomic integration of proviral DNA: accordingly, a number of up-regulated genes are associated with double-stand DNA repair (XRCC5; XRN1). The authors put particular emphasis on gene expression changes in genes that would result in reduced proteolytic activity in the lysosomal compartments. The up-regulated genes at 48 h also clearly show increased transcriptional activity by NFκB as well as increased expression of a number of genes downstream from innate sensing pathways and/or Type I interferon signaling. The expression analysis at 48 h also shows the up-regulation of MHC II genes has not persisted implying that the maturation process of the MDDC has been altered following HIV-1 cytoplasmic entry and genomic integration.

In summary, the systems level analyses indicate that HIV-1 or SIV infection of CD4 + lympohoid cells triggers no or minimal PAMP sensing, and the first signs of altered gene expression by the host are merely the beginning of a vast reprogramming of the infected CD4 + lymphocyte. This eventually results in necrotic or apoptotic death of the infected cells, but with few hallmarks of inflammation arising from the infected cells themselves. The covert nature of the infection could be due to the sequestering of the viral genome and reverse transcription intermediates within the capsid as it is transported to the nuclear membrane, whereupon the pre-integration dsDNA complex is delivered to the nucleus (Arhel 2010). Once this has occurred, the opportunity for detection by cytoplasmic nucleic acid sensors has passed. No triggering occurs thereafter, since transcripts for viral gene products look like conventional capped, polyadenylated mRNAs. There is more evidence of an early, innate antiviral response upon HIV-1 infection of macrophages and myeloid dendritic cells, which could lead to inflammatory signaling. The interaction of HIV-1 with dendritic cells appears unique inasmuch as initial virus attachment via C-type lectins does not target it exclusively for degradation within lysosomal compartments. Even if HIV does initially trigger TLR signaling in DCs, the ensuing infection of the cell by engulfed, active virus alters the process of DC maturation that TLR signaling would have initiated.

2.6 In vivo Investigations with Nonhuman Primates

2.6.1 Contrast of SIV Infection in Natural Hosts versus Pathogenic SIV Models

In striking contrast to HIV infection in humans and SIV infection in rhesus macaques, nonhuman primates that have been naturally infected with SIV for thousands of years do not progress to AIDS (Chahroudi et al. 2012). SIV infection in these natural hosts produces high viral loads, but is nonpathogenic with animals maintaining healthy CD4 + T cell counts (Apetrei et al. 2011; Silvestri et al. 2003). This is in contrast to progressive HIV/SIV infection, where high viral load

leads to loss of CD4 + T cells, immune dysfunction, and progression to AIDS. SIV infection in natural hosts also differs from HIV/SIV infection in rare elite controllers that do not progress to AIDS by maintaining durable control of viral replication at very low levels (Deeks and Walker 2007). Contrasting the mechanisms contributing to protection from AIDS in natural hosts to mechanisms driving progression to AIDS in pathogenic SIV models could lead to new insights for HIV therapy or prevention.

Natural hosts currently under study include sooty mangabeys (SM) and African green monkeys (AGM) (Chahroudi et al. 2012). Like pathogenic HIV/SIV infections, SIV infection in these species leads to high viremia, early loss of mucosal CD4 + T cells, and high levels of immune activation during acute infection (Silvestri et al. 2003). However, a key distinguishing feature in natural hosts is that the initial immune activation that occurs after infection resolves within about 4–8 weeks despite ongoing viral replication, whereas progressive SIV infection in macaques is characterized by unresolved chronic immune activation that leads to impairment in immune function and CD4 decline. Thus, rapid resolution of immune activation in natural hosts may be the key factor that allows these animals to avoid progression to AIDS. Understanding the mechanisms underlying the resolution of immune activation in natural hosts is currently a subject of intense study by systems biology approaches that may lead to identifying new agents that can suppress chronic immune activation in HIV infection.

2.6.2 Functional Genomics and Immunological Characterization of $SIV_{agm.sab}$ in African Green Monkeys Versus Pigtail Macaques

Our own investigations into the host response in natural SIV infections compared the African Green Monkeys (*Chlorocebus sabeus*, AGMs) versus the Asian macaque species *Macaca nemestrina* (pigtail macaques, PTs) after intravenous infection with the same inoculum of $SIV_{agm.sab92018}$ (Favre et al. 2009; Lederer et al. 2009). While strains of SIV_{agm} are broadly disseminated in AGM populations on the African continent, they are absent from the subspecies resident in the Caribbean islands. Experimental infections with this strain conducted in AGMs derived from these Caribbean populations have been well characterized, with no evident pathogenesis, despite high viral loads at both the acute peak and chronic stages of the infection (Diop et al. 2000; Pandrea et al. 2006). Animals experience transient CD4 + lymphocyte depletion in the gut; however, longer term these populations are restored; CD4 + T cell numbers in other compartments are unaffected by the infection (Pandrea et al. 2007). This same strain in *M. nemestrina* results in viral replication kinetics and viral load levels similar to the natural hosts; however, this species experiences a rapid decline of CD4 + T cells in all compartments (Goldstein et al. 2005). Early after infection, PTs establish a persistent slate of immune activation and many animals succumb to AIDS-like symptoms.

For a detailed immunological and functional genomics study, we performed a longitudinal analysis with four animals of each species, with blood, lymph node, and colon samples obtained at -14, +10, and +45 dpi. Ratiometric array measurements were obtained for the post-infection samples, where the expression levels in an animal's tissue were compared to the individually matched baseline sample from day-14. Both species exhibited robust gene expression signatures post-infection; this was especially pronounced in the lymph nodes (LNs) where $\sim$2,500 genes showed a $\geq$ twofold change relative to baseline regardless of species or time post-challenge. Despite this nominal equivalence in the number of regulated transcripts, statistical comparisons at each time point showed the greatest number of differences between species for the day 10 LNs (610 genes). The most prominent functional categories identified by gene ontology analysis were immune responses and cell death. Subsets within these broader categories did not reveal a species bias towards enhanced or suppressed FAS-mediated apoptosis. However, the PTs clearly exhibit greater up-regulation of genes involved in caspase activation, DNA damage, and oxidative stress. In the pathogenic context of the pigtail macaques, the genes in the immune response category distinctly show increased expression; these implicate a Th1 response, cytotoxic T cell activity, and Ifnγ signaling. In contrast, AGMs showed up-regulation of the anti-inflammatory cytokine IL10 and the anti-inflammatory regulator NLRP3, with overall expression changes implicating a more active control of the inflammatory response and a shift to homeostasis of the lymphoid compartment.

In assessing aspects of the innate host response, scrutiny of the Type I interferon genes reveals a stark contrast, with substantially increased expression of these transcripts in the lymph nodes of AGMs on day 10 post-challenge versus significantly lower levels in this same compartment on 45 dpi. (cf. Fig. 3a). The AGM lymph nodes do not show this pattern of temporal regulation for interferon-γ (IFNG), nor for the Type I or Type II interferon receptors. The lymph nodes from the infected pigtail macaques do not show this consistent up-regulation at the earlier time point, although as noted earlier the PTs exhibit great expression of IFNG than do the AGMs at 10 dpi. Moreover, this pattern does not recur in the expression patterns for blood or colon for either species, highlighting the unique kinetics and localization in the AGMs. However, as shown in Fig. 3b, in examining the Type I interferon-stimulated genes (ISGs) in the tissues of the infected AGMs, we do observe similar kinetics with generally elevated expression levels on day 10 and a decline by 45 dpi; the pattern appears in lymph nodes, blood, and is particularly conspicuous in colon. On day 10, the expression ratios of these genes in the natural hosts are comparable to (for blood), or greater than (for colon and lymph node) the observations for the pathogenic context. The PTs do show increased expression in these ISGs relative to their pre-challenge state, but this elevated expression level persists even after the viral load has declined to set point at 45 dpi.

The observed Ifnα plasma levels for both AGMs and PTs are in accord with these observations. For AGMs, plasma Ifnα spikes at 10 dpi and then returns to baseline, whereas with PTs the levels peak with similar kinetics immediately

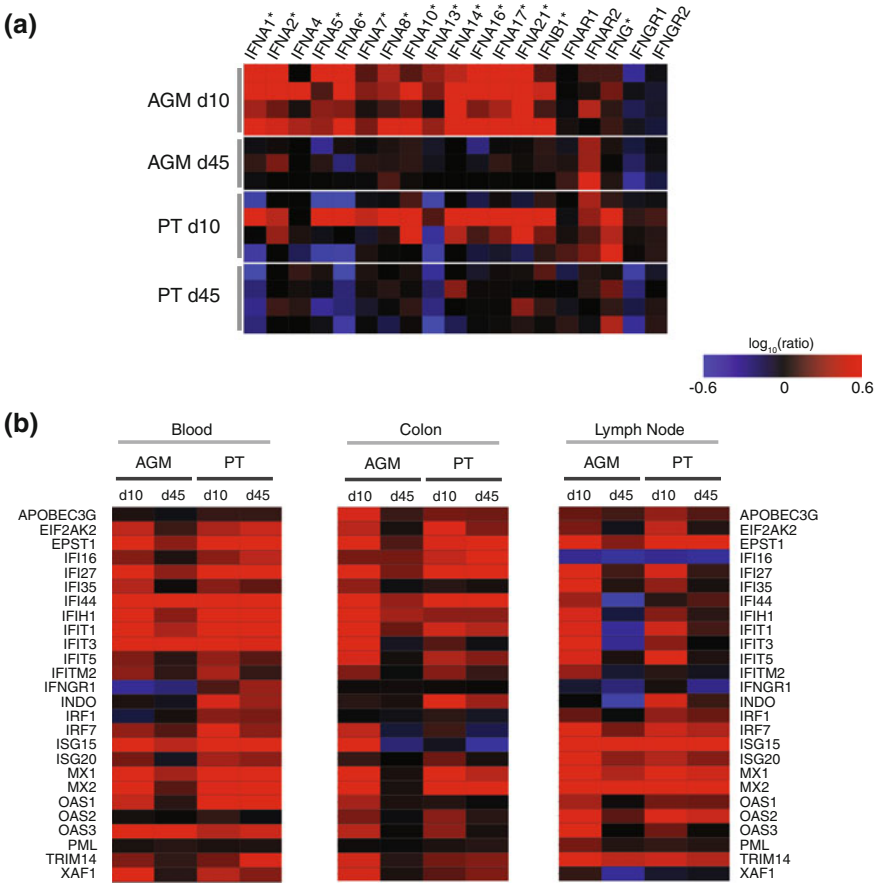


Fig. 3 Early induction of Type I interferon in African green monkeys following infection with SIVagm. Heatmap representations showing relative gene expression ratios for African green monkeys (AGMs) and pigtail macaques (PTs) following intravenous challenge with $SIV_{agm.sab92018}$. Expression ratios are expressed relative to baseline samples collected from the animals 2 weeks prior to infection. **a** Relative gene expression levels for Type I and Type II interferons and corresponding receptors, from lymph node samples at 10 and 45 days post-infection. Results are organized by species and day, with individual animal replicates shown. Gene denoted with * had p-values < 0.05 in a one-sided *t* test comparing AGM expression levels on the two time points. In addition, a Fishers exact test for this high proportion of significance outcomes returns a *p* value < 0.001. **b** Temporal profiles of indicated interferon-stimulated genes in blood, colon, and lymph node samples for infected AGMs and PTs. Shown are the weighted averages for the animals at the indicated times. Both panels are rendered as $\log_{10}$(ratios), with saturation in the color scheme at $\pm$ fourfold. See ref. (Lederer et al. 2009) for experimental details

post-challenge then decline only slightly to plateau at an elevated state through the rest of the time course. The inference from the expression data in Fig. 3a is that the Type I interferon primarily originates within the lymph nodes for AGMs. This is

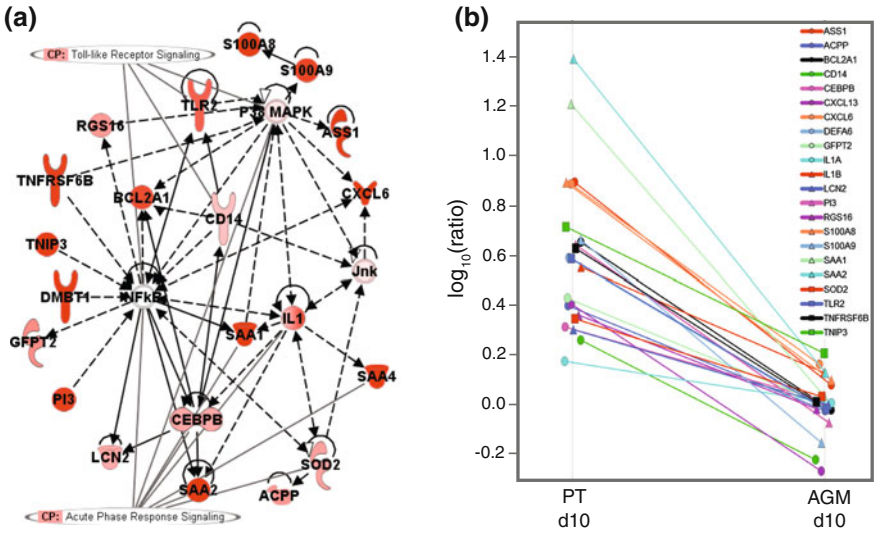


Fig. 4 Expression of genes associated with acute phase response/NFκB signaling in colon of pigtail macaques on day 10 following infection with $SIV_{agm.sab92018}$. **a** Network analysis showing connections between genes related to transcription factor NFkB, overlaid with expression data from in silico averages for PTs at day 10. **b** Line drawing showing differences in expression values for the networked genes, illustrating the elevated levels of these inflammatory genes in PTs vs. quiescent character for AGMs at the same time point. Panel A is reproduced with permission from Lederer et al. 2009

supported by other published studies that determined that plasmacytoid dendritic cells increase in the lymph nodes of AGMs during the acute phase of SIV infection and that during this interval the pDCs have matured to a state highly competent for Ifnα production (Diop et al. 2008). The situation for the PTs is more cryptic inasmuch no Type I interferon transcripts appear up-regulated in any compartment, in contradistinction to the persistently elevated protein abundance in plasma. This may be the consequence of increased protein production attendant to small increase in transcript abundance, with the latter changes falling within the noise threshold of the microarray experiments.

It also bears noting that the early Type I interferon signaling in the pigtail macaques happens in concert with the up-regulation of many acute phase and inflammatory response genes. As described earlier, in the lymph nodes many of these expression changes conform to the developing Th1 response. This inflammatory response in PTs is also starkly evident in the statistical comparison of the transcript levels in the colon at 10 dpi; 33 such acute phase/inflammatory genes are up-regulated at this stage and the majority are sustained at these increased levels to 45 dpi. Network analysis of these genes finds many of them associated with NFκB signaling (Fig. 4), and the presence of TLR2 and CD14 suggests a role for myeloid cells in this response.

Two other systems level, functional genomics investigations of natural infection models have yielded very similar outcomes. The study of Jacquelin et al. contrasted the responses of AGMs intravenously infected with $SIV_{agm.sab92018}$ vs. the course for rhesus macaques (RMs) challenged IV with SIV_{mac251} (Jacquelin et al. 2009). These investigators were able to obtain expression results on blood and lymph node samples taken as early as 1 day post-challenge, when the infected AGMs already showed strong up-regulation of a large number of ISGs typically associated with Type I interferon response. However in the RMs, increased expression of this category of genes was delayed until the ensuing time point at 6 dpi. Using a highly sensitive functional assay for Type I interferon, the authors were able to show AGMs exhibited an initial small peak in plasma levels at 2 dpi, followed by a second much higher peak at 9 dpi before returning to baseline within days. Rapid control of the innate response in the natural infection of sooty mangabeys inoculated with SIV_{smm} was also the conclusion reached by Bosinger et al. using an animal model that contrasted the natural infection versus pathogenic challenges of rhesus macaques with either the same inoculum of SIV_{smm} or with highly virulent SIV_{mac239} (Bosinger et al. 2009). Unlike SIV-infected PTs, infected rhesus macaques do not exhibit chronically high plasma levels of Ifnα, and instead manifest one peak at $\sim$10 days post-challenge. As with the PTS, the origin of the initial Type I interferon is less evident, and the underlying expression changes may proceed with differing kinetics or compartmentalization than were accessed in these studies. Nonetheless in these experiments, as in the aforementioned studies with pigtail macaques, animals continue to show persistent up-regulation of ISGs and acute inflammatory response genes likely driven by NFκB transcriptional control. We observed a similar pattern in our own longitudinal comparison of gene expression changes in the blood of rhesus macaques infected with SIV_{mac251} (Palermo et al. 2011).

3 Further Exploiting Systems Level Investigations in NHP Models for AIDS Pathogenesis and Immunity

An effective HIV vaccine or therapeutic will need to induce multiple immune defenses that can synergistically interfere with the ability of the virus to gain or maintain a foothold in the host. For therapy, systems biology approaches are currently being employed to define the mechanisms underlying progression versus lack of progression to AIDS in natural hosts versus pathogenic SIV models and in HIV/SIV progressors versus elite controllers. The results from these experiments will likely reveal novel immune targets or pathways that could be manipulated to either dampen immune activation/inflammation (natural hosts) or enhance immune control of viral replication (elite controllers). Referring again to Fig. 1, and the differing scales of biological systems we have examined here, the in vitro experiments suggest that the infection of CD4 + lymphocytes is not the driver of

the protracted inflammatory signaling that results in bystander T cell death and immune exhaustion for pathogenic infections of HIV/SIV. Certainly there are inflammatory consequences from engagement of the adaptive immune system in targeting infected CD4 + lymphocytes. But is this persistent engagement a sufficient explanation for the chronic immune activation since the immune exhaustion is not a consequence of other chronic viral infections?. From the AGM/PT comparison, there is the indication that regulation of the Type I interferon response particularly from DCs differs between the species. Therefore, it may be worth examining if the difference between natural versus pathogenic infections rests on distinctions in responses of the antigen-presenting cells that can initiate and orchestrate the immune response (i.e. DCs and macrophages), with the further refinement whether such differences are inherent or particular to the infected APCs. Recent reports concerning reduced levels of Tnfα production by monocytes in SIV-infected sooty mangabeys bear on this question (Mir et al. 2012).

A systems approach may be especially important for interrogating the pivotal host–virus interactions and immune responses at the initial mucosal site of HIV/SIV exposure. HIV infection is spread primarily via vaginal or rectal sexual transmission, and the gut is the primary reservoir of virus that persists even during therapy with potent antiretroviral drugs. It is now widely accepted that protection from mucosal infection or prevention of immune dysfunction in the gut may be essential for an effective vaccine or therapeutic. Indeed, a critical advantage in natural hosts that do not progress to AIDS is maintenance of mucosal immunity and gut integrity. Experiments in nonhuman primates have shown that immune responses to vaccination and correlates of viral control in the SIV model can differ substantially between the mucosa and blood (Fuller et al. 2012; Loudon et al. 2010). However, to date we still have only a limited understanding of what happens in the mucosa during the critical earliest stages of infection when the virus is trying to gain its first foothold into the host or its interactions with gut cells once infection is established. An inherent obstacle, especially in humans, is access to sufficient mucosal samples. This has limited our ability to adequately interrogate this compartment by traditional methods. In contrast, systems biology enables investigation of a broad range of immune functions and pathways with small samples. Studies are now underway in nonhuman primates to study the very first responses to SIV infection in the mucosa. The results from these studies may prove especially relevant to efforts aimed at developing new therapeutic drugs, microbicides, and vaccines that can eliminate or block the virus in the mucosa.

The RV144 trial and vaccine successes in nonhuman primates suggest we may already have some of the right tools to prevent HIV infection. What is now needed are new strategies, such as adjuvants, that can enhance the efficacy of these promising vaccines. However, identifying an adjuvant that increases host defenses without producing undue inflammatory responses, which would benefit the virus, is a tricky balancing act. Use of systems biology approaches to characterize the host response to vaccination and to contrast the differences between vaccinated individuals that are protected versus those that fail to be protected are needed to identify adjuvants that can walk that line.

References

Aderem A, Adkins JN, Ansong C, Galagan J, Kaiser S., Korth MJ, Law GL, McDermott JG, Proll SC, Rosenberger C, Schoolnik G, Katze MG (2011) A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. mBio 2: e00325–10

Agy MB, Wambach M, Foy K, Katze MG (1990) Expression of cellular genes in CD4 positive lymphoid cells infected by the human immunodeficiency virus, HIV-1: evidence for a host protein synthesis shut-off induced by cellular mRNA degradation. Virology 177:251–8

Agy MB, Frumkin LR, Corey L, Coombs RW, Wolinsky SM, Koehler J, Morton WR, Katze MG (1992) Infection of Macaca nemestrina by human immunodeficiency virus type-1. Science 257:103–6

Apetrei C, Sumpter B, Souquiere S, Chahroudi A, Makuwa M, Reed P, Ribeiro RM, Pandrea I, Roques P, Silvestri G (2011) Immunovirological analyses of chronically simian immunodeficiency virus SIVmnd-1- and SIVmnd-2-infected mandrills (Mandrillus sphinx). J Virol 85:13077–87

Arhel N (2010) Revisiting HIV-1 uncoating. Retrovirology 7:96

Barnett SW, Burke B, Sun Y, Kan E, Legg H, Lian Y, Bost K, Zhou F, Goodsell A, Zur Megede J, Polo J, Donnelly J, Ulmer J, Otten GR, Miller CJ, Vajdy M, Srivastava IK (2010) Antibody-mediated protection against mucosal simian-human immunodeficiency virus challenge of macaques immunized with alphavirus replicon particles and boosted with trimeric envelope glycoprotein in MF59 adjuvant. J Virol 84:5975–85

Barouch DH, Liu J, Li H, Maxfield LF, Abbink P, Lynch DM, Iampietro MJ, SanMiguel A, Seaman MS, Ferrari G, Forthal DN, Ourmanov I, Hirsch VM, Carville A, Mansfield KG, Stablein D, Pau MG, Schuitemaker H, Sadoff JC, Billings EA, Rao M, Robb ML, Kim JH, Marovich MA, Goudsmit J, Michael NL (2012) Vaccine protection against acquisition of neutralization-resistant SIV challenges in rhesus monkeys. Nature 482:89–93

Baskin CR, Bielefeldt-Ohmann H, Garcia-Sastre A, Tumpey TM, Van HN, Carter VS, Thomas MJ, Proll S, Solorzano A, Billharz R, Fornek JL, Thomas S, Chen CH, Clark EA, Murali-Krishna K, Katze MG (2007) Functional genomic and serological analysis of the protective immune response resulting from vaccination of macaques with an NS1-truncated influenza virus. J Virol 81:11817–11827

Batten CJ, De Rose R, Wilson KM, Agy MB, Chea S, Stratov I, Montefiori DC, Kent SJ (2006) Comparative evaluation of simian, simian-human, and human immunodeficiency virus infections in the pigtail macaque (Macaca nemestrina) model. AIDS Res Hum Retroviruses 22:580–8

Belyakov IM, Hel Z, Kelsall B, Kuznetsov VA, Ahlers JD, Nacsa J, Watkins DI, Allen TM, Sette A, Altman J, Woodward R, Markham PD, Clements JD, Franchini G, Strober W, Berzofsky JA (2001) Mucosal AIDS vaccine reduces disease and viral load in gut reservoir and blood after mucosal infection of macaques. Nat Med 7:1320–6

Bosinger SE, Li Q, Gordon SN, Klatt NR, Duan L, Xu L, Francella N, Sidahmed A, Smith AJ, Cramer EM, Zeng M, Masopust D, Carlis JV, Ran L, Vanderford TH, Paiardini M, Isett RB, Baldwin DA, Else JG, Staprans SI, Silvestri G, Haase AT, Kelvin DJ (2009) Global genomic analysis reveals rapid control of a robust innate response in SIV-infected sooty mangabeys. J Clin Invest 119:3556–72

Brugger B, Krautkramer E, Tibroni N, Munte CE, Rauch S, Leibrecht I, Glass B, Breuer S, Geyer M, Krausslich HG, Kalbitzer HR, Wieland FT, Fackler OT (2007) Human immunodeficiency virus type 1 Nef protein modulates the lipid composition of virions and host cell membrane microdomains. Retrovirology 4:70

Chahroudi A, Bosinger SE, Vanderford TH, Paiardini M, Silvestri G (2012) Natural SIV hosts: showing AIDS the door. Science 335:1188–93

Chan EY, Qian WJ, Diamond DL, Liu T, Gritsenko MA, Monroe ME, Camp DG 2nd, Smith RD, Katze MG (2007) Quantitative analysis of human immunodeficiency virus type 1-infected

CD4+ cell proteome: dysregulated cell cycle progression and nuclear transport coincide with robust virus production. J Virol 81:7571–83

Chan EY, Sutton JN, Jacobs JM, Bondarenko A, Smith RD, Katze MG (2009) Dynamic host energetics and cytoskeletal proteomes in human immunodeficiency virus type 1-infected human primary CD4 cells: analysis by multiplexed label-free mass spectrometry. J Virol 83:9283–95

Chang ST, Sova P, Peng X et al (2011) Next-generation sequencing reveals HIV-1-mediated suppression of T cell activation and RNA processing and regulation of noncoding RNA expression in a CD4 + T Cell line. mBio 2(5):doi:10.1128/mBio.00134-11

Chiu YH, Macmillan JB, Chen ZJ (2009) RNA polymerase III detects cytosolic DNA and induces type I interferons through the RIG-I pathway. Cell 138:576–91

Cilloniz C, Shinya K, Peng X, Korth MJ, Proll SC, Aicher LD, Carter VS, Chang JH, Kobasa D, Feldmann F, Strong JE, Feldmann H, Kawaoka Y, Katze MG (2009) Lethal influenza virus infection in macaques is associated with early dysregulation of inflammatory related genes. PLoS Pathog 5:e1000604

Cilloniz C, Ebihara H, Ni C, Neumann G, Korth MJ, Kelly SM, Kawaoka Y, Feldmann H, Katze MG (2011) Functional genomics reveals the induction of inflammatory response and metalloproteinase gene expression during lethal Ebola virus infection. J Virol 85:9060–8

Corbeil J, Sheeter D, Genini D, Rought S, Leoni L, Du P, Ferguson M, Masys DR, Welsh JB, Fink JL, Sasik R, Huang D, Drenkow J, Richman DD, Gingeras T (2001) Temporal gene regulation during HIV-1 infection of human CD4+ T cells. Genome Res 11:1198–1204

Daniel MD, Kirchhoff F, Czajak SC, Sehgal PK, Desrosiers RC (1992) Protective effects of a live attenuated SIV vaccine with a deletion in the nef gene. Science 258:1938–41

de Lang A, Baas T, Teal T, Leijten LM, Rain B, Osterhaus AD, Haagmans BL, Katze MG (2007) Functional genomics highlights differential induction of antiviral pathways in the lungs of SARS-CoV-infected macaques. PLoS Pathog 3:e112

Deeks SG, Walker BD (2007) Human immunodeficiency virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy. Immunity 27:406–16

Diop OM, Gueye A, Dias-Tavares M, Kornfeld C, Faye A, Ave P, Huerre M, Corbet S, Barre-Sinoussi F, Muller-Trutwin MC (2000) High levels of viral replication during primary simian immunodeficiency virus SIVagm infection are rapidly and strongly controlled in African green monkeys. J Virol 74:7538–47

Diop OM, Ploquin MJ, Mortara L, Faye A, Jacquelin B, Kunkel D, Lebon P, Butor C, Hosmalin A, Barre-Sinoussi F, Muller-Trutwin MC (2008) Plasmacytoid dendritic cell dynamics and alpha interferon production during Simian immunodeficiency virus infection with a nonpathogenic outcome. J Virol 82:5145–52

Favre D, Lederer S, Kanwar B, Ma ZM, Proll S, Kasakow Z, Mold J, Swainson L, Barbour JD, Baskin CR, Palermo R, Pandrea I, Miller CJ, Katze MG, McCune JM (2009) Critical loss of the balance between Th17 and T regulatory cell populations in pathogenic SIV infection. PLoS Pathog 5:e1000295

Frumkin LR, Agy MB, Coombs RW, Panther L, Morton WR, Koehler J, Florey MJ, Dragavon J, Schmidt A, Katze MG et al (1993) Acute infection of Macaca nemestrina by human immunodeficiency virus type 1. Virology 195:422–31

Fuller DH, Rajakumar PA, Wilson LA, Trichel AM, Fuller JT, Shipley T, Wu MS, Weis K, Rinaldo CR, Haynes JR, Murphey-Corb M (2002) Induction of mucosal protection against primary, heterologous simian immunodeficiency virus by a DNA vaccine. J Virol 76:3309–17

Fuller DH, Rajakumar P, Che JW, Narendran A, Nyaundi J, Michael H, Yager EJ, Stagnar C, Wahlberg B, Taber R, Haynes JR, Cook FC, Ertl P, Tite J, Amedee AM, Murphey-Corb M (2012) Therapeutic DNA vaccine induces broad T cell responses in the gut and sustained protection from viral rebound and AIDS in SIV-infected rhesus macaques. PLoS One 7:e33715

Gale M Jr, Katze MG (1998) Molecular mechanisms of interferon resistance mediated by viral-directed inhibition of PKR, the interferon-induced protein kinase. Pharmacol Ther 78:29–46

Gaucher D, Therrien R, Kettaf N, Angermann BR, Boucher G, Filali-Mouhim A, Moser JM, Mehta RS, Drake DR 3rd, Castro E, Akondy R, Rinfret A, Yassine-Diab B, Said EA, Chouikh Y, Cameron MJ, Clum R, Kelvin D, Somogyi R, Greller LD, Balderas RS, Wilkinson P, Pantaleo G, Tartaglia J, Haddad EK, Sekaly RP (2008) Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. J Exp Med 205:3119–31

Geiss GK, Bumgarner RE, An MC, Agy MB, van 't Wout AB, Hammersmark E, Carter VS, Upchurch D, Mullins JI, Katze MG (2000) Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. Virology 266:8–16

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu YS, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers YH, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csuros M, Glasscock J, Harris RA, Havlak P, Jackson AR, Jiang H et al (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234

Goldstein S, Ourmanov I, Brown CR, Plishka R, Buckler-White A, Byrum R, Hirsch VM (2005) Plateau levels of viremia correlate with the degree of CD4+-T-cell loss in simian immunodeficiency virus SIVagm-infected pigtailed macaques: variable pathogenicity of natural SIVagm isolates. J Virol 79:5153–62

Gupta A, Nagilla P, Le HS, Bunney C, Zych C, Thalamuthu A, Bar-Joseph Z, Mathavan S, Ayyavoo V (2011) Comparative expression profile of miRNA and mRNA in primary peripheral blood mononuclear cells infected with human immunodeficiency virus (HIV-1). PLoS One 6:e22730

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223–7

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol. 19:1720–1730

Henriet S, Mercenne G, Bernacchi S, Paillart JC, Marquet R (2009) Tumultuous relationship between the human immunodeficiency virus type 1 viral infectivity factor (Vif) and the human APOBEC-3G and APOBEC-3F restriction factors. Microbiol Mol Biol Rev 73:211–32

Imbeault M, Ouellet M, Tremblay MJ (2009) Microarray study reveals that HIV-1 induces rapid type-I interferon-dependent p53 mRNA up-regulation in human primary CD4+ T cells. Retrovirology 6:5

Jacquelin B, Mayau V, Targat B, Liovat AS, Kunkel D, Petitjean G, Dillies MA, Roques P, Butor C, Silvestri G, Giavedoni LD, Lebon P, Barre-Sinoussi F, Benecke A, Muller-Trutwin MC (2009) Nonpathogenic SIV infection of African green monkeys induces a strong but rapidly controlled type I IFN response. J Clin Invest 119:3544–55

Katze MG (1995) Regulation of the interferon-induced PKR: can viruses cope? Trends Microbiol 3:75–8

Katze MG (2002) Interferon, PKR, virology, and genomics: what is past and what is next in the new millennium? J Interferon Cytokine Res 22:283–6

Katze MG, Agy MB (1990) Regulation of viral and cellular RNA turnover in cells infected by eukaryotic viruses including HIV-1. Enzyme 44:332–46

Katze MG, He Y, Gale M Jr (2002) Viruses and interferon: a fight for supremacy. Nat Rev Immunol 2:675–87

Kawai T, Akira S (2010) The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. Nat Immunol 11:373–84

Kobasa D, Jones SM, Shinya K, Kash JC, Copps J, Ebihara H, Hatta Y, Kim JH, Halfmann P, Hatta M, Feldmann F, Alimonti JB, Fernando L, Li Y, Katze MG, Feldmann H, Kawaoka Y (2007) Aberrant innate immune response in lethal infection of macaques with the 1918 influenza virus. Nature 445:319–23

Korth MJ, Kash JC, Furlong JC, Katze MG (2005) Virus infection and the interferon response: a global view through functional genomics. Methods Mol Med 116:37–55

Lai L, Kwa S, Kozlowski PA, Montefiori DC, Ferrari G, Johnson WE, Hirsch V, Villinger F, Chennareddi L, Earl PL, Moss B, Amara RR, Robinson HL (2011) Prevention of infection by a granulocyte-macrophage colony-stimulating factor co-expressing DNA/modified vaccinia Ankara simian immunodeficiency virus vaccine. J Infect Dis 204:164–73

Lederer S, Favre D, Walters KA, Proll S, Kanwar B, Kasakow Z, Baskin CR, Palermo R, McCune JM, Katze MG (2009) Transcriptional profiling in pathogenic and non-pathogenic SIV infections reveals significant distinctions in kinetics and tissue compartmentalization. PLoS Pathog 5:e1000296, 10.1371/journal.ppat.1000296

Li Y, Chan EY, Katze MG (2007) Functional genomics analyses of differential macaque peripheral blood mononuclear cell infections by human immunodeficiency virus-1 and simian immunodeficiency virus. Virology 366:137–49

Loudon PT, Yager EJ, Lynch DT, Narendran A, Stagnar C, Franchini AM, Fuller JT, White PA, Nyuandi J, Wiley CA, Murphey-Corb M, Fuller DH (2010) GM-CSF increases mucosal and systemic immunogenicity of an H1N1 influenza DNA vaccine administered into the epidermis of non-human primates. PloS one 5:e11021

Magness CL, Fellin PC, Thomas MJ, Korth MJ, Agy MB, Proll SC, Fitzgibbon M, Scherer CA, Miner DG, Katze MG, Iadonato SP (2005) Analysis of the Macaca mulatta transcriptome and the sequence divergence between Macaca and human. Genome Biol 6:R60

Manrique M, Kozlowski PA, Cobo-Molinos A, Wang SW, Wilson RL, Montefiori DC, Mansfield KG, Carville A, Aldovini A (2011) Long-term control of simian immunodeficiency Virus mac251 viremia to undetectable levels in half of infected female rhesus macaques nasally vaccinated with simian immunodeficiency virus DNA/recombinant modified vaccinia virus Ankara. J Immunol 186:3581–93

McDermott JE, Shankaran H, Eisfeld AJ, Belisle SE, Neuman G, Li C, McWeeney S, Sabourin C, Kawaoka Y, Katze MG, Waters KM (2011) Conserved host response to highly pathogenic avian influenza virus infection in human cell culture, mouse and macaque model systems. BMC Syst Biol 5:190

Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10:155–9

Mir KD, Bosinger SE, Gasper M, Ho O, Else JG, Brenchley JM, Kelvin DJ, Silvestri G, Hu SL, Sodora DL (2012) SIV-induced alterations in monocyte TNF-alpha production contribute to reduced immune activation in sooty mangabeys. J Virol 86:7605–7615

Moresco EM, LaVine D, Beutler B (2011) Toll-like receptors. Curr Biol 21:R488–93

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–8

Nakaya HI, Wrammert J, Lee EK, Racioppi L, Marie-Kunze S, Haining WN, Means AR, Kasturi SP, Khan N, Li GM, McCausland M, Kanchan V, Kokko KE, Li S, Elbein R, Mehta AK, Aderem A, Subbarao K, Ahmed R, Pulendran B (2011) Systems biology of vaccination for seasonal influenza in humans. Nat Immunol 12:786–95

Napolitani G, Rinaldi A, Bertoni F, Sallusto F, Lanzavecchia A (2005) Selected Toll-like receptor agonist combinations synergistically trigger a T helper type 1-polarizing program in dendritic cells. Nat Immunol 6:769–76

Navare AT, Sova P, Purdy DE, Weiss JM, Wolf-Yadlin A, Korth MJ, Chang ST, Proll SC, Jahan TA, Krasnoselsky AL, Palermo RE, Katze MG (2012) Quantitative proteomic analysis of HIV-1 infected CD4+ T cells reveals an early host response in important biological pathways: Protein synthesis, cell proliferation, and T-cell activation. Virology 429:37–46

Palermo RE, Patterson LJ, Aicher LD, Korth MJ, Robert-Guroff M, Katze MG (2011) Genomic analysis reveals pre- and postchallenge differences in a rhesus macaque AIDS vaccine trial: insights into mechanisms of vaccine efficacy. J Virol 85:1099–116

Pandrea I, Apetrei C, Dufour J, Dillon N, Barbercheck J, Metzger M, Jacquelin B, Bohm R, Marx PA, Barre-Sinoussi F, Hirsch VM, Muller-Trutwin MC, Lackner AA, Veazey RS (2006) Simian immunodeficiency virus SIVagm.sab infection of Caribbean African green monkeys: a new model for the study of SIV pathogenesis in natural hosts. J Virol 80:4858–67

Pandrea IV, Gautam R, Ribeiro RM, Brenchley JM, Butler IF, Pattison M, Rasmussen T, Marx PA, Silvestri G, Lackner AA, Perelson AS, Douek DC, Veazey RS, Apetrei C (2007) Acute loss of intestinal CD4+ T cells is not predictive of simian immunodeficiency virus virulence. J Immunol 179:3035–3046

Pang KC, Dinger ME, Mercer TR, Malquori L, Grimmond SM, Chen W, Mattick JS (2009) Genome-wide identification of long noncoding RNAs in CD8+ T cells. J Immunol 182: 7738–48

Patterson LJ, Malkevitch N, Venzon D, Pinczewski J, Gomez-Roman VR, Wang L, Kalyanaraman VS, Markham PD, Robey FA, Robert-Guroff M (2004) Protection against mucosal simian immunodeficiency virus SIV(mac251) challenge by using replicating adenovirus-SIV multigene vaccine priming and subunit boosting. J Virol 78:2212–2221

Peng X, Gralinski L, Armour CD, Ferris MT, Thomas MJ, Proll S, Bradel-Tretheway BG, Korth MJ, Castle JC, Biery MC, Bouzek HK, Haynor DR, Frieman MB, Heise M, Raymond CK, Baric RS, Katze MG (2010) Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling. MBio 1:e00206–10

Querec TD, Akondy RS, Lee EK, Cao W, Nakaya HI, Teuwen D, Pirani A, Gernert K, Deng J, Marzolf B, Kennedy K, Wu H, Bennouna S, Oluoch H, Miller J, Vencio RZ, Mulligan M, Aderem A, Ahmed R, Pulendran B (2009) Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol 10:116–25

Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, Premsri N, Namwat C, de Souza M, Adams E, Benenson M, Gurunathan S, Tartaglia J, McNeil JG, Francis DP, Stablein D, Birx DL, Chunsuttiwat S, Khamboonruang C, Thongcharoen P, Robb ML, Michael NL, Kunasol P, Kim JH (2009) Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med 361:2209–20

Safronetz D, Rockx B, Feldmann F, Belisle SE, Palermo RE, Brining D, Gardner D, Proll SC, Marzi A, Tsuda Y, Lacasse RA, Kercher L, York A, Korth MJ, Long D, Rosenke R, Shupert WL, Aranda CA, Mattoon JS, Kobasa D, Kobinger G, Li Y, Taubenberger JK, Richt JA, Parnell M, Ebihara H, Kawaoka Y, Katze MG, Feldmann H (2011) Pandemic swine-origin H1N1 influenza A virus isolates show heterogeneous virulence in macaques. J Virol 85:1214–23

Sheehy AM, Gaddis NC, Choi JD, Malim MH (2002) Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature 418:646–50

Silvestri G, Sodora DL, Koup RA, Paiardini M, O'Neil SP, McClure HM, Staprans SI, Feinberg MB (2003) Nonpathogenic SIV infection of sooty mangabeys is characterized by limited bystander immunopathology despite chronic high-level viremia. Immunity 18:441–52

Staprans SI, Barry AP, Silvestri G, Safrit JT, Kozyr N, Sumpter B, Nguyen H, McClure H, Montefiori D, Cohen JI, Feinberg MB (2004) Enhanced SIV replication and accelerated progression to AIDS in macaques primed to mount a CD4 T cell response to the SIV envelope protein. Proc. Natl. Acad. Sci. U.S.A. 101:13026–13031

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–60

Szabo G, Dolganiuc A (2008) The role of plasmacytoid dendritic cell-derived IFN alpha in antiviral immunity. Crit Rev Immunol 28:61–94

Taylor HE, Linde ME, Khatua AK, Popik W, Hildreth JE (2011) Sterol regulatory element-binding protein 2 couples HIV-1 transcription to cholesterol homeostasis and T cell activation. J Virol 85:7699–709

Thomas MJ, Agy MB, Proll SC, Paeper BW, Li Y, Jensen KL, Korth MJ, Katze MG (2006) Functional gene analysis of individual response to challenge of SIVmac239 in M. mulatta PBMC culture. Virology 348:242–252

Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, Yi EC, Dai H, Thorsson V, Eng J, Goodlett D, Berger JP, Gunter B, Linseley PS, Stoughton RB, Aebersold R, Collins SJ, Hanlon WA, Hood LE (2004) Integrated genomic and proteomic analyses of gene expression in Mammalian cells. Mol Cell Proteomics 3:960–969

Tisoncik JR, Belisle SE, Diamond DL, Korth MJ, Katze MG (2009) Is systems biology the key to preventing the next pandemic? Future Virol 4:553–561

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–5

Turville SG, Santos JJ, Frank I, Cameron PU, Wilkinson J, Miranda-Saksena M, Dable J, Stossel H, Romani N, Piatak M Jr, Lifson JD, Pope M, Cunningham AL (2004) Immunodeficiency virus uptake, turnover, and 2-phase transfer in human dendritic cells. Blood 103:2170–9

Vahey MT, Nau ME, Jagodzinski LL, Yalley-Ogunro J, Taubman M, Michael NL, Lewis MG (2002) Impact of viral infection on the gene expression profiles of proliferating normal human peripheral blood mononuclear cells infected with HIV type 1 RF. AIDS Res Hum. Retroviruses 18:179–192

van 't Wout AB, Lehrman GK, Mikheeva SA, O'Keeffe GC, Katze MG, Bumgarner RE, Geiss GK, Mullins JI (2003) Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4(+)-T-cell lines. J Virol 77:1392–402

van 't Wout AB, Swain JV, Schindler M, Rao U, Pathmajeyan MS, Mullins JI, Kirchhoff F (2005) Nef induces multiple genes involved in cholesterol synthesis and uptake in human immunodeficiency virus type 1-infected T cells. J Virol 79:10053–10058

Vazquez N, Greenwell-Wild T, Marinos NJ, Swaim WD, Nares S, Ott DE, Schubert U, Henklein P, Orenstein JM, Sporn MB, Wahl SM (2005) Human immunodeficiency virus type 1-induced macrophage gene expression includes the p21 gene, a target for viral regulation. J Virol 79:4479–91

Venkatachari NJ, Buchanan WG, Ayyavoo V (2008) Human immunodeficiency virus (HIV-1) infection selectively downregulates PD-1 expression in infected cells and protects the cells from early apoptosis in vitro and in vivo. Virology 376:140–53

Wallace JC, Korth MJ, Paeper B, Proll SC, Thomas MJ, Magness CL, Iadonato SP, Nelson C, Katze MG (2007) High-density rhesus macaque oligonucleotide microarray design using early-stage rhesus genome sequence information and human genome annotations. BMC Genomics 8:28

Wilkinson J, Cunningham AL (2006) Mucosal transmission of HIV-1: first stop dendritic cells. Curr Drug Targets 7:1563–9

Woelk CH, Ottones F, Plotkin CR, Du P, Royer CD, Rought SE, Lozach J, Sasik R, Kornbluth RS, Richman DD, Corbeil J (2004) Interferon gene expression following HIV type 1 infection of monocyte-derived macrophages. AIDS Res Hum Retroviruses 20:1210–22

Yin J, Chen MF, Finkel TH (2004) Differential gene expression during HIV-1 infection analyzed by suppression subtractive hybridization. AIDS 18:587–96

Yoneyama M, Fujita T (2009) RNA recognition and signal transduction by RIG-I-like receptors. Immunol Rev 227:54–65

Systems Biology of Vaccination in the Elderly

Sai S. Duraisingham, Nadine Rouphael, Mary M. Cavanagh, Helder I. Nakaya, Jorg J. Goronzy and Bali Pulendran

Abstract Aging population demographics, combined with suboptimal vaccine responses in the elderly, make the improvement of vaccination strategies in the elderly a developing public health issue. The immune system changes with age, with innate and adaptive cell components becoming increasingly dysfunctional. As such, vaccine responses in the elderly are impaired in ways that differ depending on the type of vaccine (e.g., live attenuated, polysaccharide, conjugate, or subunit) and the mediators of protection (e.g., antibody and/or T cell). The rapidly progressing field of systems biology has been shown to be useful in predicting immunogenicity and offering insights into potential mechanisms of protection in young adults. Future application of systems biology to vaccination in the elderly may help to identify gene signatures that predict suboptimal responses and help to identify more accurate correlates of protection. Moreover, the identification of specific defects may be used to target novel vaccination strategies that improve efficacy in elderly populations.

S. S. Duraisingham · H. I. Nakaya · B. Pulendran (✉)
Emory Vaccine Center, Yerkes National Primate Research Center,
Emory University, 954 Gatewood Road, Atlanta, GA 30329, USA
e-mail: bpulend@emory.edu

N. Rouphael
Division of Infectious Diseases, Department of Medicine,
Emory University School of Medicine, Atlanta, GA 30329, USA

M. M. Cavanagh · J. J. Goronzy
Department of Medicine, Stanford University, Stanford, CA 94305, USA

J. J. Goronzy
Department of Medicine, Palo Alto Veteran Administration Health Care System,
Palo Alto, CA 94304, USA

Current Topics in Microbiology and Immunology (2013) 363: 117–142
DOI: 10.1007/82_2012_250

Published Online: 18 August 2012

Contents

1 Introduction

The immune system continuously transforms itself throughout life. Imprints from encounters with infectious organisms accumulate over a lifetime and in parallel, the host environment changes with age, rendering the system increasingly dysfunctional. This immunosenescence is of importance because of evolving population demographics. In industrialized countries, the percentage of individuals over 65 years, a few percent in 1900, will exceed 25% by 2050 (WHO 2002). Infections are a major cause of morbidity in the elderly; vaccinations, previously a cornerstone of public health policies targeting children, are increasingly developed for adults, thus the spectrum of routine and travel adult vaccinations has widened. Despite the success of childhood vaccination in reshaping the infectious landscape of youth, vaccination in older adults has been partly successful at best. Importantly, mechanisms of protection can be very different, ranging from the production of neutralizing antibodies to prevent infection, to cellular immunity to control latent infection. The rapidly progressing field of systems biology offers opportunities to delineate the mechanisms of immunosenescence in the context of vaccine responses, and may help to more accurately predict immunogenicity in the elderly (Nakaya et al. 2012; Pulendran 2009; Pulendran et al. 2010). Such global insights may be used to tailor vaccination strategies specifically to the aging population.

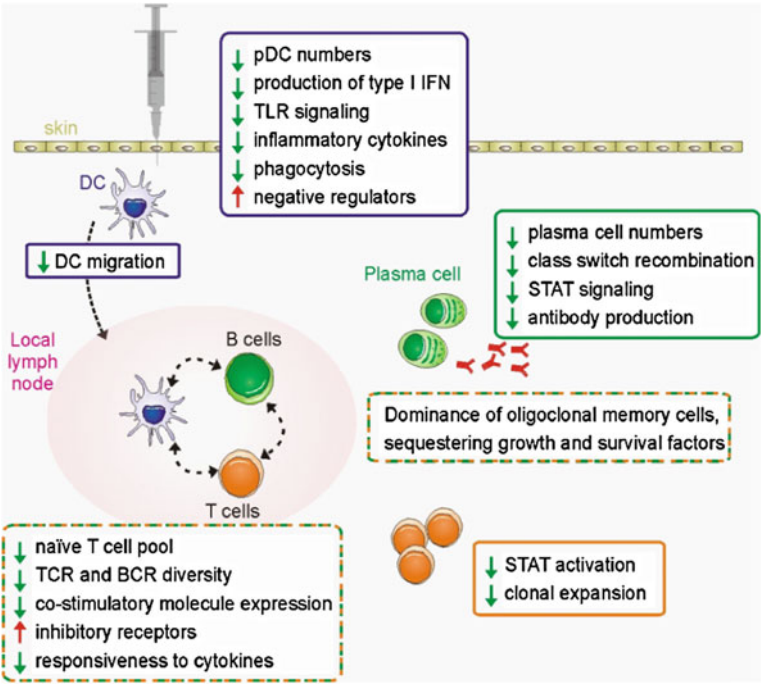


Fig. 1 Immunosenescence-associated defects in vaccine responses. Potential vulnerabilities in DCs, T cells, and B cells during immune responses to vaccination in the elderly. BCR, B cell receptor; DC, dendritic cell; IFN, interferon; TCR, T cell receptor; TLR, Toll-like receptor

2 Problems with Immune Responses in the Elderly

In the most simplified model, successful vaccine responses require activation of dendritic cells (DC), T cell activation and differentiation into effector cells and long-lived memory T cells. T cell help can also regulate B cell activation, differentiation, and antibody production. Obviously, the system has many potential vulnerabilities that could be the cause of defective immune responses in the elderly (Fig. 1).

2.1 Immune Cell Generation

The immune system is highly dynamic, with regenerative and homeostatic mechanisms subject to change with age. Assessment of whether cell subset numbers decline with age relies almost entirely on peripheral blood analysis in humans. With this reservation, numbers of myeloid DC (mDC) do not appear to decline (Jing et al. 2009), consistent with the observation that aging hematopoietic

stem cells (HSCs) are biased toward the production of myeloid cells (Pang et al. 2011; Wang et al. 2012). In contrast, plasmacytoid (pDC) numbers are reduced (Jing et al. 2009). T cells are most affected by aging, as thymic function dramatically decreases throughout life. Unlike in mice, thymic function in the healthy adult human is not a prerequisite for maintaining a naïve T cell compartment; homeostatic proliferation accounts for most T cell generation, likely even in the young adult (den Braber et al. 2012). Consequently, the total number of naïve $CD4^+$ T cells declines only moderately with age, and most individuals maintain substantial numbers into their 70s. $CD8^+$ naïve T cells are lost more rapidly than $CD4^+$ T cells (Czesnikiewicz-Guzik et al. 2008). This more rapid decline was initially thought to be related to the expansion of end-differentiated effector T cells specific for cytomegalovirus (CMV) (Sauce et al. 2009). However, recent data show that this is independent of CMV infection (Nikolich-Žugich et al. 2012). The difference between $CD4^+$ and $CD8^+$ T cells extends to the memory compartment, with numbers of central memory $CD4^+$ T cells being relatively robust, while $CD8^+$ T cells show a shift toward end-differentiated effector cells (Czesnikiewicz-Guzik et al. 2008).

Unlike thymic production of T cells, percentages and numbers of B cell precursors in the bone marrow remain relatively stable with age (McKenna et al. 2001; Rossi et al. 2003), but there is a significant decrease in mature B cells in the peripheral blood (Ademokun et al. 2010). As with T cells, there is a shift in the composition of the B cell pool—the percentage and number of switch memory B cells is reduced in elderly individuals. Whilst the percentages of naive and IgM memory B cells increase or remain constant, absolute numbers are decreased (Frasca et al. 2008).

In addition to total numbers, repertoire diversity is an important determinant of protective immunity. Based on clonal population sizes of human naïve T cells expressing identical TCRs, the development of holes in the repertoire, as observed in mice following infection (Yager et al. 2008), is unlikely to occur in humans. $CD4^+$ TCR diversity is maintained for many years without any contraction. However, at a later age, naïve $CD4^+$ T cell turnover increases, numbers decline, and the repertoire sharply contracts (Naylor et al. 2005). Peripheral selection during homeostatic proliferation may bias and contract the naïve repertoire, as has been shown for murine $CD8^+$ T cells, thereby influencing the quality of immune memory (Rudd et al. 2011). There is also a decrease in BCR diversity, which has been associated with poor health and frailty in old age (Gibson et al. 2009), and which may result in imperfect antigen fit and suboptimal responses to vaccination.

In summary, despite thymic involution, the numbers and diversity of $CD4^+$ T cells are relatively unchanged in early aging, whereas subset imbalances in the $CD8^+$ compartment are more obvious. In contrast, in older age (>75 years), T cell numbers and diversity appear to be severely limiting. Obviously, such a defect presents a difficult challenge to overcome by improving vaccine strategies.

2.2 Immune Cell Activation

The early innate stages of a vaccine response include the maturation of DCs, and the activation of T cells supported by co-stimulatory signals and cytokines. Toll-like receptor (TLR)-mediated activation of DCs is reduced in elderly individuals in terms of cytokine production and co-stimulatory molecule expression (Nyugen et al. 2010; Panda et al. 2010). It is unclear whether the defects extend to other DC functions, such as antigen presentation, cross-presentation, and DC migration, although in vitro chemokine-induced migration of DCs has been shown to be impaired (Agrawal et al. 2007). Notably, DCs in the elderly are already constitutively activated and secrete cytokines (Della Bella et al. 2007), suggesting that the host environment contains activating mediators that may induce negative feedback loops and desensitize DCs.

Depending on the tissue, recruitment of memory T cells to the site of antigen presentation may be impaired in the elderly. In the case of skin, which is an immunogenic microenvironment increasingly considered for vaccinations, the observed defect was not inherent to the homing ability of T cells. Rather, defective production of TNF-α by dermal macrophages led to a failure to activate dermal blood vessels (Agius et al. 2009).

In contrast to aged murine T cells which exhibit severe defects in calcium fluxes during activation (Miller et al. 1987), age-related differences in human T cell activation are subtle and are frequently explained by differences in T cell differentiation. Effector $CD8^+$ T cells lack co-stimulatory molecules and express co-inhibitory receptors (Ouyang et al. 2003; Tarazona et al. 2000; Xu et al. 2005). Unlike naïve cells, memory $CD4^+$ T cells generally disfavor the ZAP70-LAT-ERK pathway upon TCR stimulation, which has important implications for negative and positive feedback loops (Adachi and Davis 2011). There are, however, changes to T cell function with age that are independent of differentiation. With age, the ability of naïve $CD4^+$ T cells to generate an ERK response is reduced owing to the increased expression of the dual specific phosphatase DUSP6, resulting in reduced TCR sensitivity (Goronzy et al. 2012; Li et al. 2007). The defect is less pronounced for $CD4^+$ memory T cells, which already disfavor the ERK pathway. Reduced TCR sensitivity can be critical if antigen or co-stimulatory signals are limiting, or if the diversity of the TCR repertoire limits the availability of suitable TCRs.

2.3 Immune Cell Proliferation and Differentiation

Vaccine responses depend on the ability of T and B cells to expand and differentiate. Several mechanisms contribute to reduced proliferative capacity in age, including telomere attrition, reduced telomerase expression, and increased expression of co-inhibitory receptors (Cavanagh et al. 2011; Honda et al. 2001; Valenzuela and Effros 2002; Vaziri et al. 1993; Voehringer et al. 2002). Gene expression arrays have

identified metallothioneins as one protective mechanism that preserves proliferative capacity. Expression of metallothioneins is regulated by the zinc concentration-dependent transcription factor MTF-1, suggesting that zinc metabolism could be targeted to improve lymphocyte proliferation in the elderly (Lee et al. 2008).

Factors regulating T cell differentiation include initial TCR signal strength (Rogers and Croft 1999) and cytokine-mediated STAT signals (O'Shea and Plenge 2012). STAT signaling changes with age (Fulop et al. 2006; Longo et al. 2012), but the consequences for T cell differentiation are currently unclear. Recently, age-dependent expression of DUSP4 has been shown to affect T effector cell function, in particular to impair helper cell activity for B cell responses (Yu et al. 2012). The expression of this phosphatase is regulated by AMPK and therefore the metabolic state of the cell, which may be reduced in the elderly. Interestingly, recently identified signatures predicting vaccine responses included regulators of glycolysis and protein synthesis (Querec et al. 2008). Together with the recent recognition that T cell differentiation correlates with metabolic pathways (Finlay and Cantrell 2011; Pearce et al. 2009), these results from systems biology approaches identify cellular metabolism as a focus of interest.

Several studies have shown a decrease in antibody titers in elderly individuals (Sasaki et al. 2011; Stiasny et al. 2012). This suggests that the reduced antibody response observed in the elderly is primarily due to quantitative rather than qualitative antibody defects. However, many molecular studies have shown intrinsic B cell deficiencies that accumulate with age. These include a loss of the B lineage-specific effector molecule EBF, and decreased binding ability of B cell specific activator protein (BASP). These defects were reversed following transfection of precursor cells with active STAT5, again indicating that the STAT pathway may be affected in the elderly (Lescale et al. 2010). It is likely that quantitative and qualitative changes to B cell populations both contribute to suboptimal responses, depending on the vaccine and the recipient.

3 Systems Vaccinology

Systems biology approaches can be used with the aim of understanding the complex interactions between all components of a biological system, and using this information to generate rules that predict subsequent behavior of the system (Kitano 2002). With the advent of high-throughput technologies that allow us to assess perturbations in the transcriptome (sets of transcripts), proteome (proteins), and metabolome (metabolites) after vaccination, large amounts of data can be collected and integrated using computational methods, in order to understand the response of the system to vaccination as a whole, rather than as individual parts. The goal of systems vaccinology is to gain a more global representation of the immune response to vaccination, with the hopes of identifying mechanisms of action of current successful vaccines and to use this information for the rational design of novel vaccines (Nakaya et al. 2012; Pulendran et al. 2010). This is

particularly pertinent to elderly populations, where current vaccines are often sub-optimal in preventing disease. Given the multiple potential vulnerabilities of the aged immune system described above, understanding the specific points of weakness of different types of vaccines in the elderly may allow more specific tailoring of vaccines to improve efficacy in the elderly.

3.1 Proof of Principle

Yellow fever virus (YFV) vaccine is one of the most effective vaccines ever made. A single immunization is known to induce neutralizing antibody titers that last up to four decades, as well as cytotoxic T cells, and a balanced Th1 and Th2 cell cytokine profile. Its efficacy in protecting against infection with yellow fever is roughly 90%. Two pioneering studies helped to unravel the molecular mechanisms associated with YFV vaccination (Gaucher et al. 2008; Querec et al. 2008). Microarray analyses of peripheral blood mononuclear cells (PBMCs) isolated from the blood of healthy adults vaccinated with YFV revealed a molecular signature, induced 3–7 days after vaccination. This signature was composed of genes encoding proteins involved in the antiviral response, typified by activation of the type I interferon pathway, as well as complement and inflammasome-related genes (Gaucher et al. 2008; Querec et al. 2008). Moreover, an early innate signature that was able to predict the magnitude of the $CD8^+$ T cell response, in an independent study using a separate cohort of vaccinees who received YFV, was identified; this signature included EIF2AK4, which is involved in the cellular stress response (Querec et al. 2008). Subsequent mechanistic studies have revealed a critical role for EIF2AK4 in regulating $CD8^+$ T cell responses to YFV and some other viruses (Nair et al.—in preparation). We also identified signatures that were capable of predicting the neutralizing antibody response to YFV. For the antibody response predictive signature, TNFRSF17 a gene that encodes for BCMA, a protein known to regulate plasma cell differentiation, was identified (Querec et al. 2008). This study establishes the proof of concept of using systems biological approaches to identify signatures that predict the immunogenicity of a vaccine.

We also performed a subsequent study using the seasonal influenza vaccines, live attenuated virus vaccine (LAIV), or trivalent inactivated vaccine (TIV); as distinct from the YFV study, this represents responses that are likely to be recall responses. Similar to YFV, LAIV induced a type I interferon signature, which may be common to live attenuated viral vaccines, whereas TIV induced a signature comprising genes known to be induced during the plasma B cell response. This signature was able to predict the subsequent hemagglutinin antibody (HAI) titers (Nakaya et al. 2011). Notably, the expression of the gene encoding TNFRSF17, which was a component of the predictive signature to YFV, was also identified as a signature for TIV. This suggests that there may be common signatures that predict antibody responses to several vaccines. Furthermore, we identified that the expression of the gene encoding CAMKIV a few days after vaccination was

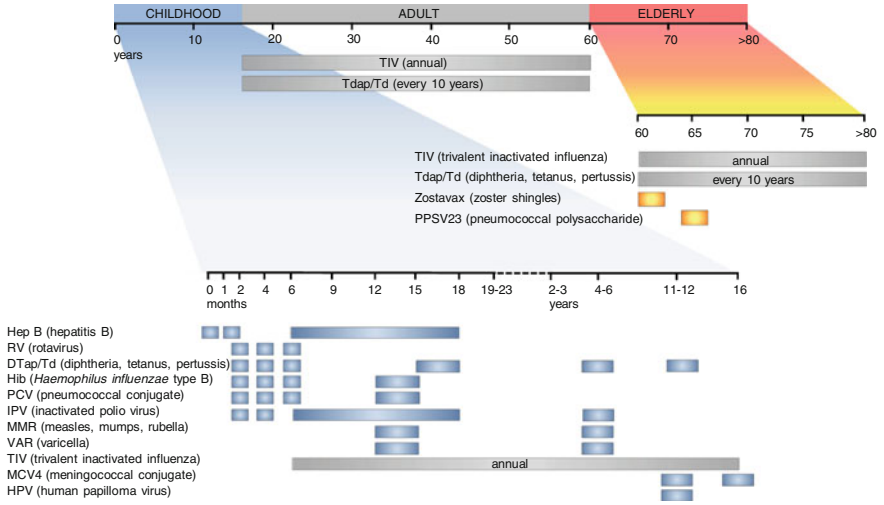


Fig. 2 Lifetime routine vaccination schedule. Adapted from the Centers for Disease Control and prevention (CDC)

inversely correlated with the magnitude of the later HAI titers (Nakaya et al. 2011). Subsequent mechanistic studies using mice deficient in CAMKIV revealed a key role for this molecule in regulating antibody responses to TIV.

These studies demonstrate the concept of how systems biology approaches can be used to predict the immunogenicity of vaccines, and generate ideas for novel hypotheses regarding the mechanisms of action of these vaccines.

4 Vaccines for the Elderly

The vaccination schedule recommended for US populations by the Centers for Disease Control (CDC) for all age groups is depicted in Fig. 2. The majority of routine vaccinations are given in childhood, with adults receiving an annual influenza vaccination and a tetanus/diphtheria booster every 10 years. The elderly additionally receive zoster and pneumococcal vaccines. Several vaccines, such as yellow fever, meningococcal, and hepatitis A vaccines, are also recommended for all individuals traveling to specific endemic areas. Since an estimated 15% of travelers are >65 years old (Hill 2000), travel vaccinations are also an important preventative health measure in some older individuals. Here we explore suboptimal responses in the elderly to different types of vaccines, including inactivated virus, polysaccharide, conjugate, live attenuated virus, and subunit vaccines.

4.1 Influenza Vaccination

Influenza results in 3,000–49,000 deaths annually in the US, with 90% of deaths occurring in those >65 years (CDC 2010a). Vaccination with trivalent inactivated influenza vaccine (TIV), containing split virus from the circulating strains of influenza A (H1N1 and H3N2) and B, is currently recommended for the elderly. A recent meta-analysis indicated that clinical efficacy of TIV in healthy adults is around 60% (Osterholm et al. 2012), whereas in elderly populations it is thought to be lower, with estimates ranging from 17 to 53% (Goodwin et al. 2006). TIV represents a model of an inactivated vaccine used in elderly individuals that have likely had prior exposure to antigenically similar virus strains, thus responses are a mixture of primary and recall responses. TIV mainly elicits serum antibodies against the HA protein (HAI titers) which correlates with protection against influenza (Hirota et al. 1997). Older vaccinees have been shown to have significantly lower serum HAI titers post-vaccination (Goodwin et al. 2006).

Evaluating immune responses to most vaccines in the elderly is confounded by the fact that most vaccinees have had prior exposure to pathogen or vaccine, which will almost certainly impact the magnitude of response to vaccination. Individuals that had received TIV one year prior to re-vaccination had higher baseline serum HAI responses, which negatively correlated with the post-vaccination number of antibody-secreting cells (ASCs) and HAI titer change (Sasaki et al. 2008). Vaccine responses are therefore a function of both immunosenescence and history of pathogen exposure.

Recent studies have suggested that a decrease in the number of ASCs post-vaccination is responsible for the reduced serum HAI titers observed in elderly individuals. No difference in antibody avidity or antibody secretion on a per cell basis was observed, suggesting that weaker humoral responses in the elderly may be due to a quantitative defect in ASC numbers, rather than qualitative differences in antibody functionality (Sasaki et al. 2011). However, other studies have found that elderly vaccinees receiving seasonal or pandemic influenza vaccinees, had lower proportions of switched memory B cells in the blood and lower B cell expression of activation-induced cytidine deaminase (AID) mRNA, which is involved in class switching and somatic hypermutation. Moreover, pre-vaccination AID expression, induced by CpG stimulation, correlated with the subsequent HAI response, suggesting that an intrinsic defect in B cell function in the elderly may contribute to poor humoral responses (Frasca et al. 2010, 2012).

Although HAI titers are useful as a correlate of protection at the population level, this may not be the best predictor for protection in the elderly. As such, HAI titers were unable to distinguish between elderly individuals that developed influenza and those that did not; some that developed influenza had ‘protective’ titers (McElhaney et al. 2006). TIV mostly elicits antibody responses; however, cytotoxic T cells have been implicated in controlling the severity of influenza infection (McMichael et al. 1983). Additionally, pre-existing $CD4^+$ T cells specific for influenza internal antigens were found to negatively correlate with influenza

illness severity and virus shedding. These individuals were all seronegative to the challenge strain prior to infection, suggesting that memory T cells may limit illness severity even in the absence of pre-existing antibodies (Wilkinson et al. 2012). Elderly individuals were also found to have fewer IFNγ-secreting influenza-specific $CD8^+$ T cells, and many of these cells had a senescent/late differentiation phenotype (Wagar et al. 2011). Whether the increased susceptibility to influenza disease in the elderly is also a consequence of diminished T cell responses is uncertain. Other immunological parameters identified by a more global systems biology approach may provide a more accurate predictor of protection, especially in elderly populations.

4.2 Pneumococcal Polysaccharide Vaccination

In the US in 2009 there were 5,000 *Streptococcus pneumoniae*-related deaths, with a disproportionate effect on the elderly; vaccination is recommended for >65 year olds. Until recently, only Pneumovax-23 (PPSV23) was licensed for use in the elderly. PPSV23 contains capsular polysaccharides from 23 bacterial serotypes that cause 80% of invasive disease (CDC 2010b). Efficacy against invasive disease, characterized by bacteremia and meningitis, in the general elderly population is 60–80%; however, PPSV23 does not appear to be effective in preventing pneumonia (Ortqvist et al. 1998). Efficacy is also dramatically lower in older (>85 years) or immunocompromised elderly individuals (Melegaro and Edmunds 2004; Shapiro et al. 1991).

Although the exact correlates of protection are unclear, serum IgG and opsonophagocytic antibody titers (OPA titers) are used as surrogate markers (Jodar et al. 2003). The OPA assay measures the ability of serum anti-capsular antibodies to opsonize pneumococci in order to be killed by phagocytes. Opsonizing antibodies have been implicated in protection from pneumonia in human patients (Musher et al. 2000) and animal models (Guckian et al. 1980). Passive transfer of opsonizing human antibodies to neonatal mice correlated with protection from bacteremia (Johnson et al. 1999). Capsular polysaccharides, which are T-independent antigens, cross-link B cell receptors to stimulate clonal expansion and antibody production, but lack the ability to induce memory (Mazmanian and Kasper 2006). Several studies have shown that PPSV23 vaccination induces similar levels of serum IgG in young and elderly vaccinees. However, aging seems to affect the functional quality of antibodies, as demonstrated by a decrease in OPA titers (Romero-Steiner et al. 1999; Rubins et al. 1998; Schenkein et al. 2008). Although serum IgM levels induced after vaccination are low, a significantly lower level of IgM has been described in elderly vaccinees (Park and Nahm 2011; Shi et al. 2005). Interestingly, depletion of IgM from young sera eliminated OPA titer differences between the young and elderly (Park and Nahm 2011), suggesting that decreased IgM levels in the elderly may account for the reduced opsonizing ability of their serum. The PPS-specific IgG variable heavy chain (V_H) gene repertoire was also found to differ between the young and elderly after vaccination (Kolibab

et al. 2005). The functional consequences of this are unclear, but may be a sign of differences in both intrinsic gene rearrangement mechanisms, and lifetime exposure to multiple serotypes, which would shape the memory B cell antibody repertoire through clonal expansion, and therefore affect post-vaccination antibody diversity.

Immunity induced by PPSV23 appears to decline 3–5 years after vaccination (Shapiro et al. 1991), which is unsurprising given the T-independent nature of the antigen. Re-vaccination of elderly individuals with PPSV23 resulted in an increase in antibody titers that was less than that observed after the first dose, suggesting hyporesponsiveness rather than boosting (Torling et al. 2003). Furthermore, PPSV23 vaccination resulted in a decrease in blood memory B cell frequency. One possible explanation of this may be that repeated administration of a T-independent antigen drives memory B cells to terminal differentiation without replenishing the memory B cell pool, leading to hyporesponsiveness (Clutterbuck et al. 2012).

4.3 Pneumococcal Polysaccharide-Protein Conjugate Vaccination

The 7- and 13-valent polysaccharide-protein conjugate vaccines, Prevnar (PCV7 and PCV13), have been previously licensed for use in children; recently PCV13 has also been licensed for use in the elderly. PCV13 contains capsular polysaccharides from 13 serotypes conjugated to diphtheria-CRM_{197} protein to form a T-dependent antigen. The carbohydrate component can stimulate B cell receptors, and antigen presentation of the protein component by the same B cell can activate T cell help for B cell class-switching, affinity maturation and B cell memory (Mazmanian and Kasper 2006). Recently, it has been shown that TCRs can directly recognize carbohydrate fragments when they are anchored to MHCII by a conjugate-peptide, offering a new explanation of how conjugate vaccines may recruit antigen-specific T cell help (Avci et al. 2011). Vaccination of elderly individuals with PCV7 led to higher OPA titers compared to PPSV23 (de Roux et al. 2008). Re-vaccination of PPSV23-primed elderly individuals with PCV7 induced greater OPA titers compared to PPSV23 (Jackson et al. 2007), suggesting that T-dependent antigens may elicit qualitatively superior antibodies and may be used as a booster without inducing hyporesponsiveness.

Despite these advances, questions still remain unresolved as to precisely how these vaccines work. Systems biology approaches may be used to compare signatures induced by polysaccharide versus conjugate vaccines, thereby offering mechanistic insights into how T-independent and T-dependent antigens elicit immunity. Additionally, evaluation of these vaccines in the elderly may help to explain the molecular basis for the decline in antibody functionality with age, which may be used to improve vaccine efficacy in the most vulnerable elderly populations.

4.4 Varicella Zoster Vaccination

Primary infection with varicella zoster virus (VZV) causes chickenpox—a disease which affected 4 million people annually in the USA prior to the introduction of childhood vaccination (Jumaan et al. 2005). On clearance of disease symptoms, the virus establishes life-long latency within dorsal root or trigeminal ganglia which typically remains subclinical. A decline in cell-mediated immunity (CMI), either following immunosuppression or as a result of aging, can result in a loss of viral control, reactivation of the virus, and herpes zoster (shingles) (Arvin 2005). Lasting post-herpetic neuralgia (PHN) is the most burdensome aspect of shingles and occurs in 20% of cases. As the lifetime risk of zoster is between 22 and 32% (Chapman et al. 2003), PHN is a significant cause of morbidity.

Unlike immunity to influenza infection, which requires high levels of neutralizing antibodies, protection against reactivation of latent VZV and the development of shingles is thought to be independent of antibody and instead requires robust CMI. A vaccine targeting shingles is approved for use in adults >60 years. Zostavax utilizes the same strain of live attenuated virus as in the chickenpox vaccine, but at approximately 14 times the dose. Results from clinical trials suggest that the vaccine is effective, reducing incidence of shingles by up to 51.3% and incidence of PHN by up to 66.5% (Oxman et al. 2005). The efficacy of this vaccine in a population, which typically exhibits low immune responses to vaccination, suggests that it will be useful as a model to monitor vaccine responses in the elderly and identify correlates and predictors of efficacy. An important caveat is that this vaccine relies on the reactivation of already extant immune memory rather than the initiation of a primary response to previously unseen pathogenic antigens.

The mechanisms of action of the vaccine and unequivocal correlates of protection have not been established to date. The vaccine does induce the production of VZV-specific antibodies and the expansion of VZV-specific T cells (Weinberg et al. 2009). The kinetics of CMI after vaccination have not been fully studied despite the importance of CMI in maintaining viral latency. Preliminary data suggest that the age of the vaccinee may strongly influence duration of protection. Results from our lab suggest that there is little, if any, viremia and low innate immune activation following VZV vaccination (Goronzy et al. unpublished observations).

4.5 Tetanus Toxoid, Diphtheria Toxoid, and Acellular Pertussis Vaccination

Tdap/Td vaccines consist of purified tetanus and diphtheria toxoids and pertussis antigens, adjuvanted with aluminum hydroxide, combined in a single vaccine. Immunity to tetanus, diphtheria, and pertussis antigens wanes over time, with those >40 years of age having lower serum antibody titers (Theeten et al. 2007); as such, a booster is recommended for adults every 10 years. This vaccine elicits

boosting of a pre-existing memory response established by previous vaccinations. Protection against tetanus and diphtheria is considered to be mediated by toxin-neutralizing antibodies, whereas the mechanisms of protection against pertussis are not clear. One booster dose is sufficient to result in 95–100% seroprotection rates in adults >40 years (Van Damme and Burgess 2004); however, the magnitude of the response tends to decrease with age. A recent study showed that a proportion of Tdap recipients >65 years did not achieve a response to two pertussis antigens and that tetanus and diphtheria antibody responses were lower in the >75-year-old subset (Weston et al. 2012). Thus, even booster responses to adjuvanted subunit antigens are diminished in the elderly, which may suggest a fundamental change in the way antigen is recognized by the innate immune system, or changes in maintenance of the memory compartment.

4.6 Other Vaccinations Given to the Elderly

4.6.1 Yellow Fever Vaccination

The YFV vaccine containing the live attenuated YF-17D strain is recommended for individuals traveling to endemic areas (Staples et al. 2010). In contrast to TIV, pneumococcal, and Tdap vaccines, YFV represents a model of primary immune responses, since most people in nonendemic areas will be immunologically naive to YFV. Since YF-17D is a live virus, it likely locally replicates and induces a strong innate immune response characterized by triggering several TLRs, as well as RIG-I and MDA-5, on multiple DC subsets and a type I IFN response (Querec et al. 2008, 2006). Neutralizing antibodies are induced and seroconversion is achieved in 97–100% of vaccinees 30 days after vaccination, with no significant difference between the young and elderly (Monath et al. 2005). However, closer examination reveals that antibody titers are slower to develop in the elderly, such that 10 days after vaccination there are significantly lower antibody titers in the elderly (Roukens et al. 2011).

Although elderly vaccinees eventually seroconvert, individuals >60 years carry an increased risk of systemic adverse effects (sysAE), which in more serious cases is characterized by viral dissemination to vital organs (Khromava et al. 2005; Lindsey et al. 2008; Martin et al. 2001). SysAE were found to be associated with higher viremia in the elderly (Roukens et al. 2011). Several cases of sysAE have also been reported in individuals with thymic disease (Barwick 2004), suggesting that T cells may play a role in preventing sysAE. A potential role for innate responses in sysAE was also demonstrated in a case report where the individual had elevated numbers of blood inflammatory $CD14^{+}CD16^{++}$ monocytes, increased plasma cytokine/chemokine levels (IL-1α, IL-6, CXCL10, CCL2, CCL5), and mutations in the CCR5 gene (expressed on monocytes) and its ligand CCL5. The patient also had increased viremia that persisted for 34 days after vaccination, whereas the virus is usually cleared by days 7–11. Increased antibody and antigen-specific $CD8^{+}$ T cell responses

were also observed, and persisted longer than in healthy vaccinees. Moreover, the numbers of inflammatory monocytes remained ten-fold higher compared to healthy vaccinees well after viral clearance (Pulendran et al. 2008). Given that the frequency of na $CD8^+$ T cells substantially decreases during aging, a dysfunctional innate response, a suboptimal primary $CD8^+$ T cell response and a delayed humoral response may prevent efficient control of viral replication, resulting in a greater risk of high viremia and disseminated sysAE. This suggests that immunosenescence may affect not only vaccine immunogenicity, but also vaccine safety.

4.6.2 Meningococcal Vaccination

A quadrivalent polysaccharide vaccine (Menomune) is available for use against the encapsulated bacteria *Neisseria meningitides.* Conjugate vaccines (Menactra and Menveo) are available but are not yet licensed for use in those >55 years old. The principle of how these vaccines work in terms of eliciting a T-independent or a T-dependent response are thought to be similar to the pneumococcal vaccines described previously, with a notable exception. In contrast to PPSV23 where the antibody response is qualitatively (opsonophagocytic titers) but not quantitatively (binding antibody titers by ELISA) affected, a MenC polysaccharide vaccine was shown to elicit lower levels of antibody measured by ELISA as well as lower functional serum bactericidal activity (SBA) titers in older vaccinees (Hutchins et al. 1999). Therefore, some caution must be used when grouping vaccines into broad 'types' (e.g. polysaccharide), as clearly immunosenescent responses differ for each pathogen.

4.6.3 Hepatitis A and B Vaccination

Hepatitis A vaccine, comprised of inactivated hepatitis A virus (HAV) adsorbed to aluminum hydroxide, induces seroconversion rates of nearly 100%. Typically, in older adults one HAV dose leads to lower antibody responses and requires boosting (Reuman et al. 1997). Although seroconversion is achieved after 2–3 doses, mean antibody titers are still lower in vaccinees >40 years old (McMahon et al. 1995).

Subunit hepatitis B vaccines containing the viral surface antigen (HBsAg) adjuvanted with aluminum hydroxide are given in a series of three doses. The proportion of individuals achieving seroprotective levels (>10 mIU/ml) and the titers of anti-HBsAg antibodies are considerably lower in older individuals (Tohme et al. 2011; Wolters et al. 2003). Consequently, inactivated virus and subunit vaccine responses also appear to be impaired in the elderly. However, whether there is any clinical significance of the seroprotective, but lower levels of anti-HAV and anti-HBsAg antibody titers in terms of duration of immunity is unknown. It is thought that even when circulating anti-HBsAg antibodies drop below seroprotective levels, some protection may be afforded by the presence of

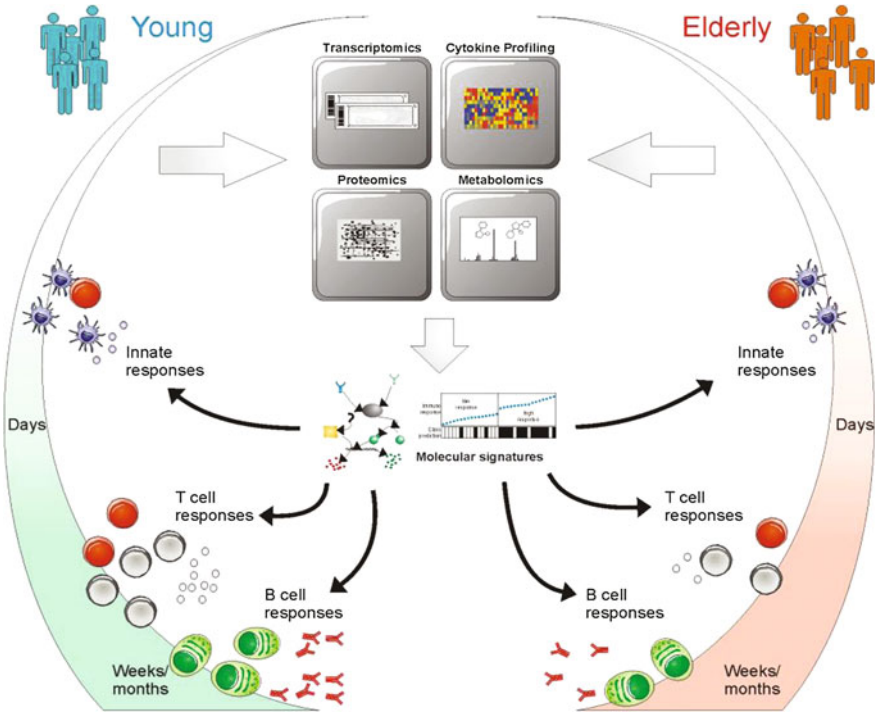


Fig. 3 A potential approach to systems biology studies in human vaccinees. Transcriptomic, proteomic, metabolomic, and cytokine-profiling information collected from young and elderly vaccinees can be used to identify early molecular signatures that may predict the subsequent T and B cell responses

memory B cells; assessment of memory B cell induction in the elderly may, therefore, also be useful (West and Calandra 1996).

5 Systems Biology Approaches to Identifying Signatures of Immunogenicity in the Elderly

Future studies aim to integrate information from transcriptomic, metabolomic, and proteomic approaches as well as immunological assays such as multi-parameter flow cytometry, ELISpots, and multiplex-cytokine profiling collected from young and elderly human vaccinees (Fig. 3). Utilizing this information, there are several key questions that can be addressed using systems vaccinology.

There are profound changes in many aspects of the immune system with age, including changes in cell numbers, receptor repertoire, activation, differentiation, and function (See Sect. 2). What are the fundamental differences in the immune systems of younger and older individuals at baseline? And can these differences be

combined into a signature that can predict the vaccine response? The baseline immune status of an older individual will be the result of both intrinsic immunosenescent defects as well as the history of exposure to pathogens or previous vaccines. As with influenza vaccination, where the baseline HAI titer negatively correlates with the subsequent HAI response (Sasaki et al. 2008), a more sophisticated algorithm that incorporates indicators of prior exposure (e.g. HAI titer) as well as transcriptomic information on intrinsic cellular defects may be used to more accurately predict the vaccine response.

In addition to baseline differences between young and elderly individuals, what are the factors that come into play during an ongoing immune response that contribute to poorer vaccine efficacy in the elderly? The immune response to vaccination in the elderly can be thought of in terms of the following equation:

Immune responses in the elderly = function [Innate sensing and response] [Precursor frequency of antigen-specific T and B cells] [Proliferative capacity of antigen-specific T and B cells] [Functional capacity of antigen-specific T and B cells] [Host microenvironment].

DCs from elderly individuals may differ in their ability to sense pathogens and in their functional response. A detailed understanding of the pattern recognition receptor (PRR) signaling networks in response to vaccines in young versus aged DCs may pinpoint specific molecular defects that can be targeted to restore or improve DC function and vaccine efficacy in the elderly. The antigen-specific T and B cell precursor frequency, which is affected by the na cell output, previous antigen experience and homeostatic proliferation and survival, would differ in the elderly. Sequencing of Ig heavy chain genes from antigen-specific B cells isolated from young and elderly vaccinees, could be used to explore changes in antibody repertoire diversity that may affect antibody function such as binding specificity and affinity (Wiley et al. 2011). The proliferative ability and functional capacity of T and B cells, such as the type and quantity of cytokines secreted, cytotoxic ability, and the quantity and quality of antibodies secreted may also contribute to decreased vaccine efficacy. These factors will also depend on changes at the population and single-cell level. For example, how much of the inferior response seen in the elderly is due to defects in proliferative and functional ability on an individual cell basis, and how much is due to a decrease in the number of otherwise functionally competent cells? Antigen-specific tetramer^{+} T cells could be isolated from young and elderly vaccinees and deep sequencing carried out to understand the molecular basis for any intrinsic functional defects. Multiple-parameter single-cell mass cytometry could also be used to compare the quantitative differences in cell subsets and differences in phenotype and signaling of T and B cells from young and elderly vaccinees (Bendall et al. 2011). The host microenvironment in terms of cytokine/survival factor secretion by stromal cells, and the architecture of lymphoid and peripheral tissues may change with aging, which could affect cellular migration and in situ activation of immune cells. Systems vaccinology approaches often rely on analysis of peripheral blood, which makes the effect of tissue factors difficult to assess directly. However, insights

from peripheral blood (e.g. changes in homing receptor expression) may help to generate hypotheses that could be tested further in animal models.

One application of systems vaccinology would be to identify the early robust signatures that could predict suboptimal responses in the elderly. For example, how do the molecular signatures identified with TIV vaccination in healthy young adults (Nakaya et al. 2011) compare to those in the elderly, and can they still predict HAI responses? Furthermore, identification of genes and pathways elicited by vaccination that are unique to the young or the elderly, or common to both groups may offer insights into immunosenescence. Such signature differences could highlight the specific vulnerabilities in the aged immune system that could be the basis for generating novel hypotheses for future mechanistic studies to explore. For example, the YF-17D vaccine is known to trigger the cellular stress response, and genes of this pathway form part of the signature that predicted the $CD8^+$ T cell response (Querec et al. 2008). Given that elderly individuals are known to have dysfunctional naive $CD8^+$ T cell responses, which may contribute to the increased risk of systemic adverse effects seen after YFV vaccination, it would be interesting to ascertain whether the cellular stress response is also impaired in the elderly. Ultimately, the aim is to identify specific defects in the aged immune system that could be targeted with novel vaccination strategies (see Sect. 6) that improve efficacy in the elderly; such approaches may need to be customized for individual vaccines or certain 'types' of vaccines (e.g. live attenuated viral), depending on the mechanism of action of each vaccine.

Correlates of protection are often defined by a certain quantity of neutralizing or opsonizing antibody (e.g. influenza and pneumococcal); however, for some vaccines the correlates are less well defined and most likely include T cell responses too. In many cases, it is likely that a combination of antibody, $CD8^+$ and diverse $CD4^+$ responses are involved; systems biology could be used to define more sophisticated correlates of protection, based on the combination of multiple parameters. This is particularly important in elderly populations where traditional correlates may not be as reliable, such as HAI titers (McElhaney et al. 2006), and may help to improve our understanding of the link between immunogenicity and efficacy of each vaccine. Additionally, the identification of early innate signatures that predict immunogenicity, and ultimately efficacy in the elderly, could be incorporated into a 'vaccine chip' that would predict nonresponders within a few days of vaccination (Pulendran et al. 2010). This could be a crucial public health tool to identify the most vulnerable elderly individuals who could be targeted for re-vaccination or closer follow-up care.

Another advantage of collecting '-omics', data from large numbers of elderly vaccine trial participants, is that this information may eventually also be used to answer more fundamental questions about immunosenescence, as well as human aging in general. Identification of intrinsic defects in cellular function that accumulate with age could be understood at a more molecular level. However, a caveat of such vaccine studies is that the chronological age of an individual may not be equivalent to their biological age. Study participants are more likely to be healthy elderly individuals and as such, immune responses may differ significantly to

responses in the frail elderly; caution must be used when extending study findings to the frail elderly, who are the population most in need of more successful vaccination strategies.

6 Strategies to Overcome Defective Vaccine Responses in the Elderly

The rational design of novel vaccination strategies will depend on identifying critical defects, which will differ depending on the age of the recipient and the type of vaccine. Most obviously, aged individuals with a severely diminished and diversity-contracted repertoire of T cells will require a different approach than individuals who have reduced innate immunity. The type of vaccine response, whether neutralizing antibodies or CMI, is equally important. Systems biological analysis of vaccine responses is an important tool to customize these approaches and improve and widen the strategies that are currently envisioned.

6.1 Innate Immune Activation by Adjuvants

Most current preclinical and clinical studies aim to identify adjuvants that can activate innate immune cells, including DCs. Innate immune activation shapes the quality and magnitude of the ensuing adaptive immune response. The attenuated YF-17D strain, one of the most powerful available vaccines, is an extremely potent activator of innate immunity (Querec et al. 2006). mDC responses to TLRs are diminished with age, type I interferon-producing pDCs may be reduced (Jing et al. 2009), and vascular activation in the skin is impaired (Agius et al. 2009). Numerous targets also exist among pattern recognition receptors and other danger sensing receptors. As a proof-of-concept, the CpG-adjuvanted hepatitis B vaccine stimulating TLR9, boosted vaccine efficacy compared to the conventional vaccine in those age >40 years (Sablan et al. 2012). However, given that age-related defects in TLR signaling may involve the triggering of negative feedback loops by constitutive stimulation, it is currently unclear whether adjuvants will be able to overcome this defect.

6.2 Improving Vaccine Delivery

Strategies to increase vaccine doses or to promote sustained antigen delivery have the potential to overcome defects such as delayed recruitment of immune cells or reduced sensitivity of TCRs and BCRs. Aluminum-based adjuvants are

widely used to prolong release of antigen. The squalene MF59 adjuvant has been shown to be superior to unadjuvanted influenza vaccine in the elderly (Faenzi et al. 2012; Ikematsu et al. 2012). Viral and nonviral self-amplifying vaccines are under development and may eventually find application in the elderly, despite their compromised immunity. Increased vaccine doses have proven beneficial for the prevention of herpes zoster. However, a second dose 6 weeks following primary vaccination has no further effect (Vermeulen et al. 2012). Higher influenza vaccine doses have been explored as a means of increasing responsiveness in the elderly—high-dose TIV increases HAI titers in adults >65 years (Chen et al. 2011; Couch et al. 2007), but these responses are still lower than young adult responses to standard-dose TIV, and increased titers may only be observed for a limited set of antigens and not confer improved protection (Palache et al. 1993).

6.3 *Improving T and B Cell Activation, Expansion, Differentiation, and Survival*

Increased vaccine doses may be in part necessary to bypass neutralization by pre-existing antibodies, which are frequent in the elderly in particular for influenza vaccination (Feng et al. 2009), but they will also have a direct effect on T cells by increasing signal strength and overcoming decreased TCR sensitivity or imperfect fit owing to repertoire restriction. Similarly, improved co-stimulatory signals following DC activation will have positive effects on T cell activation and differentiation. Direct targeting of T or B cell defects may be possible, as already exemplified by the success of direct CD28 stimulation to activate melanoma-specific T cells, albeit with currently unacceptable autoimmune side effects (Hodi et al. 2010). More specific targeting of T cells, through inducible rather than constitutive co-stimulatory molecules, may overcome some of the risks of autoimmunity. Recent success has been documented with the BCG vaccine, the efficacy of which can be increased with the co-administration of OX40L fusion protein (Snelgrove et al. 2012). Obviously, a balance between the risk of autoreactivity and vaccine efficacy has to be found, which may be different for the elderly than for young adults. Signatures identified by systems biology approaches hold the promise to define defects and pathways that can be directly targeted. Possible examples already identified are the inhibition of DUSP6 to improve T cell activation and T helper cell function for B cell activation, and induction of metallothioneins by zinc supplementation to improve clonal expansion. Intriguing is the observation that metabolic pathways influence the outcome of a vaccine response (Pearce et al. 2009; Querec et al. 2008; Yu et al. 2012), which could be a promising target for intervention.

7 Conclusions

Aging of the immune system is accompanied by changes in the frequency, repertoire diversity, activation, and differentiation of cell subsets including DCs, T cells and B cells. As such, many vaccines that are effective in younger adults are suboptimal in the elderly, which is particularly important for diseases that are of major public health concern, such as influenza and pneumonia. Responses to other types of vaccines (live attenuated, polysaccharide, conjugate or subunit) are also reduced in the elderly. Systems biology approaches, which have been shown to be useful in predicting immunogenicity in the young, could be a useful tool when applied to the elderly. Future systems biology studies in the elderly may offer insights into mechanisms of immunosenescent responses and help to identify better correlates of protection. This could also be used in the rational design of novel vaccines that target specific defects in the elderly, thus improving vaccine efficacy.

Acknowledgments The work in the laboratory of B.P. was supported by grants U19AI090023, HHSN266200700006C, U54AI057157, R37AI48638, R37DK057665, U19AI057266, and NO1 AI50025 from the National Institutes of Health and a grant from the Bill & Melinda Gates Foundation.

References

Adachi K, Davis M (2011) T-cell receptor ligation induces distinct signaling pathways in naive vs. antigen-experienced T cells. Procl Natl Acad Sci USA 108:1549–1603

Ademokun A, Wu Y-C, Dunn-Walters D (2010) The ageing B cell population: composition and function. Biogerontology 11:125–162

Agius E, Lacy K, Vukmanovic-Stejic M, Jagger A, Papageorgiou A-P, Hall S, Reed J, Curnow S, Fuentes-Duculan J, Buckley C, Salmon M, Taams L, Krueger J, Greenwood J, Klein N, Rustin M, Akbar A (2009) Decreased TNF-alpha synthesis by macrophages restricts cutaneous immunosurveillance by memory CD4 $^{+}$ T cells during aging. J Exp Med 206:1929–1969

Agrawal A, Agrawal S, Cao J-N, Su H, Osann K, Gupta S (2007) Altered innate immune functioning of dendritic cells in elderly humans: a role of phosphoinositide 3-kinase-signaling pathway. J Immunol 178:6912–6934

Arvin A (2005) Aging, immunity, and the varicella-zoster virus. N Engl J Med 352:2266–2273

Avci FY, Li X, Tsuji M, Kasper DL (2011) A mechanism for glycoconjugate vaccine activation of the adaptive immune system and its implications for vaccine design. Nat Med 17:1602–1609

Barwick R (2004) History of thymoma and yellow fever vaccination. Lancet 364:936

Bendall SC, Simonds EF, Qiu P, Amir el AD, Krutzik PO, Finck R, Bruggner RV, Melamed R, Trejo A, Ornatsky OI, Balderas RS, Plevritis SK, Sachs K, Pe'er D, Tanner SD, Nolan GP (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 332:687–696

Cavanagh MM, Qi Q, Weyand CM, Goronzy JJ (2011) Finding balance: T cell regulatory receptor expression during aging. Aging Dis 2:398–413

CDC (2010a) Estimates of deaths associated with seasonal influenza—United States, 1976–2007. MMWR Morb Mortal Wkly Rep 59:1057–1062

CDC (2010b) Updated recommendations for prevention of invasive pneumococcal disease among adults using the 23-valent pneumococcal polysaccharide vaccine (PPSV23). MMWR Morb Mortal Wkly Rep 59:1102–1106

Chapman R, Cross K, Fleming D (2003) The incidence of shingles and its implications for vaccination policy. Vaccine 21:2541–2548

Chen W, Cross A, Edelman R, Sztein M, Blackwelder W, Pasetti M (2011) Antibody and Th1-type cell-mediated immune responses in elderly and young adults immunized with the standard or a high dose influenza vaccine. Vaccine 29:2865–2938

Clutterbuck EA, Lazarus R, Yu LM, Bowman J, Bateman EA, Diggle L, Angus B, Peto TE, Beverley PC, Mant D, Pollard AJ (2012) Pneumococcal conjugate and plain polysaccharide vaccines have divergent effects on antigen-specific B cells. J Infect Dis 205:1408–1416

Couch RB, Winokur P, Brady R, Belshe R, Chen WH, Cate TR, Sigurdardottir B, Hoeper A, Graham IL, Edelman R, He F, Nino D, Capellan J, Ruben FL (2007) Safety and immunogenicity of a high dosage trivalent influenza vaccine among elderly subjects. Vaccine 25:7656–7663

Czesnikiewicz-Guzik M, Lee W-W, Cui D, Hiruma Y, Lamar D, Yang Z-Z, Ouslander J, Weyand C, Goronzy J (2008) T cell subset-specific susceptibility to aging. Clin Immunol 127:107–125

de Roux A, Schmole-Thoma B, Siber GR, Hackell JG, Kuhnke A, Ahlers N, Baker SA, Razmpour A, Emini EA, Fernsten PD, Gruber WC, Lockhart S, Burkhardt O, Welte T, Lode HM (2008) Comparison of pneumococcal conjugate polysaccharide and free polysaccharide vaccines in elderly adults: conjugate vaccine elicits improved antibacterial immune responses and immunological memory. Clin Infect Dis 46:1015–1023

Della Bella S, Bierti L, Presicce P, Arienti R, Valenti M, Saresella M, Vergani C, Villa M (2007) Peripheral blood dendritic cells and monocytes are differently regulated in the elderly. Clin Immunol 122:220–228

den Braber I, Mugwagwa T, Vrisekoop N, Westera L, Mögling R, de Boer A, Willems N, Schrijver E, Spierenburg G, Gaiser K, Mul E, Otto S, Ruiter A, Ackermans M, Miedema F, Borghans J, de Boer R, Tesselaar K (2012) Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. Immunity 36:288–385

Faenzi E, Zedda L, Bardelli M, Spensieri F, Borgogni E, Volpini G, Buricchi F, Pasini F, Capecchi P, Montanaro F, Belli R, Lattanzi M, Piccirella S, Montomoli E, Ahmed S, Rappuoli R, Del Giudice G, Finco O, Castellino F, Galli G (2012) One dose of an MF59-adjuvanted pandemic A/H1N1 vaccine recruits pre-existing immune memory and induces the rapid rise of neutralizing antibodies. Vaccine 30:4086–4094

Feng J, Gulati U, Zhang X, Keitel W, Thompson D, James J, Thompson L, Air G (2009) Antibody quantity versus quality after influenza vaccination. Vaccine 27:6358–6420

Finlay D, Cantrell D (2011) Metabolism, migration and memory in cytotoxic T cells. Nat Rev Immunol 11:109–126

Frasca D, Diaz A, Romero M, Landin AM, Phillips M, Lechner SC, Ryan JG, Blomberg BB (2010) Intrinsic defects in B cell response to seasonal influenza vaccination in elderly humans. Vaccine 28:8077–8084

Frasca D, Diaz A, Romero M, Phillips M, Mendez NV, Landin AM, Blomberg BB (2012) Unique biomarkers for B-cell function predict the serum response to pandemic H1N1 influenza vaccine. Int Immunol 24:175–182

Frasca D, Landin A, Lechner S, Ryan J, Schwartz R, Riley R, Blomberg B (2008) Aging down-regulates the transcription factor E2A, activation-induced cytidine deaminase, and Ig class switch in human B cells. J Immunol 180:5283–5373

Fulop T, Larbi A, Douziech N, Levesque I, Varin A, Herbein G (2006) Cytokine receptor signalling and aging. Mech Ageing Dev 127:526–563

Gaucher D, Therrien R, Kettaf N, Angermann BR, Boucher G, Filali-Mouhim A, Moser JM, Mehta RS, Drake DR, 3rd, Castro E, Akondy R, Rinfret A, Yassine-Diab B, Said EA, Chouikh Y, Cameron MJ, Clum R, Kelvin D, Somogyi R, Greller LD, Balderas RS, Wilkinson P, Pantaleo G, Tartaglia J, Haddad EK, Sekaly RP (2008) Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. J Exp Med 205:3119–3131

Gibson K, Wu Y-C, Barnett Y, Duggan O, Vaughan R, Kondeatis E, Nilsson B-O, Wikby A, Kipling D, Dunn-Walters D (2009) B-cell diversity decreases in old age and is correlated with poor health status. Aging cell 8:18–43

Goodwin K, Viboud C, Simonsen L (2006) Antibody response to influenza vaccination in the elderly: a quantitative review. Vaccine 24:1159–1169

Goronzy JJ, Li G, Yu M, Weyand CM (2012) Signaling pathways in aged T cells—a reflection of T cell differentiation, cell senescence and host environment. Semin Immunol (in press)

Guckian JC, Christensen GD, Fine DP (1980) The role of opsonins in recovery from experimental pneumococcal pneumonia. J Infect Dis 142:175–190

Hill DR (2000) Health problems in a large cohort of Americans traveling to developing countries. J Travel Med 7:259–266

Hirota Y, Kaji M, Ide S, Kajiwara J, Kataoka K, Goto S, Oka T (1997) Antibody efficacy as a keen index to evaluate influenza vaccine effectiveness. Vaccine 15:962–967

Hodi F, O'Day S, McDermott D, Weber R, Sosman J, Haanen J, Gonzalez R, Robert C, Schadendorf D, Hassel J, Akerley W, van den Eertwegh A, Lutzky J, Lorigan P, Vaubel J, Linette G, Hogg D, Ottensmeier C, Lebbé C, Peschel C, Quirt I, Clark J, Wolchok J, Weber J, Tian J, Yellin M, Nichol G, Hoos A, Urba W (2010) Improved survival with ipilimumab in patients with metastatic melanoma. N Engl J Med 363:711–734

Honda M, Mengesha E, Albano S, Nichols W, Wallace D, Metzger A, Klinenberg J, Linker-Israeli M (2001) Telomere shortening and decreased replicative potential, contrasted by continued proliferation of telomerase-positive CD8 + CD28(lo) T cells in patients with systemic lupus erythematosus. Clin Immunol 99:211–432

Hutchins WA, Carlone GM, Westerink MA (1999) Elderly immune response to a TI-2 antigen: heavy and light chain use and bactericidal activity to Neisseria meningitidis serogroup C polysaccharide. J Infect Dis 179:1433–1440

Ikematsu H, Nagai H, Kawashima M, Kawakami Y, Tenjinbaru K, Li P, Walravens K, Gillard P, Roman F (2012) Characterization and long-term persistence of immune response following two doses of an AS03A-adjuvanted H1N1 influenza vaccine in healthy Japanese adults. Human Vacc Immunother 8:260–266

Jackson LA, Neuzil KM, Nahm MH, Whitney CG, Yu O, Nelson JC, Starkovich PT, Dunstan M, Carste B, Shay DK, Baggs J, Carlone GM (2007) Immunogenicity of varying dosages of 7-valent pneumococcal polysaccharide-protein conjugate vaccine in seniors previously vaccinated with 23-valent pneumococcal polysaccharide vaccine. Vaccine 25:4029–4037

Jing Y, Shaheen E, Drake R, Chen N, Gravenstein S, Deng Y (2009) Aging is associated with a numerical and functional decline in plasmacytoid dendritic cells, whereas myeloid dendritic cells are relatively unaltered in human peripheral blood. Human Immunol 70:777–861

Jodar L, Butler J, Carlone G, Dagan R, Goldblatt D, Kayhty H, Klugman K, Plikaytis B, Siber G, Kohberger R, Chang I, Cherian T (2003) Serological criteria for evaluation and licensure of new pneumococcal conjugate vaccine formulations for use in infants. Vaccine 21:3265–3272

Johnson SE, Rubin L, Romero-Steiner S, Dykes JK, Pais LB, Rizvi A, Ades E, Carlone GM (1999) Correlation of opsonophagocytosis and passive protection assays using human anticapsular antibodies in an infant mouse model of bacteremia for Streptococcus pneumoniae. J Infect Dis 180:133–140

Jumaan A, Yu O, Jackson L, Bohlke K, Galil K, Seward J (2005) Incidence of herpes zoster, before and after varicella-vaccination-associated decreases in the incidence of varicella, 1992-2002. J Infect Dis 191:2002−2009

Khromava AY, Eidex RB, Weld LH, Kohl KS, Bradshaw RD, Chen RT, Cetron MS (2005) Yellow fever vaccine: an updated assessment of advanced age as a risk factor for serious adverse events. Vaccine 23:3256–3263

Kitano H (2002) Computational systems biology. Nature 420:206–210

Kolibab K, Smithson SL, Rabquer B, Khuder S, Westerink MA (2005) Immune response to pneumococcal polysaccharides 4 and 14 in elderly and young adults: analysis of the variable heavy chain repertoire. Infect Immun 73:7465–7476

Lee W-W, Cui D, Czesnikiewicz-Guzik M, Vencio R, Shmulevich I, Aderem A, Weyand C, Goronzy J (2008) Age-dependent signature of metallothionein expression in primary CD4 T cell responses is due to sustained zinc signaling. Rejuvenation Res 11:1001–1012

Lescale C, Dias S, Maës J, Cumano A, Szabo P, Charron D, Weksler M, Dosquet C, Vieira P, Goodhardt M (2010) Reduced EBF expression underlies loss of B-cell potential of hematopoietic progenitors with age. Aging cell 9:410–419

Li Q-J, Chau J, Ebert P, Sylvester G, Min H, Liu G, Braich R, Manoharan M, Soutschek J, Skare P, Klein L, Davis M, Chen C-Z (2007) miR-181a is an intrinsic modulator of T cell sensitivity and selection. Cell 129:147–208

Lindsey NP, Schroeder BA, Miller ER, Braun MM, Hinckley AF, Marano N, Slade BA, Barnett ED, Brunette GW, Horan K, Staples JE, Kozarsky PE, Hayes EB (2008) Adverse event reports following yellow fever vaccination. Vaccine 26:6077–6082

Longo DM, Louie B, Putta S, Evensen E, Ptacek J, Cordeiro J, Wang E, Pos Z, Hawtin RE, Marincola FM, Cesano A (2012) Single-cell network profiling of peripheral blood mononuclear cells from healthy donors reveals age- and race-associated differences in immune signaling pathway activation. J Immunol 188:1717–1725

Martin M, Weld LH, Tsai TF, Mootrey GT, Chen RT, Niu M, Cetron MS (2001) Advanced age a risk factor for illness temporally associated with yellow fever vaccination. Emerg Infect Dis 7:945–951

Mazmanian SK, Kasper DL (2006) The love-hate relationship between bacterial polysaccharides and the host immune system. Nat Rev Immunol 6:849–858

McElhaney JE, Xie D, Hager WD, Barry MB, Wang Y, Kleppinger A, Ewen C, Kane KP, Bleackley RC (2006) T cell responses are better correlates of vaccine protection in the elderly. J Immunol 176:6333–6339

McKenna R, Washington L, Aquino D, Picker L, Kroft S (2001) Immunophenotypic analysis of hematogones (B-lymphocyte precursors) in 662 consecutive bone marrow specimens by 4-color flow cytometry. Blood 98:2498–3005

McMahon BJ, Williams J, Bulkow L, Snowball M, Wainwright R, Kennedy M, Krause D (1995) Immunogenicity of an inactivated hepatitis a vaccine in Alaska Native children and Native and non-Native adults. J Infect Dis 171:676–679

McMichael AJ, Gotch FM, Noble GR, Beare PA (1983) Cytotoxic T-cell immunity to influenza. N Engl J Med 309:13–17

Melegaro A, Edmunds WJ (2004) The 23-valent pneumococcal polysaccharide vaccine. Part I. Efficacy of PPV in the elderly: a comparison of meta-analyses. Eur J Epidemiol 19:353–363

Miller R, Jacobson B, Weil G, Simons E (1987) Diminished calcium influx in lectin-stimulated T cells from old mice. J Cell Physiol 132:337–379

Monath TP, Cetron MS, McCarthy K, Nichols R, Archambault WT, Weld L, Bedford P (2005) Yellow fever 17D vaccine safety and immunogenicity in the elderly. Hum Vaccin 1:207–214

Musher DM, Phan HM, Watson DA, Baughn RE (2000) Antibody to capsular polysaccharide of Streptococcus pneumoniae at the time of hospital admission for Pneumococcal pneumonia. J Infect Dis 182:158–167

Nakaya HI, Li S, Pulendran B (2012) Systems vaccinology: learning to compute the behavior of vaccine induced immunity. Wiley Interdiscip Rev Syst Biol Med 4:193–205

Nakaya HI, Wrammert J, Lee EK, Racioppi L, Marie-Kunze S, Haining WN, Means AR, Kasturi SP, Khan N, Li GM, McCausland M, Kanchan V, Kokko KE, Li S, Elbein R, Mehta AK, Aderem A, Subbarao K, Ahmed R, Pulendran B (2011) Systems biology of vaccination for seasonal influenza in humans. Nat Immunol 12:786–795

Naylor K, Li G, Vallejo A, Lee W-W, Koetz K, Bryl E, Witkowski J, Fulbright J, Weyand C, Goronzy J (2005) The influence of age on T cell generation and TCR diversity. J Immunol 174:7446–7498

Nikolich-Žugich J, Li G, Uhrlaub J, Renkema K, Smithey M (2012) Age-related changes in CD8 T cell homeostasis and immunity to infection. Semin Immunol (in press)

Nyugen J, Agrawal S, Gollapudi S, Gupta S (2010) Impaired functions of peripheral blood monocyte subpopulations in aged humans. J Clin Immunol 30:806–819

O'Shea J, Plenge R (2012) JAK and STAT Signaling Molecules in Immunoregulation and Immune-Mediated Disease. Immunity 36:542–550

Ortqvist A, Hedlund J, Burman LA, Elbel E, Hofer M, Leinonen M, Lindblad I, Sundelof B, Kalin M (1998) Randomised trial of 23-valent pneumococcal capsular polysaccharide vaccine in prevention of pneumonia in middle-aged and elderly people. Swedish Pneumococcal Vaccination Study Group. Lancet 351:399–403

Osterholm MT, Kelley NS, Sommer A, Belongia EA (2012) Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis. Lancet Infect Dis 12: 36–44.

Ouyang Q, Wagner W, Voehringer D, Wikby A, Klatt T, Walter S, Müller C, Pircher H, Pawelec G (2003) Age-associated accumulation of CMV-specific CD8 + T cells expressing the inhibitory killer cell lectin-like receptor G1 (KLRG1). Exp Gerontol 38:911–931

Oxman M, Levin M, Johnson G, Schmader K, Straus S, Gelb L, Arbeit R, Simberkoff M, Gershon A, Davis L, Weinberg A, Boardman K, Williams H, Zhang J, Peduzzi P, Beisel C, Morrison V, Guatelli J, Brooks P, Kauffman C, Pachucki C, Neuzil K, Betts R, Wright P, Griffin M, Brunell P, Soto N, Marques A, Keay S, Goodman R, Cotton D, Gnann J, Loutit J, Holodniy M, Keitel W, Crawford G, Yeh SS, Lobo Z, Toney J, Greenberg R, Keller P, Harbecke R, Hayward A, Irwin M, Kyriakides T, Chan C, Chan I, Wang W, Annunziato P, Silber J, Shingles Prevention Study G (2005) A vaccine to prevent herpes zoster and postherpetic neuralgia in older adults. N Engl J Med 352:2271–2355

Palache A, Beyer W, Lüchters G, Völker R, Sprenger M, Masurel N (1993) Influenza vaccines: the effect of vaccine dose on antibody response in primed populations during the ongoing interpandemic period. A review of the literature. Vaccine 11:892–1800

Panda A, Qian F, Mohanty S, van Duin D, Newman F, Zhang L, Chen S, Towle V, Belshe R, Fikrig E, Allore H, Montgomery R, Shaw A (2010) Age-associated decrease in TLR function in primary human dendritic cells predicts influenza vaccine response. J Immunol 184:2518–2545

Pang W, Price E, Sahoo D, Beerman I, Maloney W, Rossi D, Schrier S, Weissman I (2011) Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. Procl Natl Acad Sci USA 108:20012–20019

Park S, Nahm MH (2011) Older adults have a low capacity to opsonize pneumococci due to low IgM antibody response to pneumococcal vaccinations. Infect Immun 79:314–320

Pearce E, Walsh M, Cejas P, Harms G, Shen H, Wang L-S, Jones R, Choi Y (2009) Enhancing CD8 T-cell memory by modulating fatty acid metabolism. Nature 460:103–110

Pulendran B (2009) Learning immunology from the yellow fever vaccine: innate immunity to systems vaccinology. Nat Rev Immunol 9:741–747

Pulendran B, Li S, Nakaya HI (2010) Systems vaccinology. Immunity 33:516–529

Pulendran B, Miller J, Querec TD, Akondy R, Moseley N, Laur O, Glidewell J, Monson N, Zhu T, Zhu H, Staprans S, Lee D, Brinton MA, Perelygin AA, Vellozzi C, Brachman P, Jr., Lalor S, Teuwen D, Eidex RB, Cetron M, Priddy F, del Rio C, Altman J, Ahmed R (2008) Case of yellow fever vaccine–associated viscerotropic disease with prolonged viremia, robust adaptive immune responses, and polymorphisms in CCR5 and RANTES genes. J Infect Dis 198:500–507

Querec T, Akondy R, Lee E, Cao W, Nakaya H, Teuwen D, Pirani A, Gernert K, Deng J, Marzolf B, Kennedy K, Wu H, Bennouna S, Oluoch H, Miller J, Vencio R, Mulligan M, Aderem A, Ahmed R, Pulendran B (2008) Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol 10:116–141

Querec T, Bennouna S, Alkan S, Laouar Y, Gorden K, Flavell R, Akira S, Ahmed R, Pulendran B (2006) Yellow fever vaccine YF-17D activates multiple dendritic cell subsets via TLR2, 7, 8, and 9 to stimulate polyvalent immunity. J Exp Med 203:413–437

Reuman PD, Kubilis P, Hurni W, Brown L, Nalin D (1997) The effect of age and weight on the response to formalin inactivated, alum-adjuvanted hepatitis A vaccine in healthy adults. Vaccine 15:1157–1161

Rogers P, Croft M (1999) Peptide dose, affinity, and time of differentiation can contribute to the Th1/Th2 cytokine balance. J Immunol 163:1205–1218

Romero-Steiner S, Musher DM, Cetron MS, Pais LB, Groover JE, Fiore AE, Plikaytis BD, Carlone GM (1999) Reduction in functional antibody activity against Streptococcus pneumoniae in vaccinated elderly individuals highly correlates with decreased IgG antibody avidity. Clin Infect Dis 29:281–288

Rossi M, Yokota T, Medina K, Garrett K, Comp P, Schipul A, Kincade P (2003) B lymphopoiesis is active throughout human life, but there are developmental age-related changes. Blood 101:576–660

Roukens AH, Soonawala D, Joosten SA, de Visser AW, Jiang X, Dirksen K, de Gruijter M, van Dissel JT, Bredenbeek PJ, Visser LG (2011) Elderly subjects have a delayed antibody response and prolonged viraemia following yellow fever vaccination: a prospective controlled cohort study. PLoS One 6:e27753

Rubins JB, Puri AK, Loch J, Charboneau D, MacDonald R, Opstad N, Janoff EN (1998) Magnitude, duration, quality, and function of pneumococcal vaccine responses in elderly adults. J Infect Dis 178:431–440

Rudd B, Venturi V, Li G, Samadder P, Ertelt J, Way S, Davenport M, Nikolich-Žugich J (2011) Nonrandom attrition of the naive CD8 $^{+}$ T-cell pool with aging governed by T-cell receptor:pMHC interactions. Procl Natl Acad Sci USA 108:13694–13703

Sablan B, Kim D, Barzaga N, Chow W, Cho M, Ahn S, Hwang S, Lee J, Namini H, Heyward W (2012) Demonstration of safety and enhanced seroprotection against hepatitis B with investigational HBsAg-1018 ISS vaccine compared to a licensed hepatitis B vaccine. Vaccine 30:2689–2785

Sasaki S, He XS, Holmes TH, Dekker CL, Kemble GW, Arvin AM, Greenberg HB (2008) Influence of prior influenza vaccination on antibody and B-cell responses. PLoS One 3:e2975

Sasaki S, Sullivan M, Narvaez CF, Holmes TH, Furman D, Zheng NY, Nishtala M, Wrammert J, Smith K, James JA, Dekker CL, Davis MM, Wilson PC, Greenberg HB, He XS (2011) Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies. J Clin Invest 121:3109–3119

Sauce D, Larsen M, Fastenackels S, Duperrier A, Keller M, Grubeck-Loebenstein B, Ferrand C, Debré P, Sidi D, Appay V (2009) Evidence of premature immune aging in patients thymectomized during early childhood. J Clin Invest 119:3070–3078

Schenkein JG, Park S, Nahm MH (2008) Pneumococcal vaccination in older adults induces antibodies with low opsonic capacity and reduced antibody potency. Vaccine 26:5521–5526

Shapiro ED, Berg AT, Austrian R, Schroeder D, Parcells V, Margolis A, Adair RK, Clemens JD (1991) The protective efficacy of polyvalent pneumococcal polysaccharide vaccine. N Engl J Med 325:1453–1460

Shi Y, Yamazaki T, Okubo Y, Uehara Y, Sugane K, Agematsu K (2005) Regulation of aged humoral immune defense against pneumococcal bacteria by IgM memory B cell. J Immunol 175:3262–3267

Snelgrove R, Cornere M, Edwards L, Dagg B, Keeble J, Rodgers A, Lyonga D, Stewart G, Young D, Walker B, Hussell T (2012) OX40 ligand fusion protein delivered simultaneously with the BCG vaccine provides superior protection against murine Mycobacterium tuberculosis infection. J Infect Dis 205:975–1058

Staples JE, Gershman M, Fischer M (2010) Yellow fever vaccine: recommendations of the Advisory Committee on Immunization Practices (ACIP). MMWR Recomm Rep 59:1–27

Stiasny K, Aberle J, Keller M, Grubeck-Loebenstein B, Heinz F (2012) Age affects quantity but not quality of antibody responses after vaccination with an inactivated flavivirus vaccine against tick-borne encephalitis. PLoS One 7:e34145

Tarazona R, DelaRosa O, Alonso C, Ostos B, Espejo J, Peña J, Solana R (2000) Increased expression of NK cell markers on T lymphocytes in aging and chronic activation of the immune system reflects the accumulation of effector/senescent T cells. Mech Ageing Dev 121:77–165

Theeten H, Rumke H, Hoppener FJ, Vilatimo R, Narejos S, Van Damme P, Hoet B (2007) Primary vaccination of adults with reduced antigen-content diphtheria-tetanus-acellular pertussis or dTpa-inactivated poliovirus vaccines compared to diphtheria-tetanus-toxoid vaccines. Curr Med Res Opin 23:2729–2739

Tohme RA, Awosika-Olumo D, Nielsen C, Khuwaja S, Scott J, Xing J, Drobeniuc J, Hu DJ, Turner C, Wafeeg T, Sharapov U, Spradling PR (2011) Evaluation of hepatitis B vaccine immunogenicity among older adults during an outbreak response in assisted living facilities. Vaccine 29:9316–9320

Torling J, Hedlund J, Konradsen HB, Ortqvist A (2003) Revaccination with the 23-valent pneumococcal polysaccharide vaccine in middle-aged and elderly persons previously treated for pneumonia. Vaccine 22:96–103

Valenzuela H, Effros R (2002) Divergent telomerase and CD28 expression patterns in human CD4 and CD8 T cells following repeated encounters with the same antigenic stimulus. Clin Immunol 105:117–142

Van Damme P, Burgess M (2004) Immunogenicity of a combined diphtheria-tetanus-acellular pertussis vaccine in adults. Vaccine 22:305–308

Vaziri H, Schächter F, Uchida I, Wei L, Zhu X, Effros R, Cohen D, Harley C (1993) Loss of telomeric DNA during aging of normal and trisomy 21 human lymphocytes. Am J Hum Genet 52:661–668

Vermeulen J, Lange J, Tyring S, Peters P, Nunez M, Poland G, Levin M, Freeman C, Chalikonda I, Li J, Smith J, Caulfield M, Stek J, Chan I, Vessey R, Schödel F, Annunziato P, Schlienger K, Silber J (2012) Safety, tolerability, and immunogenicity after 1 and 2 doses of zoster vaccine in healthy adults $\geq$ 60 years of age. Vaccine 30:904–914

Voehringer D, Koschella M, Pircher H (2002) Lack of proliferative capacity of human effector and memory T cells expressing killer cell lectinlike receptor G1 (KLRG1). Blood 100:3698–4400

Wagar LE, Gentleman B, Pircher H, McElhaney JE, Watts TH (2011) Influenza-specific T cells from older people are enriched in the late effector subset and their presence inversely correlates with vaccine response. PLoS One 6:e23698

Wang J, Sun Q, Morita Y, Jiang H, Gross A, Lechel A, Hildner K, Guachalla L, Gompf A, Hartmann D, Schambach A, Wuestefeld T, Dauch D, Schrezenmeier H, Hofmann W-K, Nakauchi H, Ju Z, Kestler H, Zender L, Rudolph K (2012) A differentiation checkpoint limits hematopoietic stem cell self-renewal in response to DNA damage. Cell 148:1001–1015

Weinberg A, Zhang J, Oxman M, Johnson G, Hayward A, Caulfield M, Irwin M, Clair J, Smith J, Stanley H, Marchese R, Harbecke R, Williams H, Chan I, Arbeit R, Gershon A, Schödel F, Morrison V, Kauffman C, Straus S, Schmader K, Davis L, Levin M, Investigators USDoVACSPSPS (2009) Varicella-zoster virus-specific immune responses to herpes zoster in elderly participants in a trial of a clinically effective zoster vaccine. J Infect Dis 200:1068–1145

West DJ, Calandra GB (1996) Vaccine induced immunologic memory for hepatitis B surface antigen: implications for policy on booster vaccination. Vaccine 14:1019–1027

Weston WM, Friedland LR, Wu X, Howe B (2012) Vaccination of adults 65 years of age and older with tetanus toxoid, reduced diphtheria toxoid and acellular pertussis vaccine (Boostrix((R))): results of two randomized trials. Vaccine 30:1721–1728

WHO (2002) World population ageing: 1950–2050. World assembly on ageing report, pp 11–13

Wiley SR, Raman VS, Desbien A, Bailor HR, Bhardwaj R, Shakri AR, Reed SG, Chitnis CE, Carter D (2011) Targeting TLRs expands the antibody repertoire in response to a malaria vaccine. Sci Transl Med 3:93ra69

Wilkinson TM, Li CK, Chui CS, Huang AK, Perkins M, Liebner JC, Lambkin-Williams R, Gilbert A, Oxford J, Nicholas B, Staples KJ, Dong T, Douek DC, McMichael AJ, Xu XN (2012) Preexisting influenza-specific CD4 + T cells correlate with disease protection against influenza challenge in humans. Nat Med 18:274–280

Wolters B, Junge U, Dziuba S, Roggendorf M (2003) Immunogenicity of combined hepatitis A and B vaccine in elderly persons. Vaccine 21:3623–3628

Xu J, Vallejo A, Jiang Y, Weyand C, Goronzy J (2005) Distinct transcriptional control mechanisms of killer immunoglobulin-like receptors in natural killer (NK) and in T cells. J Biol Chem 280:24277–24362

Yager E, Ahmed M, Lanzer K, Randall T, Woodland D, Blackman M (2008) Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus. J Exp Med 205:711–734

Yu M, Li G, Lee W-W, Yuan M, Cui D, Weyand C, Goronzy J (2012) Signal inhibition by the dual-specific phosphatase 4 impairs T cell-dependent B-cell responses with age. Procl Natl Acad Sci USA 109:88

Systems Biology Analyses to Define Host Responses to HCV Infection and Therapy

Reneé C. Ireton and Michael Gale Jr.

Abstract While 170 million people worldwide are chronically infected with HCV, the response rate to the current treatment regimens of pegylated IFN-α (IFN) in combination with ribavirin is only approximately 55 % of all HCV patients undergoing therapy. This IFN-based therapy is now slated to serve as the backbone for future combination therapeutics involving direct-acting antiviral compounds, including HCV protease inhibitors, viral polymerase inhibitors, and other small molecules. It is essential that the application of IFN be improved for overall enhancement of therapy outcome to effectively cure HCV infection. Systems approaches, including genomics and network modeling, are particularly powerful tools that are now being used to dissect the underlying mechanisms of successful or failed treatment response in an effort to design improved IFN-based therapeutic regimens. Furthermore, systems applications can be used to define virus-host interactions and map their variation within viral and host genomes, leading to identification of targets for novel therapy strategies. Using these approaches, we have defined distinct hepatic expression and tissue distribution of innate immune signaling molecules and gene networks that associate with IFN-based treatment outcome for HCV infection. This chapter will focus on using systems approaches to understand the host response to both HCV infection and therapy to drive the development of improved HCV therapeutics.

R. C. Ireton · M. Gale Jr. (✉)
Department of Immunology, University of Washington School of Medicine,
1959 NE Pacific Street, Box 357650 Seattle, WA 98195, USA
e-mail: mgale@u.washington.edu

Current Topics in Microbiology and Immunology (2013) 363: 143–167
DOI: 10.1007/82_2012_251

Published Online: 18 August 2012

Contents

1 HCV Infection and Current Therapy

The millions of Hepatitis C virus (HCV) infections that occur annually around the globe are difficult to treat due to the ability of the virus to adapt to its only known natural hosts—humans. Hundreds and perhaps thousands of years of evolutionary time have allowed HCV to develop sophisticated and efficient ways to evade the human immune response, despite its small genome (Pybus et al. 2001, 2009; Smith et al. 1997). A positive strand RNA virus of the Flaviviridae family, HCV has a 9.6 kb genome that encodes a single 3,000 aa polyprotein. The translated HCV polyprotein is processed by host peptidase and viral proteases to produce: (1) structural proteins (Core, E1, and E2) that form new viral particles and, (2) nonstructural (NS) proteins that support viral RNA replication (Wieland and Chisari 2005). HCV infections are often insidious to the host, causing progressive liver damage, and are the leading indication of liver transplantation in the Western World (Hoofnagle 2002).

1.1 HCV Infection: Pathogenesis and Disease Outcome

HCV infections impact global health significantly, with an estimated 170 million people around the world chronically infected. Only 15–25 % of HCV infected

individuals spontaneously clear the virus. Most people who are exposed to the virus and develop an acute infection will progress to a chronic infection, which can have devastating repercussions on the health of the individual. Decades of uninhibited, ongoing virus replication in host hepatocytes typically results in chronic hepatitis, fibrosis, progressive cirrhosis, and an increased risk of liver failure and liver cancer (Seeff 2002). Liver tissue of chronically infected HCV patients typically contains infiltrates of CD4+ and CD8+ T lymphocytes, B lymphocytes, NK cells, NK T cells, and myloid cells, including Dendritic cells (DCs) and plasmacytoid DCs, which create a necro-inflammatory environment. Fibrosis develops within the areas of necro-inflammation, leading to cirrhosis and liver failure (Lloyd et al. 2007). Up to one-fifth of chronically infected patients develop end-stage liver disease, and are at high risk of developing hepatocellular carcinoma (Ikeda et al. 1998).

While HCV appears to primarily infect hepatocytes, traces of the virus have been detected in Kupffer and endothelial liver cells (Blight et al. 1994). Other peripheral tissues have also been found to contain the virus in infected hosts: including peripheral blood leukocytes, lymph nodes, and epithelial cells of the gut and brain (Forton et al. 2004; Cabot et al. 2000; Laskus et al. 2000; Deforges et al. 2004). However, infections occurring in nonliver tissue are not robust, as the virus does not appear to efficiently replicate to the point of effectively producing infectious virus in these peripheral tissues. Many potential HCV receptors and co-factors have been identified (LDLR, DC-SIGN, GAG, SRBI, tetraspanin CD81, claudin-1, and occludin) (Tan and He 2011). Despite a large effort to determine the precise cellular receptors and factors required for productive infection, only supplying all the known HCV receptors on murine cells results in virus entry, but not replication (Ploss et al. 2009). Therefore, host factors or replication conditions found specifically in human hepatocytes must contribute to HCV replication during infection.

1.2 Biology and Efficacy of Current Antiviral Therapy

Acute HCV infections are often not clinically diagnosed due to the lack of symptoms that would promote a clinic visit by the infected individual. However, once HCV infection has been identified, typically after clinical symptoms of liver dysfunction appear some years after an acute exposure, the current standard of care is to treat the patient with parenteral injection of pegylated interferon-α and the oral nucleoside analog Ribavirin over a typical 48-week treatment course. This treatment is based on the known antiviral activity of Type 1 Interferons (IFN) in general, and has been aggressively applied in various forms to HCV patients since 1986. IFNs are cytokines naturally produced by the host during virus infection, and they serve to trigger antiviral, anti-proliferative, and immunomodulatory host responses through the induction of hundreds of interferon-stimulated genes (ISGs). During treatment, administration of exogenous IFN does not directly act on the virus, but instead triggers the production of various ISGs that have antiviral

activity or impact lipid metabolism, proteolysis, apoptosis, and inflammation (Feld and Hoofnagle 2005). Furthermore, type-1 IFNs can promote cellular immune responses such as memory T cell proliferation, dendritic cell maturation, Natural killer cell proliferation, and have anti-apoptotic affects on T cells. The therapeutic effects of IFN are greatly enhanced by HCV patients who concurrently receive Ribavirin capsules orally on a daily basis. Ribavirin is a guanosine analog that is phosphorylated within the host cell. The mechanism of action of Ribavirin in combination therapy is not well-defined, but studies have indicated that it can accelerate the clearance of infected cells, reduce HCV infectiousness, amplify the IFN-α responses, and shift host immune responses to infection toward a Th1 response and away from a Th2 response (Pawlotsky 2009). Recent studies have also suggested that Ribavirin increases virus mutation rates to such an extent that it is thought to force the production of less fit viral species that are less able to escape host immune responses. However, clinical studies have found no evidence of accelerated HCV mutagenesis in HCV patients undergoing Ribavirin therapy (Chevaliez et al. 2007; Lutchman et al. 2007), suggesting that other, nonmutagenic mechanisms of action likely confer Ribavirin antiviral properties against HCV in vivo.

A perplexing complication of the current standard of care is that treatment response is impacted by the infecting HCV genotype of a given patient. To date, six major genotypes of HCV (HCV 1–6) have been classified and generally differ from each other by 30–35 % on the nucleotide level (Simmonds et al. 2005). Patients with genotype 1 infection exhibit a lower response rate to therapy (about 40 % response, at best), wherein patients infected with HCV genotype 2 or 3 exhibit a response rate of nearly 80 % (Hnatyszyn 2005). The factors determining treatment outcome among patients infected with different HCV genotypes are not well understood, but have been associated with specific virus-host interactions that control immune defenses against infection (Li et al. 2011; Asselah et al. 2010; Jouan et al. 2012).

In 2011, the FDA approved the use of two new direct-acting antiviral therapeutics, telaprevir and boceprevir, which inhibit the HCV NS3/4A protease, in the treatment of HCV infection. The inclusion of either of these two inhibitors into the standard IFN-α plus ribavarin treatment was able to significantly increase the sustained virological response rates in patients infected with HCV genotype 1 (McHutchison et al. 2009; Jacobson et al. 2011; Poordad et al. 2011; Kwo et al. 2010). However, despite these improved treatment effects, both drugs aggravate the standard of care treatment side effects (see below) and only seem to improve the responses of patients infected with genotype 1, thus limiting the overall improvements to HCV therapy.

1.3 Improved HCV Therapies are Needed

The need for developing improved HCV treatment therapies is obvious when considering infections from all HCV genotypes—overall only 55 % of HCV patients respond to treatment, while improvements to therapy using the new direct-

acting antivirals are currently limited to certain patient subsets. This disappointing efficacy rate is compounded by the harsh, undesirable side effects experienced by patients undergoing treatment. IFN-based, treatment-induced side effects such as fever, chills, muscle aches, joint pain, headaches, nausea, diarrhea, hair loss, and mental depression often make it exceedingly difficult for patients to complete the full recommended standard of therapy over the 48-week regimen. Whereas a shorter treatment duration of 24 weeks may have similar efficacy in some patients (see below), the side effects of treatment remain daunting and still serve to reduce therapy compliance. The indicated standard of therapy today is based on the genotype of the infecting virus. HCV patients infected with genotype 1 or 4 receive 1–3 injections of pegylated IFN weekly for 48 weeks, while patients infected with genotype 2 or 3 are recommended to undergo treatment for 24 weeks. Often patients cannot tolerate the unwanted side effects of interferon therapy, and prematurely discontinue the course of therapy, putting them at a higher risk of relapse than if they completed the therapy (Shiffman et al. 2007). With the advent of new direct-acting antivirals as therapeutics, it has become evident that the high mutation frequency within the population of virions in the host (see below, Sect. 1.4) can cause an HCV infection that is initially responsive to treatment to eventually become resistant (Pawlotsky 2011). Thus, despite the recent advances in therapy, novel HCV therapeutics are still urgently needed in the clinic to aid the millions of people worldwide who are coping with HCV infections.

1.4 HCV Evasion of the Host Response

The low frequency of HCV-infected individuals who are able to fully resolve the infection underscores the remarkable success of HCV to subvert the host response to infection. The host response overall comprises the innate and adaptive immune responses, and HCV infection dysregulates each to mediate a chronic infection course. Remarkably, by using only its genome and 10 mature proteins, HCV has the ability to establish long-term chronic infections in hepatocytes by interplaying with the host cellular machinery to replicate and produce progeny virions while actively evading the host innate and adaptive immune responses. HCV uses both genomic variability and its multifunctional proteome to evade and inhibit innate and adaptive immunity components of the host responses.

Extreme variation in the HCV genome is one powerful immune evasion mechanism. This large variability is caused by replication errors that generate the production of genetically-distinct, but closely related, viral genomes or "quasi-species". Like other RNA viruses, HCV replication is an error-prone process that generates 10^{-4} to 10^{-5} mutations per nucleotide per replication cycle (Pawlotsky 2006), which can produce an average of 10^{12} virions per day during infection (Neumann et al. 1998). Variable quasispecies along with high viral turnover allow HCV to evade host immune detection by helping prevent host immune

surveillance factors from detecting the virus and generating a strong immune response. Furthermore, the constant generation of genetically distinct quasispecies provides HCV a remarkable means to readily adapt to selective pressure applied by the natural host response or treatment with antiviral therapy. Recent studies have documented the ability of viral quasispecies in the HCV viral protease NS3/4A to evade immune detection and impart resistance to treatment, while also detailing individual variant impact on overall viral fitness (Verbinnen et al. 2010; Lopez-Labrador et al. 2008; Soderholm et al. 2006; Soderholm and Sallberg 2006; Susser et al. 2009; Xue et al. 2012; Welsch et al. 2012; Ruhl et al. 2011; Uebelhoer et al. 2008; Romano et al. 2010; Shimakami et al. 2011). In addition, viral quasispecies that have mutations in MHC class I-restricted epitopes can impact the ability of TCR to bind the epitope-bound MHC complex, thus blocking the host's ability to mount a significant T cell response (Timm et al. 2004; Bowen and Walker 2005b; Cox et al. 2005; Tester et al. 2005; Ray et al. 2005). Finally, variation in the HCV quasispecies may permit the virus to evade humoral immunity by containing mutations that prevent the generation of neutralizing antibody responses (Zhang et al. 2009). The extreme variation and continuous evolution of the HCV genome generates a high level of complexity when considering host–virus interactions and developing appropriately targeted therapeutics.

Another mechanism by which HCV evades host immunity is by antagonizing innate immune signaling to modulate host inflammatory and cytokine responses. HCV specifically targets the RIG-I dependent viral sensing program in a mechanism that ultimately inhibits the expression of α/β IFNs and ISGs generally responsible for limiting HCV replication and initiating mature humoral and cellular host immune responses (Liu and Gale 2010). RIG-I, a specific pathogen recognition receptor (PRR), detects a short ds RNA and/or poly-uridine motif of HCV (Saito et al. 2008) and induces downstream signaling via the CARD adaptor protein, termed MAVS (Sumpter et al. 2005) to induce IRF-3 activation and subsequent IFN production. HCV viral NS3/4A protease directly antagonizes this process cleaving MAVS and ablating RIG-I-dependent production of IFNα/β (Foy et al. 2003, 2005; Loo et al. 2006; Meylan et al. 2005).

HCV infection also imparts dysregulation of adaptive immunity through alteration of humoral immune programs, epitope drift among viral quasispecies, and imposing a state of immune exhaustion among antigen-responsive T cells (Bowen and Walker 2005a; Walker 2010). In terms of humoral immunity, constant genetic drift of HCV quasispecies leads to the outgrowth of antibody escape variants that can persist even in the face of a robust humoral immune response. Moreover, the virus-indued generation of anti-HCV antibodies leads to a HCV-typical pathology called cryoglobulin anemia in which viral antigen–antibody complexes deposit in the joints of the HCV patient where they form precipitates that mediate an inflammatory response. Engagement of HCV with surface CD81 and likely surface immunoglobulin on B cells also can potentiate B cell signaling and disposition to a proliferative phenotype linked to Non-Hodgins lymphoma that can appear in patients with chronic HCV infection (Hartridge-Lambert et al. 2012).

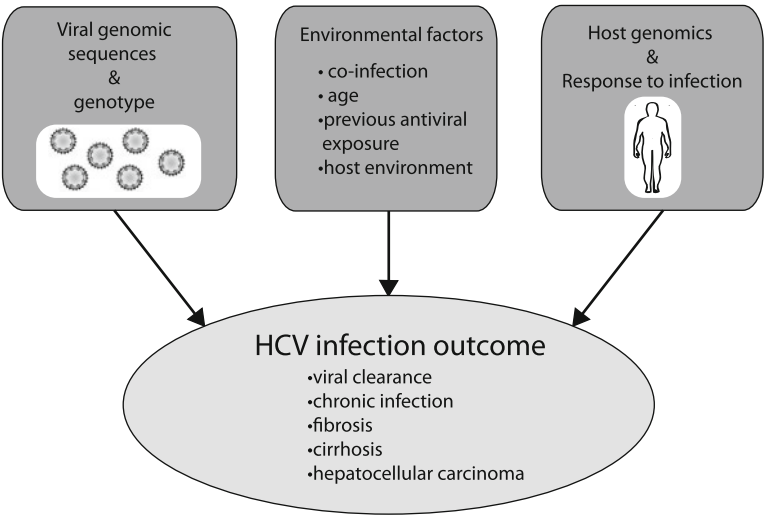


Fig. 1 Sources of biological variation during HCV infection that impact infection outcomes

Coupled with T cell exhaustion, the immunoregulatory features of HCV upon B cell and T cell effector actions serve to support chronic infection.

2 Overview of Systems Approaches to Understanding the Host Response to HCV Infection

In the decade since systems biology was first incorporated into the formal lexicon of biomedical research, use of systems approaches has expanded exponentially as considerable progress has been made in developing methods to generate measurements on a global scale and in advancing the computational framework to support systems analyses (Chaung et al. 2010). Arguably, the complex integration between virus and host that defines HCV infection outcome is ideal for the application of systems approaches for understanding the virus–host interactions that support HCV infection. While a systems approach is not always appropriate for all research applications in biomedicine, the encompassing global nature of such an approach holds the promise of significantly advancing our understanding of HCV infection and developing new therapies. HCV infections and outcomes are impacted by variations induced by a wide variety of biological sources (Fig. 1), each which can complicate the abstracting of conclusions made from nonsystems approaches from bench top to the clinic. An advantage of using a systems approach to understand HCV is that such an approach can embrace these large sources of biological variation and, at times, can harness it to generate a deeper understanding of the system. For example, a study comparing host genomes and

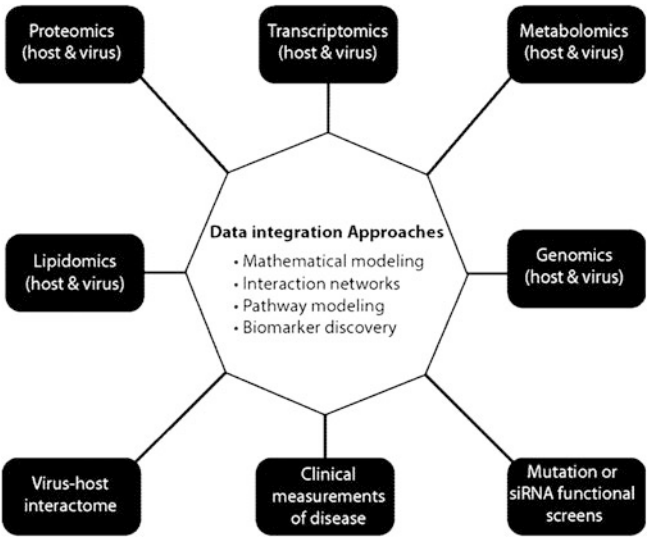


Fig. 2 Design of systems approaches

HCV infection outcomes can reveal novel host factors that impact HCV disease progression. By viewing HCV infection on a global scale, we can gain an expanded view of virus–host interactions— information that can be used in the practical development of improved HCV therapeutics. Systems approaches can help us fully characterize the mechanisms HCV uses to evade the host immune response. Furthermore, by globally defining the host response, we can broaden our perspective of the essential host response, which may lead to new avenues for HCV targeted therapy. Likewise, the study of biological variation during HCV infection, when viewed in a broad context, can allow us to define patterns that can be used clinically to predict disease course and clinical outcomes. As discussed in the sections below, systems approaches when applied to understanding the host response to HCV infection can contribute immensely to our ability to rationally design improved HCV therapeutics.

The definition of systems biology can vary widely (Chaung et al. 2010). Therefore, for our purposes here we will define systems biology as the generation of biological systems networks from the integration of data generated from conducting measurements across a system-wide level. Systems-wide measurements of HCV infection can be generated from a wide variety of platforms (Fig. 2), including genomics, proteomics, transcriptomics, and so on. This data can then be integrated to provide global models by generating network and mathematical models as well as to aid in biomarker discovery. In particular, our most pressing needs are to understand HCV-host interactions within the context of: (1) time (active versus chronic infection; innate response vs. cellular response during infection; and disease state), (2) geography and ethnicity, and (3) cellular

populations/tissues within the host. As described below, systems approaches are already being used to help define therapy outcome in HCV patients and improve our understanding of virus–host interactions.

3 Using Systems Biology Applications to Define Virus–host Interactions of Host Response Control

Our level of understanding HCV–host interactions has the potential to blossom through the implementation of systems approaches to address fundamental questions in HCV research. Few studies have taken a true systems approach as defined in the section above; however, as more scientists become fluent in using systems approaches, we expect this number to change dramatically. Below we describe some examples of how systems approaches or studies that incorporate systems-wide measurements have advanced our understanding of HCV–host interactions.

3.1 Defining Host Factors that Correlate with Disease Stage/Pathogenesis

One of the most perplexing aspects of HCV infection in the human population is the wide variation of disease outcomes post-infection. Systems approaches are ideal for teasing out this variation and can be used to define the underlying host factors that are associated with disease. In a study by Diamond et al. (2007), a quantitative nanoproteomics platform was used to identify differentially expressed proteins in HCV-infected liver tissue at different stages of fibrosis. After the identification of 210 proteins with expression profiles that associated with fibrosis stage, a functional pathway and network mapping of these proteins identified the dysregulation of two key cellular processes that were associated with fibrosis: the mitochondria processes of oxidative phosphorylation and fatty acid oxidation and the host response to oxidative stress. While both of these host cellular process have been linked to HCV infection by other approaches, the systems-base dataset can lay the foundation for additional systems studies that focus on the molecular mechanisms that link these pathway networks to liver disease progression.

Genomics platforms have also been used to identify host factors associated with disease progression. A functional genomics study of HCV and HCV/HIV co-infected individuals was able to identify a gene expression signature that separated a subset of patients from the group (Walters et al. 2006). Interestingly, the gene expression patterns in this subset were similar to the expression profiles obtained from HCV patients who developed fibrosis within 1 year of liver transplant (Smith et al. 2006). Functional analysis of the gene networks in this expression signature identified a downregulation of the FAS pathway and impaired type I and II IFN responses. A follow-up study by this group using a microarray

platform that allowed a wider transcriptomic coverage of samples from core needle liver biopsies was able to identify specific intrahepatic expression signatures that were associated with HIV/HCV co-infection in patients (Rasmussen et al. 2012b). Additional recent studies have focused on using genomic, proteomic, and computational analyses to identify molecular signatures that associate with disease progression in transplantation patients. Interestingly, a longitudinal transcriptional profiling study of liver biopsies from 57 HCV-infected patients by Rasmussen et al. found that patients who eventually develop the most severe liver disease demonstrated transcriptional profiles with broad repression of genes involved in immune responses, cell-cycle regulation, and antigen presentation and that these genomic alterations occur before liver disease progression could be detected by histology (Rasmussen et al. 2012a). A similar proteomics study by Diamond et al. on liver biopsy and serum samples from HCV-infected liver transplant patients found 250 differentially regulated proteins in patients with rapidly progressive fibrosis, with an enrichment of proinflammatory proteins and a decrease in proteins involved in detoxification of reactive oxidants. Furthermore, this study found that patients who develop severe liver injury have an altered amount of metabolites associated with oxidative stress in their serum, indicating a possible application for predicting early progression to fibrosis (Diamond et al. 2012). Such studies can be used as a starting point for biomarker studies that indicate disease progression or to provide the foundations for research agendas focused on understanding the underlying mechanisms that cause progression to fibrosis in select individuals.

3.2 Use of HCV Infection Model Systems to Identify Novel Host Responses to Infection

Since most acute HCV infections in humans go undetected, it is extremely difficult to study the early host response to infection in human subjects. Furthermore, applying a systems approach to human tissue samples is difficult due to the typical requirement of large amount of sample (both physical size and numbers of individual samples) to generate system-wide measurements. As an alternative, scientists have begun to use systems biology approaches to evaluate HCV infection models such as HCV replicon cells or chimpanzees and have identified novel host responses to infection. Of particular interest in these studies have been the innate immune response factors that likely shape the long-term immune response to long-term infection.

The development of HCV replicon cell lines (Lohmann et al. 1999) and HCV replication-permissive Huh-7 hepatoma cell lines revolutionized the HCV research field by enabling the in vitro evaluation of genomic HCV RNA replication. The application of high-throughput genomics characterizing these model systems has generated extensive resources for system biology approaches. Available genome-wide datasets specifically include host–virus interaction networks and gene expression changes induced by infection (Blackham et al. 2010; de Chassey et al.

2008; Nishimura-Sakurai et al. 2010). The generation of these resources provides a good foundation for implementing systems approaches to understanding HCV infection using in vitro infection model systems.

In 2011, MacPerson et al. used an interesting systems approach of the HCV replicon system to identify host factors that impact HCV replication (MacPherson et al. 2011). In this work, the group performed genomic analysis of cell lines that are hyperpermissive or resistant to HCV infection to define host factors that determine cellular permissiveness to infection. By overlaying this information with a proteomics study where they identified 236 host factors that are associated with the HCV replication complex in the membraneous web of infected cells, the authors were able to implicate that changes in the expression of APOE, DDOST, and PPIA may contribute to cellular resistance to infection. Using host–virus interaction networks, they were also able to identify a subset of the host replication factors from the proteomics screen (e.g., APOE and CALN) that interact with HCV proteins and are known to be involved in HCV production. From this work, the authors were able to put forth a candidate list of antiviral and proviral genes, some which were novel: tubulin-α (antiviral), NCEH1 (antiviral), and VSNL1 (proviral). Furthermore, functional classification of the genome-wide expression studies found that secreted glycoproteins are linked to HCV infection, and network analysis also provided evidence that host factors involved in protein folding, such as heat shock proteins, could be linked to HCV infection susceptibility and impact virus protein production. Other host factors that were linked to HCV infection susceptibility were involved in innate immune response, the secretion of signal peptides, and viral entry. Further detailed analysis of these factors and their related networks in HCV infection could lay the foundation for novel drug development platforms.

While in vitro systems can offer a platform with less variation, and thus can provide what could be considered cleaner datasets, they remain limited in scope. In vitro systems cannot provide information on how HCV functions in its natural infection environment where it is exposed to a milieu of different host cell populations nor do they provide a wide-angled view of how HCV infection interacts with the host on the level of the total organism. The application of systems approaches to chimpanzee models of HCV infection has the potential to provide a panoramic view of HCV infection over the course of infection as well as in discrete organ systems (e.g., immune system, liver etc.). Of particular interest is the fact that more than 60 % of chimpanzees inoculated with HCV are able to rapidly clear the virus, making them an interesting model for identifying host factors involved with viral clearance (Bassett et al. 1998, 1999; Lanford et al. 2001).

To date, most approaches using the chimpanzee model of HCV infection have involved only genome-wide transcriptional analyses. However, linking systems-wide transcriptional changes and viral quantities in particular tissues of the infected animal has provided new insights into the systematic host response triggered by HCV infection. An initial study of chimpanzee host response to HCV infection documented genomic changes and viral RNA in liver biopsies as well as serum levels of ALT, viral RNA, and HCV antibodies over both early (2-days post-infection) and later time points (up to 14-weeks post-infection) during the course of infection

(Bigger et al. 2001). While only a single chimpanzee, which effectively cleared the virus, was evaluated in the study, the authors were able to track the course of infection, and found that the virus was cleared from the blood between weeks 6 and 8 post-infection, an event which corresponded to seroconversion for anti-HCV antibodies. Interestingly, despite being cleared in the blood, the virus remained in infected hepatocytes until week 14 post-infection. Functional profiling of the genomics data found that IFN response genes could be detected as early as 2 days post-infection and that overall IFN response genes fell into three patterns of expression: (1) peak early at day 7 post-infection, then declined, (2) peak late at week 6 post-infection, and (3) peak early and sustained until viremia was cleared. These results suggest that different regulatory pathways and/or cell populations may be contributing to the IFN response over the course of infection. The authors observed that the peak serum ALT levels did not coincide with the declines in the viral RNA in the serum or liver, suggesting that viral clearance was not associated with extensive hepatocellular death.

A different study with a broader scope of chimpanzee response to HCV infection was conducted (Su et al. 2002) with the evaluation of host responses from three HCV-infected chimpanzees with three different infection outcomes: persistent infection, transient viral clearance, and sustained clearance. Functional analysis of the genes unique to each outcome identified IFN-γ-induced genes and genes involved with antigen processing, antigen presentation, and adaptive immune responses that were uniquely associated with transient or sustained viral clearance. When evaluating the early gene expression changes in the persistently infected chimpanzee versus the chimpanzees that had transient and sustained viral clearance, the authors of this study noted that genes associated with lipid metabolism were correlated with the onset of viremia in the transient and sustained viral clearance outcomes and the initial increase in HCV RNA levels. The group went on to validate this finding in HCV replicon systems, finding that small molecules designed to perturb lipid metabolism can modulate HCV replication and possibly identifying a new area of exploration for novel antiviral therapies.

3.3 miRNA–mRNA Host Networks in HCV Infection

A novel computational approach to understanding host networks activated during HCV infection was recently taken with the microarray and computational profiling of micro-RNA (miRNA) in liver tissues from HCV infected individuals (Peng et al. 2009). Micro-RNAs, a class of small noncoding RNA molecules that regulate gene expression, have been recently implicated in HCV infection. Some micro-RNAs, such as miR-122, have been found to be required for HCV RNA replication in the liver (Jopling et al. 2005), while others, such as miR-196 and miR-488 can directly prevent viral replication (Pedersen et al. 2007). The approach taken by Peng et al. provides a systematic profiling of host miRNA expression during HCV infection and allowed the identification of mi-RNA associated regulatory networks that were associated with HCV infection. In this approach, Peng et al., profiled the expression

of miRNAs and mRNAs in uninfected and HCV-infected human liver tissue samples. By combining the inverse expression patterns between the miRNAs and mRNAs as well as computational prediction of miRNA binding targets, they were able to identify 38 miRNA–mRNA regulatory modules. Biological functions of these identified regulatory modules were extrapolated from functional analysis of the predicted miRNA targets using a protein-interaction network. Overall, these biological functions included innate immune responses, cell cycle check point, and negative regulation of the initiation of translation. miRNA expression analyses will continue to play a part in system approaches to understanding HCV infection and even have the potential to be developed into new HCV therapies.

3.4 Exploring the Impact of Host Response on the HCV Genome

Genetic drift of HCV to escape host immune responses is a well-documented event that occurs during HCV infections, but the underlying host factors that shape that genetic drift are not as clearly defined. Recent studies using the chimpanzee model of HCV host infection have characterized the host response while simultaneously sequencing the virus genome over time. Such an experimental design has shed some light on the host-driven impact on viral evolution. A study of MHC I and II restricted epitopes in persistently infected chimpanzees revealed that despite amino acid changes in the NS3 protein that caused decreased activation, proliferation, and cytokine production by epitope-specific CD4 T cells in these animals, these changes were uncommon (Fuller et al. 2010). Instead, in each individual, the frequency of mutational escapes in MHC class II-restricted epitopes is much less common than class I-restricted epitopes, indicating that the lack of CD4 T cell response in persistent HCV infections is not caused by virus escaping CD4 detection. A similar study that characterized the CD8 T cell response and viral genome over the course of infection found that the ratio of nonsynonymous to synonymous mutations, which is a measure of selective pressure, increased 50-fold in class I-restricted epitopes compared to the rest of the HCV genome (Callendret et al. 2011). This finding suggests that CD8 T cells exert a strong selective pressure on the viral evolution of HCV during infection. Thus, wide-angled views provided by systems approaches such as those described above are contributing much to our understanding globally how host and HCV factors impact each other during the course of infection.

4 Systems Approaches to Defining/Predicting Therapy Outcomes in HCV Patients

With the revolution of high-throughput, high resolution, biological assays requiring increasingly miniscule amounts of biological specimens to generate measurements, the age of individualized medicine is looming ever closer. While the ability of the

clinician to use patient-specific genome or proteome information to determine the best treatment options is still in the future, systems biology approaches have begun to inch the concept of individualized medicine into a reality. As noted above in the introduction to this chapter, the low rate of treatment responses by HCV patients to the current approved HCV therapies, especially among those infected with specific HCV genotypes, indicates that variation within the infecting host or virus may account for HCV treatment failures. By nature, systems approaches can harness these sources of variation and use them to identify host or virus factors that may impact treatment outcomes. Below we highlight a number of studies that have used systems approaches to evaluate the various therapeutic outcomes in patients undergoing HCV therapies.

4.1 Systems Approaches to Understanding Interferon Therapy

We know that interferon-based therapies are effective in achieving sustained viral response in some, but not all, HCV patients. Therefore, understanding how the host systems respond to this therapy will provide a huge benefit to discerning how this response can effectively overcome infection. For this reason, several studies have used systems approaches to characterize IFN response in both human and chimpanzee model systems.

A seminal study by Katze and co-workers used isotope-encoded affinity tag ICAT-based proteomics to identify IFN-regulated proteins in the human hepatoma Huh7 cell line (Yan et al. 2004). In this work, they performed a global quantitative proteomic analysis of Huh7 cell extracts that had been cultured in the absence or presence of IFN. This proteomics approach identified 1,364 proteins in the cells. By overlaying the ICAT quantification data from the proteomics study with a genomics dataset of genes that contain the interferon-stimulated response element (a signature of IFN-inducible genes), the authors identified 78 proteins that were likely regulated by IFN. The application of data mining tools such as gene ontology (GO) and Cytoscape allowed them to determine that the identified proteins that increased with IFN treatment tended to be involved in antiviral defense and immune response signaling networks while proteins that decreased with IFN tended to be involved with metabolism and growth. Interestingly, a novel observation from this work was that a number of the IFN-induced proteins were involved with G-protein coupled signaling pathways, suggesting that IFN treatment may impact these pathways. Overall, this approach identified 39 novel proteins that were previously unknown to be interferon responsive. However, the findings from this study may have limitations in their practical application to clinical advances as they were based on measurements generated from immortalized cell lines that are known to have inherent cell signaling defects.

The limitation of using cell lines as an infection model system has been addressed with a similar genomics analyses that was performed in primary chimpanzee tissues treated with IFN (Lanford et al. 2006). A strength of this study

is that it assessed the kinetics of the transcriptional response to IFN-α. By conducting genomic measurements across tissue types (e.g., liver and PBMCs) and with ex vivo systems (primary chimpanzee and human hepatocytes), the authors were able to track tissue and cell-type specific transcriptional changes induced by exposure to IFN. This work demonstrated that the IFN-induced response is rapidly downregulated in vivo, is indistinguishable between chimpanzee and humans, and was tissue specific. Such a study helps us put together how antiviral therapy responses impact the viral life cycle and shape the observed kinetics of viral clearance during therapy. Studies of viral infection kinetics have noted that viral titers tend to decrease significantly in the first 24–48 h of therapy, after which they generally rise again. These two events could correlate with first the quick induction of ISGs after the initial IFN treatment, which would impact virus replication, and the gradual resurgence of virus could correspond to the downregulation of IFN responsiveness.

4.2 Harnessing Host Genomics to Predict Treatment Outcome

Systems approaches are beginning to help us link individual host responses to predict treatment response. Recent advances in genomics technologies, in which a half-million variations in individual genomes can be compared, have increased the resolution in genomics studies and enhanced our ability to detect variation. For example, in 2009, the results of genome-wide association studies designed to identify genomic variations linked to HCV treatment outcomes identified SNPs near the IL-28B gene as strong predictors of treatment success in patients infected with HCV genotype 1 (Ge et al. 2009; Tanaka et al. 2009; Suppiah et al. 2009). Since then, allelic variants in IL28B have been linked to natural clearance of HCV and the outcome of liver transplant patients (Thomas et al. 2009; Coto-Llerena et al. 2011). Controversial studies have also indicated that IL-28B alleles also may be associated with HCV viral load and progression to cirrhosis, but additional studies are needed to confirm these connections (Soriano et al. 2012).

The exact mechanism by which IL-28B alleles influence HCV treatment response is unclear, but the fact that IL-28B encodes IFNλ-3, a type III interferon that is induced during virus infection and stimulates the production of antiviral ISGs, is intriguing. Several groups using gene expression profiling have found a correlation between hepatic expression of ISGs and treatment response (Asselah et al. 2005; Chen et al. 2005; Feld and Hoofnagle 2005). Low baseline ISG expression indicates a positive response to interferon therapy and suppression of viral loads, whereas upregulation prior to therapy is predictive of a treatment failure. Indeed, pre-assessment of future nonresponder gene expression profiles found that these patients have maximal baseline ISG expression levels that remain flat with IFN treatment (Sarasin-Filipowicz et al. 2008). However, whether the relationship between ISG expression and IL-28B gene expression is a causal or independent predictor of treatment outcome remains to be fully resolved (McGilvray et al. 2012). Thus, future

studies are needed to better understand the exact link between IL-28B SNPs and HCV treatment outcomes.

Additional studies have used the plethora of genomics data to determine gene expression patterns with predictive value for HCV therapeutic outcomes. Compared to using IL-28B genotyping alone as a predictor of treatment outcome, the gene classifiers were better at indicating therapeutic response (Dill et al. 2011; McGilvray et al. 2012). Interestingly, one study found that immunostaining levels of the human myxovirus A protein 1 (MxA) in macrophages of liver biopsies from HCV patients had an exceptionally high negative predictive value for treatment outcome and was inversely correlated to ISG expression in hepatocytes, indicating the potential importance of cellular crosstalk within tissues during treatment (McGilvray et al. 2012). In the future, better predictive models of HCV therapeutic outcomes will be built as more classifiers are identified from -omics studies and plugged into the existing models or developed into new, independent models.

Few studies have truly employed systems approaches to understand HCV therapy. However, some recent work has demonstrated the power in using such an approach (Lau et al. 2012). To better understand the host response to acute IFN-α treatment in HCV-infected patients, Lau and co-workers integrated datasets from multiple data gathering platforms: gene expression profiling of hepatocytes and PBMCs from treated patients, serum viral kinetics, bioinformatic analysis, and mathematical modeling of viral decay. In this study, they profiled the acute response to IFN-α treatment in eight HCV patients who were chronically infected with HCV genotype 1. Patients in the study were classified as rapid virological responders (RVRs), early virological responders (EVR), and nonresponders (NR) based on HCV RNA kinetics that were measured in the initial 12 weeks of therapy. As noted in other studies, genomic expression profiles of the NR had a high basal level of ISG expression in pre-treatment samples compared to the EVRs and RVRs. Pathway modeling of the identified differentially expressed genes in the NR samples indicates that Stat-1 is a central regulatory node for the high ISG basal expression in these patients. Furthermore, NR patient liver samples analyzed 24 h after IFN treatment displayed very little differences in their gene expression patterns compared to the pre-treatment samples. In contrast, a similar evaluation of EVR and RVR patient samples revealed a large number of genes displayed significant changes in expression after IFN treatment in a pattern that differentiated these samples from the NR-derived samples. Importantly, the NR patients had a high "set point" ISG expression pattern pre-treatment that lacked further induction with IFN treatment. Network modeling of EVR and RVR patient responses to IFN indicates that the increase in IFN-α signaling occurs through an amplification of IRF-7 signaling (Fig. 3).

Interestingly, when Lau et al. completed a similar analysis on PBMCs and then compared these data to the genomics results obtained from the liver of the same patient, they discovered that, unlike the liver tissues, pre-treatment ISGs in PBMCs were similar between the NR and the SVR and EVR, with little ISG expression in the pre-treatment samples and then a strong induction in the first 3–12 h. Therefore, both the pre-treatment ISG set-point and the acute responses to

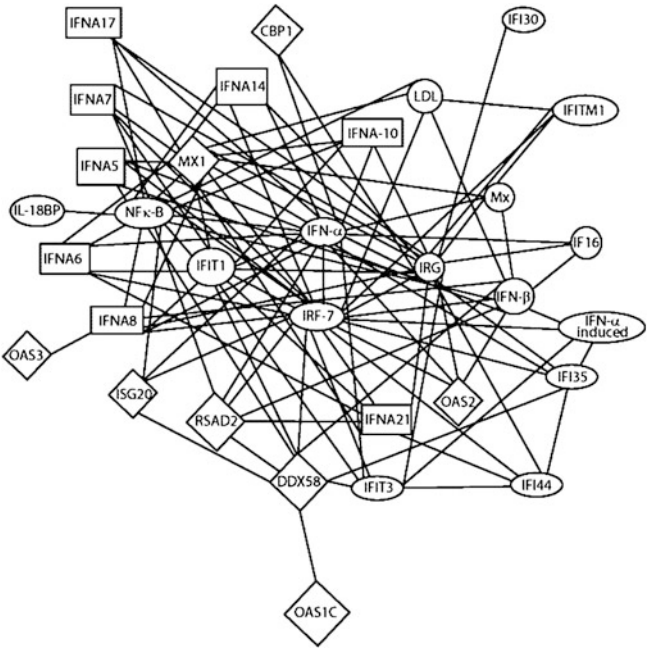


Fig. 3 Summary of pathway modeling to reveal processes of gene regulation of ISG setpoint among NR patients with chronic HCV infection

IFN treatment are different between PBMCs and liver tissue, indicating a tissue compartmentalization of the IFN-induced response.

The high expression of specific ISG pre- and post-IFN treatment in NR patients was confirmed by immunohistochemical analysis of HCV patient liver tissues. With these studies, Lau et al. also observed that the tissues had three distinct staining patterns that were associated with treatment outcome. EVR and RVR patient tissues had a strong staining pattern that occurred through the majority of the hepatocytes in the tissue sample. NR patients had one of two staining patterns: (1) a "cell specific" response where adjacent cells expressed different ISG levels, or (2) a "focal" response where only response foci displayed different levels of ISG. Importantly, the ISG expression level differences were observed between the hepatocytes and the liver-resident macrophages, called Kupffer cells. Further analysis of these tissues showed that Kupffer cells were expressing IFN-β, and when combined with the results from other studies, indicates that Kupffer cells may take up HCV and trigger IFN-β expression. From this work, it was proposed that endogenous expression of hepatic IFN drives a high ISG set point, which would permit a state of cellular tolerance to IFN and impact treatment outcome. Indeed, when constitutive IFN exposure of HCV was modeled in an HCV replicon cell system it was found that IFN-induced cellular responses were blunted with periodic IFN exposure, although the expression of IFN-α/β receptor levels remained stable. In other studies, Chisari and co-workers found that HCV can similarly stimulate IFN production from

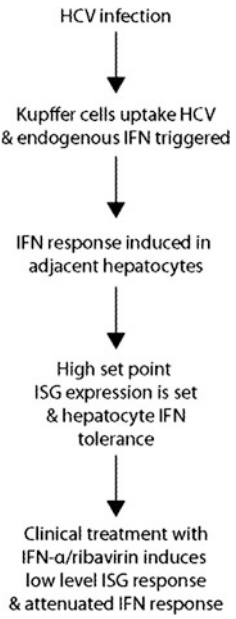


Fig. 4 Model of innate immune tolerance in chronic HCV patients undergoing treatment, based on data from Lau et al. Endogenous IFN-β is produced by Kupffer cells or other myeloid cells, such a plasmacytoid DCs, IFN then drives paracrine ISG expression in hepatocytes and within the liver. This creates a high set point of ISG expression in the liver and innate immune tolerance of IFN (Takahashi et al. 2010; Lau et al. 2012)

plasmacytoid DC via direct cell interaction and exosomal transfer of HCV RNA to the DC. In this case, the viral RNA was engaged by TLR7 to drive IFN production (Takahashi et al. 2010). This model reveals that DCs can also take up HCV RNA and produce IFN as a result. Thus, resident macrophages/Kupffer cells as well as plasmacytoid DCs and other DC subsets likely serve to produce IFN locally in the liver. This hepatic IFN then drives ISG expression to impart innate immune tolerance to the actions of therapeutic IFN, rendering a reduced efficacy of antiviral therapy against HCV infection. Together, systems biology approaches to defining hepatic host responses suggest that chronic exposure to IFN may contribute to a state of tolerance to IFN-α, preventing the hepatocyte from fully suppressing HCV and leading to treatment failure (Fig. 4). Importantly, these studies demonstrate how the careful application of systems approaches can contribute to our understanding of disease and the appropriate design of treatment options.

5 Future Impact of Systems Biology on HCV Therapy Design

Since the advent of systems biology into biomedical research, we have made significant initial progress in using this powerful approach to understanding host response to HCV infection and improving HCV therapies. This early work has generated glimpses of the systematic changes that occur in the host during HCV infection and treatment. However, much work remains to be accomplished in the HCV field and systems approaches will no doubt have a major impact in shaping

how we understand HCV infection and treat HCV disease. Likely, systems discoveries in HCV host response and infection outcome will impact: biomarker discovery, the development of novel therapies, clinical treatment procedures (i.e., diagnosis, predicting course of infection and outcome, and monitoring treatment response), and vaccine development (correlates of immunity).

Challenges remain in the effective implementation of information generated from systems approaches into our basic understanding of HCV. Data analysis using systems approaches are, by nature, complex and the huge amounts of output can make the generation of overarching conclusions difficult. Many current systems studies only provide an expanded, wide angle view of HCV infection, making it difficult to hone-in on specific areas that can practically be applied in HCV therapy development. Rather than simply providing a holistic view of infection, future studies using systems approaches need to focus on identifying specific factors that impact infection outcome. Only when this laser focus of intent is used to filter through the noisy data produced by systems approaches, will we realize the full force of systems biology in developing powerful HCV therapeutics and treatment strategies.

Acknowledgments RI and MG are supported by NIH grants AI060389, AI88778, DA024563, and NIH Contract 27220090035C.

References

Asselah T, Bieche I, Laurendeau I, Paradis V, Vidaud D, Degott C, Martinot M, Bedossa P, Valla D, Vidaud M, Marcellin P (2005) Liver gene expression signature of mild fibrosis in patients with chronic hepatitis C. Gastroenterology 129(6):2064–2075. doi:10.1053/j.gastro.2005.09.010, S0016-5085(05)01794-4 [pii]

Asselah T, Estrabaud E, Bieche I, Lapalus M, De Muynck S, Vidaud M, Saadoun D, Soumelis V, Marcellin P (2010) Hepatitis C: viral and host factors associated with non-response to pegylated interferon plus ribavirin. Liver Int 30(9):1259–1269. doi:10.1111/j.1478-3231.2010.02283.x, LIV2283 [pii]

Bassett SE, Brasky KM, Lanford RE (1998) Analysis of hepatitis C virus-inoculated chimpanzees reveals unexpected clinical profiles. J Virol 72(4):2589–2599

Bassett SE, Thomas DL, Brasky KM, Lanford RE (1999) Viral persistence, antibody to E1 and E2, and hypervariable region 1 sequence stability in hepatitis C virus-inoculated chimpanzees. J Virol 73(2):1118–1126

Bigger CB, Brasky KM, Lanford RE (2001) DNA microarray analysis of chimpanzee liver during acute resolving hepatitis C virus infection. J Virol 75(15):7059–7066. doi:10.1128/JVI.75.15.7059-7066.2001

Blackham S, Baillie A, Al-Hababi F, Remlinger K, You S, Hamatake R, McGarvey MJ (2010) Gene expression profiling indicates the roles of host oxidative stress, apoptosis, lipid metabolism, and intracellular transport genes in the replication of hepatitis C virus. J Virol 84(10):5404–5414. doi:10.1128/JVI.02529-09, JVI.02529-09 [pii]

Blight K, Lesniewski RR, LaBrooy JT, Gowans EJ (1994) Detection and distribution of hepatitis C-specific antigens in naturally infected liver. Hepatology 20(3):553–557. 0270-9139(94)90087-6 [pii]

Bowen DG, Walker CM (2005a) Adaptive immune responses in acute and chronic hepatitis C virus infection. Nature 436(7053):946–952. doi:10.1038/nature04079, nature04079 [pii]

Bowen DG, Walker CM (2005b) Mutational escape from CD8+ T cell immunity: HCV evolution, from chimpanzees to man. J Exp Med 201(11):1709–1714. doi:10.1084/jem.20050808, jem.20050808 [pii]

Cabot B, Martell M, Esteban JI, Sauleda S, Otero T, Esteban R, Guardia J, Gomez J (2000) Nucleotide and amino acid complexity of hepatitis C virus quasispecies in serum and liver. J Virol 74(2):805–811

Callendret B, Bukh J, Eccleston HB, Heksch R, Hasselschwert DL, Purcell RH, Hughes AL, Walker CM (2011) Transmission of clonal hepatitis C virus genomes reveals the dominant but transitory role of CD8(+) T cells in early viral evolution. J Virol 85(22):11833–11845. doi:10.1128/JVI.02654-10, JVI.02654-10 [pii]

Chaung HY et al (2010) A decade of systems biology. Annu Rev Dev Biol 26:721

Chen L, Borozan I, Feld J, Sun J, Tannis LL, Coltescu C, Heathcote J, Edwards AM, McGilvray ID (2005) Hepatic gene expression discriminates responders and nonresponders in treatment of chronic hepatitis C viral infection. Gastroenterology 128(5):1437–1444 S0016508505003999 [pii]

Chevaliez S, Brillet R, Lazaro E, Hezode C, Pawlotsky JM (2007) Analysis of ribavirin mutagenicity in human hepatitis C virus infection. J Virol 81(14):7732–7741. doi:10.1128/JVI.00382-07, JVI.00382-07 [pii]

Coto-Llerena M, Perez-Del-Pulgar S, Crespo G, Carrion JA, Martinez SM, Sanchez-Tapias JM, Martorell J, Navasa M, Forns X (2011) Donor and recipient IL28B polymorphisms in HCV-infected patients undergoing antiviral therapy before and after liver transplantation. Am J Transplant 11(5):1051–1057. doi:10.1111/j.1600-6143.2011.03491.x

Cox AL, Mosbruger T, Mao Q, Liu Z, Wang XH, Yang HC, Sidney J, Sette A, Pardoll D, Thomas DL, Ray SC (2005) Cellular immune selection with hepatitis C virus persistence in humans. J Exp Med 201(11):1741–1752. doi:10.1084/jem.20050121, jem.20050121 [pii]

de Chassey B, Navratil V, Tafforeau L, Hiet MS, Aublin-Gex A, Agaugue S, Meiffren G, Pradezynski F, Faria BF, Chantier T, Le Breton M, Pellet J, Davoust N, Mangeot PE, Chaboud A, Penin F, Jacob Y, Vidalain PO, Vidal M, Andre P, Rabourdin-Combe C, Lotteau V (2008) Hepatitis C virus infection protein network. Mol Syst Biol 4:230. doi:10.1038/msb.2008.66 , msb200866 [pii]

Deforges S, Evlashev A, Perret M, Sodoyer M, Pouzol S, Scoazec JY, Bonnaud B, Diaz O, Paranhos-Baccala G, Lotteau V, Andre P (2004) Expression of hepatitis C virus proteins in epithelial intestinal cells in vivo. J Gen Virol 85(Pt 9):2515–2523. doi:10.1099/vir.0.80071-0 85/9/2515 [pii]

Diamond DL, Jacobs JM, Paeper B, Proll SC, Gritsenko MA, Carithers RL Jr, Larson AM, Yeh MM, Camp DG 2nd, Smith RD, Katze MG (2007) Proteomic profiling of human liver biopsies: hepatitis C virus-induced fibrosis and mitochondrial dysfunction. Hepatology 46(3):649–657

Diamond DL, Krasnoselsky AL, Burnum KE, Monroe ME, Webb-Robertson BJ, McDermott JE, Yeh MM, Dzib JF, Susnow N, Strom S, Proll SC, Belisle SE, Purdy DE, Rasmussen AL, Walters KA, Jacobs JM, Gritsenko MA, Camp DG, Bhattacharya R, Perkins JD, Carithers RL Jr, Liou IW, Larson AM, Benecke A, Waters KM, Smith RD, Katze MG (2012) Proteome and computational analyses reveal new insights into the mechanisms of hepatitis C virus-mediated liver disease post transplantation. Hepatology 56(1):28–38 doi:10.1002/hep.25649

Dill MT, Duong FH, Vogt JE, Bibert S, Bochud PY, Terracciano L, Papassotiropoulos A, Roth V, Heim MH (2011) Interferon-induced gene expression is a stronger predictor of treatment response than IL28B genotype in patients with hepatitis C. Gastroenterology 140(3):1021–1031. doi:10.1053/j.gastro.2010.11.039, S0016-5085(10)01729-4 [pii]

Feld JJ, Hoofnagle JH (2005) Mechanism of action of interferon and ribavirin in treatment of hepatitis C. Nature 436(7053):967–972. doi:10.1038/nature04082, nature04082 [pii]

Forton DM, Karayiannis P, Mahmud N, Taylor-Robinson SD, Thomas HC (2004) Identification of unique hepatitis C virus quasispecies in the central nervous system and comparative

analysis of internal translational efficiency of brain, liver, and serum variants. J Virol 78(10):5170–5183

Foy E, Li K, Wang C, Sumpter R, Jr., Ikeda M, Lemon SM, Gale M, Jr. (2003) Regulation of interferon regulatory factor-3 by the hepatitis C virus serine protease. Science 300(5622):1145–1148. doi:10.1126/science.10826041082604 [pii]

Foy E, Li K, Sumpter R Jr, Loo YM, Johnson CL, Wang C, Fish PM, Yoneyama M, Fujita T, Lemon SM, Gale M Jr (2005) Control of antiviral defenses through hepatitis C virus disruption of retinoic acid-inducible gene-I signaling. Proc Natl Acad Sci U S A 102(8): 2986–2991. doi:10.1073/pnas.0408707102, 0408707102 [pii]

Fuller MJ, Shoukry NH, Gushima T, Bowen DG, Callendret B, Campbell KJ, Hasselschwert DL, Hughes AL, Walker CM (2010) Selection-driven immune escape is not a significant factor in the failure of CD4 T cell responses in persistent hepatitis C virus infection. Hepatology 51(2):378–387. doi:10.1002/hep.23319

Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, Heinzen EL, Qiu P, Bertelsen AH, Muir AJ, Sulkowski M, McHutchison JG, Goldstein DB (2009) Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 461(7262):399–401. doi:10.1038/nature08309, nature08309 [pii]

Hartridge-Lambert SK, Stein EM, Markowitz AJ, Portlock CS (2012) Hepatitis C and non-Hodgkin lymphoma: the clinical perspective. Hepatology 55(2):634–641. doi:10.1002/hep.25499

Hnatyszyn HJ (2005) Chronic hepatitis C and genotyping: the clinical significance of determining HCV genotypes. Antivir Ther 10(1):1–11

Hoofnagle JH (2002) Course and outcome of hepatitis C. Hepatology 36(5 Suppl 1):S21–S29. doi:10.1053/jhep.2002.36227, S0270913902001684 [pii]

Ikeda K, Saitoh S, Suzuki Y, Kobayashi M, Tsubota A, Koida I, Arase Y, Fukuda M, Chayama K, Murashima N, Kumada H (1998) Disease progression and hepatocellular carcinogenesis in patients with chronic viral hepatitis: a prospective observation of 2215 patients. J Hepatol 28(6):930–938. S0168-8278(98)80339-5 [pii]

Jacobson IM, McHutchison JG, Dusheiko G, Di Bisceglie AM, Reddy KR, Bzowej NH, Marcellin P, Muir AJ, Ferenci P, Flisiak R, George J, Rizzetto M, Shouval D, Sola R, Terg RA, Yoshida EM, Adda N, Bengtsson L, Sankoh AJ, Kieffer TL, George S, Kauffman RS, Zeuzem S (2011) Telaprevir for previously untreated chronic hepatitis C virus infection. N Engl J Med 364(25):2405–2416. doi:10.1056/NEJMoa1012912

Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P (2005) Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. Science 309(5740):1577–1581. doi:10.1126/science.1113329, 309/5740/1577 [pii]

Jouan L, Chatel-Chaix L, Melancon P, Rodrigue-Gervais IG, Raymond VA, Selliah S, Bilodeau M, Grandvaux N, Lamarre D (2012) Targeted impairment of innate antiviral responses in the liver of chronic hepatitis C patients. J Hepatol 56(1):70–77. doi:10.1016/j.jhep.2011.07.017, S0168-8278(11)00611-8 [pii]

Kwo PY, Lawitz EJ, McCone J, Schiff ER, Vierling JM, Pound D, Davis MN, Galati JS, Gordon SC, Ravendhran N, Rossaro L, Anderson FH, Jacobson IM, Rubin R, Koury K, Pedicone LD, Brass CA, Chaudhri E, Albrecht JK (2010) Efficacy of boceprevir, an NS3 protease inhibitor, in combination with peginterferon alfa-2b and ribavirin in treatment-naive patients with genotype 1 hepatitis C infection (SPRINT-1): an open-label, randomised, multicentre phase 2 trial. Lancet 376(9742):705–716. doi:10.1016/S0140-6736(10)60934-8, S0140-6736(10)60934-8 [pii]

Lanford RE, Bigger C, Bassett S, Klimpel G (2001) The chimpanzee model of hepatitis C virus infections. ILAR J 42(2):117–126

Lanford RE, Guerra B, Lee H, Chavez D, Brasky KM, Bigger CB (2006) Genomic response to interferon-alpha in chimpanzees: implications of rapid downregulation for hepatitis C kinetics. Hepatology 43(5):961–972. doi:10.1002/hep.21167

Laskus T, Radkowski M, Piasek A, Nowicki M, Horban A, Cianciara J, Rakela J (2000) Hepatitis C virus in lymphoid cells of patients coinfected with human immunodeficiency virus type 1:

evidence of active replication in monocytes/macrophages and lymphocytes. J Infect Dis 181(2):442–448. doi:10.1086/315283, JID991010 [pii]

Lau DT, Negash A, Chen J, Crochet N, Sinha M, Zhang Y, Guedj J, Holder S, Saito T, Lemon SM, Luxon BA, Perelson AS, Gale MJ (2012) Innate immune tolerance and the role of Kupffer cells in the differential response to interferon therapy in HCV genotype 1 patients. Gastroenterology (in press)

Li T, Chen Z, Zeng J, Zhang J, Wang W, Zhang L, Zheng X, Shuai L, Klenerman P, Allain JP, Li C (2011) Impact of host responses on control of hepatitis C virus infection in Chinese blood donors. Biochem Biophys Res Commun 415(3):503–508. doi:10.1016/j.bbrc.2011.10.102, S0006-291X(11)01927-9 [pii]

Liu HM, Gale M (2010) Hepatitis C virus evasion from RIG-I-dependent hepatic innate immunity. Gastroenterol Res Pract 2010:548390. doi:10.1155/2010/548390

Lloyd AR, Jagger E, Post JJ, Crooks LA, Rawlinson WD, Hahn YS, Ffrench RA (2007) Host and viral factors in the immunopathogenesis of primary hepatitis C virus infection. Immunol Cell Biol 85(1):24–32. doi:10.1038/sj.icb.7100010, 7100010 [pii]

Lohmann V, Korner F, Koch J, Herian U, Theilmann L, Bartenschlager R (1999) Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. Science 285 (5424):110–113. 7638 [pii]

Loo YM, Owen DM, Li K, Erickson AK, Johnson CL, Fish PM, Carney DS, Wang T, Ishida H, Yoneyama M, Fujita T, Saito T, Lee WM, Hagedorn CH, Lau DT, Weinman SA, Lemon SM, Gale M Jr (2006) Viral and therapeutic control of IFN-beta promoter stimulator 1 during hepatitis C virus infection. Proc Natl Acad Sci U S A 103(15):6001–6006. doi:10.1073/pnas.0601523103, 0601523103 [pii]

Lopez-Labrador FX, Moya A, Gonzalez-Candelas F (2008) Mapping natural polymorphisms of hepatitis C virus NS3/4A protease and antiviral resistance to inhibitors in worldwide isolates. Antivir Ther 13(4):481–494

Lutchman G, Danehower S, Song BC, Liang TJ, Hoofnagle JH, Thomson M, Ghany MG (2007) Mutation rate of the hepatitis C virus NS5B in patients undergoing treatment with ribavirin monotherapy. Gastroenterology 132(5):1757–1766. doi:10.1053/j.gastro.2007.03.035, S0016-5085(07)00559-8 [pii]

MacPherson JI, Sidders B, Wieland S, Zhong J, Targett-Adams P, Lohmann V, Backes P, Delpuech-Adams O, Chisari F, Lewis M, Parkinson T, Robertson DL (2011) An integrated transcriptomic and meta-analysis of hepatoma cells reveals factors that influence susceptibility to HCV infection. PLoS One 6(10):e25584. doi:10.1371/journal.pone.0025584, PONE-D-11-11546 [pii]

McGilvray I, Feld JJ, Chen L, Pattullo V, Guindi M, Fischer S, Borozan I, Xie G, Selzner N, Heathcote EJ, Siminovitch K (2012) Hepatic cell-type specific gene expression better predicts HCV treatment outcome than IL28B genotype. Gastroenterology 142(5):1122–1131 e1121. doi:10.1053/j.gastro.2012.01.028, S0016-5085(12)00145-X [pii]

McHutchison JG, Everson GT, Gordon SC, Jacobson IM, Sulkowski M, Kauffman R, McNair L, Alam J, Muir AJ (2009) Telaprevir with peginterferon and ribavirin for chronic HCV genotype 1 infection. N Engl J Med 360(18):1827–1838. doi:10.1056/NEJMoa0806104, 360/18/1827 [pii]

Meylan E, Curran J, Hofmann K, Moradpour D, Binder M, Bartenschlager R, Tschopp J (2005) Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus. Nature 437(7062):1167–1172. doi:10.1038/nature04193, nature04193 [pii]

Neumann AU, Lam NP, Dahari H, Gretch DR, Wiley TE, Layden TJ, Perelson AS (1998) Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy. Science 282(5386):103–107

Nishimura-Sakurai Y, Sakamoto N, Mogushi K, Nagaie S, Nakagawa M, Itsui Y, Tasaka-Fujita M, Onuki-Karakama Y, Suda G, Mishima K, Yamamoto M, Ueyama M, Funaoka Y, Watanabe T, Azuma S, Sekine-Osajima Y, Kakinuma S, Tsuchiya K, Enomoto N, Tanaka H, Watanabe M (2010) Comparison of HCV-associated gene expression and cell signaling pathways in cells with

or without HCV replicon and in replicon-cured cells. J Gastroenterol 45(5):523–536. doi:10.1007/s00535-009-0162-3

Pawlotsky JM (2006) Hepatitis C virus population dynamics during infection. Curr Top Microbiol Immunol 299:261–284

Pawlotsky JM (2009) Therapeutic implications of hepatitis C virus resistance to antiviral drugs. Therap Adv Gastroenterol 2(4):205–219. doi:10.1177/1756283X09336045

Pawlotsky JM (2011) Treatment failure and resistance with direct-acting antiviral drugs against hepatitis C virus. Hepatology 53(5):1742–1751. doi:10.1002/hep.24262

Pedersen IM, Cheng G, Wieland S, Volinia S, Croce CM, Chisari FV, David M (2007) Interferon modulation of cellular microRNAs as an antiviral mechanism. Nature 449(7164):919–922. doi:10.1038/nature06205, nature06205 [pii]

Peng X, Li Y, Walters KA, Rosenzweig ER, Lederer SL, Aicher LD, Proll S, Katze MG (2009) Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. BMC Genomics 10:373. doi:10.1186/1471-2164-10-373, 1471-2164-10-373 [pii]

Ploss A, Evans MJ, Gaysinskaya VA, Panis M, You H, de Jong YP, Rice CM (2009) Human occludin is a hepatitis C virus entry factor required for infection of mouse cells. Nature 457(7231):882–886. doi:10.1038/nature07684, nature07684 [pii]

Poordad F, McCone J, Jr., Bacon BR, Bruno S, Manns MP, Sulkowski MS, Jacobson IM, Reddy KR, Goodman ZD, Boparai N, DiNubile MJ, Sniukiene V, Brass CA, Albrecht JK, Bronowicki JP (2011) Boceprevir for untreated chronic HCV genotype 1 infection. N Engl J Med 364(13):1195–1206. doi:10.1056/NEJMoa1010494

Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH (2001) The epidemic behavior of the hepatitis C virus. Science 292(5525):2323–2325. doi:10.1126/science.1058321, 292/5525/2323 [pii]

Pybus OG, Barnes E, Taggart R, Lemey P, Markov PV, Rasachak B, Syhavong B, Phetsouvanah R, Sheridan I, Humphreys IS, Lu L, Newton PN, Klenerman P (2009) Genetic history of hepatitis C virus in East Asia. J Virol 83(2):1071–1082. doi:10.1128/JVI.01501-08, JVI.01501-08 [pii]

Rasmussen AL, Tchitchek N, Susnow NJ, Krasnoselsky AL, Diamond DL, Yeh MM, Proll SC, Korth MJ, Walters KA, Lederer S, Larson AM, Carithers RL, Benecke A, Katze MG (2012a) Early transcriptional programming links progression to hepatitis C virus-induced severe liver disease in transplant patients. Hepatology. doi:10.1002/hep.25612

Rasmussen AL, Wang IM, Shuhart MC, Proll SC, He Y, Cristescu R, Roberts C, Carter VS, Williams CM, Diamond DL, Bryan JT, Ulrich R, Korth MJ, Thomassen LV, Katze MG (2012b) Chronic immune activation is a distinguishing feature of liver and PBMC gene signatures from HCV/HIV coinfected patients and may contribute to hepatic fibrogenesis. Virology 430(1):43–52. doi:10.1016/j.virol.2012.04.011, S0042-6822(12)00195-X [pii]

Ray SC, Fanning L, Wang XH, Netski DM, Kenny-Walsh E, Thomas DL (2005) Divergent and convergent evolution after a common-source outbreak of hepatitis C virus. J Exp Med 201(11):1753–1759. doi:10.1084/jem.20050122, jem.20050122 [pii]

Romano KP, Ali A, Royer WE, Schiffer CA (2010) Drug resistance against HCV NS3/4A inhibitors is defined by the balance of substrate recognition versus inhibitor binding. Proc Natl Acad Sci U S A 107(49):20986–20991. doi:10.1073/pnas.1006370107, 1006370107 [pii]

Ruhl M, Knuschke T, Schewior K, Glavinic L, Neumann-Haefelin C, Chang DI, Klein M, Heinemann FM, Tenckhoff H, Wiese M, Horn PA, Viazov S, Spengler U, Roggendorf M, Scherbaum N, Nattermann J, Hoffmann D, Timm J (2011) CD8+ T-cell response promotes evolution of hepatitis C virus nonstructural proteins. Gastroenterology 140(7):2064–2073. doi:10.1053/j.gastro.2011.02.060, S0016-5085(11)00272-1 [pii]

Saito T, Owen DM, Jiang F, Marcotrigiano J, Gale M Jr (2008) Innate immunity induced by composition-dependent RIG-I recognition of hepatitis C virus RNA. Nature 454(7203):523–527. doi:10.1038/nature07106, nature07106 [pii]

Sarasin-Filipowicz M, Oakeley EJ, Duong FH, Christen V, Terracciano L, Filipowicz W, Heim MH (2008) Interferon signaling and treatment outcome in chronic hepatitis C. Proc Natl Acad Sci U S A 105(19):7034–7039. doi:10.1073/pnas.0707882105, 0707882105 [pii]

Seeff LB (2002) Natural history of chronic hepatitis C. Hepatology 36(5 Suppl 1):S35–46. doi:10.1053/jhep.2002.36806, S0270913902001702 [pii]

Shiffman ML, Suter F, Bacon BR, Nelson D, Harley H, Sola R, Shafran SD, Barange K, Lin A, Soman A, Zeuzem S (2007) Peginterferon alfa-2a and ribavirin for 16 or 24 weeks in HCV genotype 2 or 3. N Engl J Med 357(2):124–134. doi:10.1056/NEJMoa066403, 357/2/124 [pii]

Shimakami T, Welsch C, Yamane D, McGivern DR, Yi M, Zeuzem S, Lemon SM (2011) Protease inhibitor-resistant hepatitis C virus mutants with reduced fitness from impaired production of infectious virus. Gastroenterology 140(2):667–675. doi:10.1053/j.gastro.2010.10.056, S0016-5085(10)01596-9 [pii]

Simmonds P, Bukh J, Combet C, Deleage G, Enomoto N, Feinstone S, Halfon P, Inchauspe G, Kuiken C, Maertens G, Mizokami M, Murphy DG, Okamoto H, Pawlotsky JM, Penin F, Sablon E, Shin IT, Stuyver LJ, Thiel HJ, Viazov S, Weiner AJ, Widell A (2005) Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. Hepatology 42(4):962–973. doi:10.1002/hep.20819

Smith DB, Pathirana S, Davidson F, Lawlor E, Power J, Yap PL, Simmonds P (1997) The origin of hepatitis C virus genotypes. J Gen Virol 78(Pt 2):321–328

Smith MW, Walters KA, Korth MJ, Fitzgibbon M, Proll S, Thompson JC, Yeh MM, Shuhart MC, Furlong JC, Cox PP, Thomas DL, Phillips JD, Kushner JP, Fausto N, Carithers RL Jr, Katze MG (2006) Gene expression patterns that correlate with hepatitis C and early progression to fibrosis in liver transplant recipients. Gastroenterology 130(1):179–187. doi:10.1053/j.gastro.2005.08.015, S0016-5085(05)01636-7 [pii]

Soderholm J, Sallberg M (2006) A complete mutational fitness map of the hepatitis C virus nonstructural 3 protease: relation to recognition by cytotoxic T lymphocytes. J Infect Dis 194(12):1724–1728. doi:10.1086/509513, JID37047 [pii]

Soderholm J, Ahlen G, Kaul A, Frelin L, Alheim M, Barnfield C, Liljestrom P, Weiland O, Milich DR, Bartenschlager R, Sallberg M (2006) Relation between viral fitness and immune escape within the hepatitis C virus protease. Gut 55(2):266–274. doi:10.1136/gut.2005.072231, gut.2005.072231 [pii]

Soriano V, Poveda E, Vispo E, Labarga P, Rallon N, Barreiro P (2012) Pharmacogenetics of hepatitis C. J Antimicrob Chemother 67(3):523–529. doi:10.1093/jac/dkr506, dkr506 [pii]

Su AI, Pezacki JP, Wodicka L, Brideau AD, Supekova L, Thimme R, Wieland S, Bukh J, Purcell RH, Schultz PG, Chisari FV (2002) Genomic analysis of the host response to hepatitis C virus infection. Proc Natl Acad Sci U S A 99(24):15669–15674. doi:10.1073/pnas.202608199, 202608199 [pii]

Sumpter R Jr, Loo YM, Foy E, Li K, Yoneyama M, Fujita T, Lemon SM, Gale M Jr (2005) Regulating intracellular antiviral defense and permissiveness to hepatitis C virus RNA replication through a cellular RNA helicase. RIG-I. J Virol 79(5):2689–2699. doi:10.1128/JVI.79.5.2689-2699.2005, 79/5/2689 [pii]

Suppiah V, Moldovan M, Ahlenstiel G, Berg T, Weltman M, Abate ML, Bassendine M, Spengler U, Dore GJ, Powell E, Riordan S, Sheridan D, Smedile A, Fragomeli V, Muller T, Bahlo M, Stewart GJ, Booth DR, George J (2009) IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat Genet 41(10):1100–1104. doi:10.1038/ng.447, ng.447 [pii]

Susser S, Welsch C, Wang Y, Zettler M, Domingues FS, Karey U, Hughes E, Ralston R, Tong X, Herrmann E, Zeuzem S, Sarrazin C (2009) Characterization of resistance to the protease inhibitor boceprevir in hepatitis C virus-infected patients. Hepatology 50(6):1709–1718. doi:10.1002/hep.23192

Takahashi K, Asabe S, Wieland S, Garaigorta U, Gastaminza P, Isogawa M, Chisari FV (2010) Plasmacytoid dendritic cells sense hepatitis C virus-infected cells, produce interferon, and inhibit infection. Proc Natl Acad Sci U S A 107(16):7431–7436. doi:10.1073/pnas.1002301107, 1002301107 [pii]

Tan S-L, He Y (eds) (2011) Hepatitis C antiviral drug discovery and development. Caister Academic Press, Norfolk

Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, Sakamoto N, Nakagawa M, Korenaga M, Hino K, Hige S, Ito Y, Mita E, Tanaka E, Mochida S, Murawaki Y, Honda M,

Sakai A, Hiasa Y, Nishiguchi S, Koike A, Sakaida I, Imamura M, Ito K, Yano K, Masaki N, Sugauchi F, Izumi N, Tokunaga K, Mizokami M (2009) Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. Nat Genet 41(10):1105–1109. doi:10.1038/ng.449, ng.449 [pii]

Tester I, Smyk-Pearson S, Wang P, Wertheimer A, Yao E, Lewinsohn DM, Tavis JE, Rosen HR (2005) Immune evasion versus recovery after acute hepatitis C virus infection from a shared source. J Exp Med 201(11):1725–1731. doi:10.1084/jem.20042284, jem.20042284 [pii]

Thomas DL, Thio CL, Martin MP, Qi Y, Ge D, O'Huigin C, Kidd J, Kidd K, Khakoo SI, Alexander G, Goedert JJ, Kirk GD, Donfield SM, Rosen HR, Tobler LH, Busch MP, McHutchison JG, Goldstein DB, Carrington M (2009) Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. Nature 461(7265):798–801. doi:10.1038/nature08463, nature08463 [pii]

Timm J, Lauer GM, Kavanagh DG, Sheridan I, Kim AY, Lucas M, Pillay T, Ouchi K, Reyor LL, Schulze zur Wiesch J, Gandhi RT, Chung RT, Bhardwaj N, Klenerman P, Walker BD, Allen TM (2004) CD8 epitope escape and reversion in acute HCV infection. J Exp Med 200(12):1593–1604. doi:10.1084/jem.20041006, jem.20041006 [pii]

Uebelhoer L, Han JH, Callendret B, Mateu G, Shoukry NH, Hanson HL, Rice CM, Walker CM, Grakoui A (2008) Stable cytotoxic T cell escape mutation in hepatitis C virus is linked to maintenance of viral fitness. PLoS Pathog 4 (9):e1000143. doi:10.1371/journal.ppat.1000143

Verbinnen T, Van Marck H, Vandenbroucke I, Vijgen L, Claes M, Lin TI, Simmen K, Neyts J, Fanning G, Lenz O (2010) Tracking the evolution of multiple in vitro hepatitis C virus replicon variants under protease inhibitor selection pressure by 454 deep sequencing. J Virol 84(21):11124–11133. doi:10.1128/JVI.01217-10, JVI.01217-10 [pii]

Walker CM (2010) Adaptive immunity to the hepatitis C virus. Adv Virus Res 78:43–86. doi:10.1016/B978-0-12-385032-4.00002-1, B978-0-12-385032-4.00002-1 [pii]

Walters KA, Smith MW, Pal S, Thompson JC, Thomas MJ, Yeh MM, Thomas DL, Fitzgibbon M, Proll S, Fausto N, Gretch DR, Carithers RL, Jr, Shuhart MC, Katze MG (2006) Identification of a specific gene expression pattern associated with HCV-induced pathogenesis in HCV- and HCV/HIV-infected individuals. Virology 350(2):453–464. doi:10.1016/j.virol.2006.02.030, S0042-6822(06)00079-1 [pii]

Welsch C, Shimakami T, Hartmann C, Yang Y, Domingues FS, Lengauer T, Zeuzem S, Lemon SM (2012) Peptidomimetic escape mechanisms arise via genetic diversity in the ligand-binding site of the hepatitis C virus NS3/4A serine protease. Gastroenterology 142(3):654–663. doi:10.1053/j.gastro.2011.11.035, S0016-5085(11)01632-5 [pii]

Wieland SF, Chisari FV (2005) Stealth and cunning: hepatitis B and hepatitis C viruses. J Virol 79(15):9369–9380. doi:10.1128/JVI.79.15.9369-9380.2005, 79/15/9369 [pii]

Xue W, Pan D, Yang Y, Liu H, Yao X (2012) Molecular modeling study on the resistance mechanism of HCV NS3/4A serine protease mutants R155 K, A156 V and D168A to TMC435. Antiviral Res 93(1):126–137. doi:10.1016/j.antiviral.2011.11.007, S0166-3542(11)00508-0 [pii]

Yan W et al (2004) System-based proteomic analysis of the interferon response in human liver cells. Genome Biol 5:R54

Zhang P, Zhong L, Struble EB, Watanabe H, Kachko A, Mihalik K, Virata-Theimer ML, Alter HJ, Feinstone S, Major M (2009) Depletion of interfering antibodies in chronic hepatitis C patients and vaccinated chimpanzees reveals broad cross-genotype neutralizing activity. Proc Natl Acad Sci U S A 106(18):7537–7541. doi:10.1073/pnas.0902749106, 0902749106 [pii]

Systems Biology Approach for New Target and Biomarker Identification

I-Ming Wang, David J. Stone, David Nickle, Andrey Loboda, Oscar Puig and Christopher Roberts

Abstract The pharmaceutical industry is spending increasingly large amounts of money on the discovery and development of novel medicines, but this investment is not adequately paying off in an increased rate of newly approved drugs by the FDA. The post-genomic era has provided a wealth of novel approaches for generating large, high-dimensional genetic and transcriptomic data sets from large cohorts of preclinical species as well as normal and diseased individuals. This systems biology approach to understanding disease-related biology is revolutionizing our understanding of the cellular pathways and gene networks underlying the onset of disease, and the mechanisms of pharmacological treatments that ameliorate disease phenotypes. In this article, we review a number of approaches being used by pharmaceutical and biotechnology companies, e.g., high-throughput DNA genotyping, sequencing, and genome-wide gene expression profiling, to enable drug discovery and development through the identification of new drug targets and biomarkers of disease progression, drug pharmacodynamics, and predictive markers for selecting the patients most likely to respond to therapy.

Contents

I.-M. Wang (✉) · D. J. Stone · D. Nickle · A. Loboda · O. Puig · C. Roberts
Informatics and Analysis, Merck Research Laboratory, West Point, PA 19486, USA
e-mail: I_ming_wang@merck.com

I.-M. Wang
Merck Corporation, PO Box 100 Whitehouse Station, NJ 08889-0100, USA

Current Topics in Microbiology and Immunology (2013) 363: 169–199
DOI: 10.1007/82_2012_252

Published Online: 18 August 2012

1 Introduction

Systems biology is an interdisciplinary approach examining complex interactions among different components in a biological system. Its goal is to determine prediction rules governing a system's behavior under different conditions. The pharmaceutical industry is currently facing tremendous challenges from many directions including an imminent "patent cliff" (Harrison 2011), declining research and development (R&D) productivity, low public perception, and increasing regulatory hurdles (Kola 2008). To be able to successfully respond to these challenges, the industry has to adopt significant paradigm shifts and innovative approaches in drug R&D to cut the cost of development, to find targets with better efficacy, and to identify biomarkers which can predict response and adverse events (AEs). Many recent technological advancements including transcriptional profiling, next-generation sequencing (NGS), and other high-throughput genomics analysis platforms, such as the single nucleotide polymorphism (SNP) array (i.e. SNP chip) for genome-wide association studies (GWAS) and RNA interference (RNAi) screening, have further enabled the systems biology approach to impact future drug discovery and development. For example, applying the systems approach could help identify the mode of action and potential toxicity of compounds under development, which would allow companies to terminate unfavorable development projects early. The systems approach could also help to identify human subpopulations who may not respond to certain therapeutics, which would bring us closer to the goal of achieving personalized medicine (Trusheim et al. 2011).

Merck Research Laboratory (MRL) was among the first research institutes to perform integrated analysis of large scale genomics data sets, and to apply the

results toward practical drug development (Dai et al. 2005; Schadt et al. 2009; van't Veer et al. 2002, 2003; van de Vijver et al. 2002). We summarize herein the current status of selected areas within the systems biology arena, and include our recent efforts in biomarker and target identification.

2 Gene Signatures Used for Predicting and Understanding Diseases

Genome-wide gene expression profiling has launched a new era of understanding of human diseases at the molecular level (Keller and Attie 2010; van't Veer and Bernards 2008; van't Veer et al. 2002). Coherent patterns of gene expression observed across large cohorts of human samples provide an information rich source of data for understanding of molecular subtypes of human diseases and their drivers. Its utility was first recognized in oncology, where distinct tumor subtypes develop in the same histological environment and their differential response to therapies presents a significant hurdle for drug development (Bertos and Park 2011). Indeed, genome-wide gene expression of large cohorts of tumor samples revealed heterogeneous patterns of gene expression and multiple independent groups of coherently expressed genes or 'metagenes' (Huang et al. 2003). Some of these 'metagenes' are related to key physiological properties of the tumor, such as the rate of proliferation, presence of immune components, or adhesion. Other patterns can be traced to pathway activation status, such as the state of the RAS pathway (Loboda et al. 2010), Myc signaling (Huang et al. 2003), or the degree of epithelial-to-mesenchymal transition (Loboda et al. 2011) (see Sect. 2.2 below). These molecular patterns reveal complex tumor biology comprised of multiple independent dimensions that need to be captured for correct diagnostic and prognostic decisions.

2.1 Cancer Prognostic Signatures

In each tumor type, the most variable, clinically and biologically relevant gene expression patterns were used to define molecular tumor subtypes. For example, among breast tumors, luminal and basal types were defined, each of which was further divided into subtypes using a combination of key patterns such as proliferation and ER signaling (Sorlie et al. 2003). This molecular subtyping is clearly just the tip of an iceberg of a much more complex set of subtypes driven by different oncogenic events, and further characterized by additional physiological properties. Surprisingly, even a very high-level molecular characterization turns out to be clinically useful in prognosis of primary and metastatic tumors. It was not obvious that such early prognoses could be made, since most gene expression

profiles are generated from primary tumors, and it is not guaranteed that the biology captured at that early stage would be useful for making predictions of the future metastatic behavior of the tumor. Fortunately, the biology of tumor subtype and its behavior turn out to be to a high degree predetermined at first diagnosis of the tumor, which allows for prognostic predictions at least in some tumor types, such as breast tumors (van't Veer et al. 2002). Several diagnostic assays such as MammaPrint (Mook et al. 2007), PAM50, OncotypeDx (Koscielny 2008) have been developed for prognosis based on this principle, and hundreds of thousands of women have been tested with these diagnostic assays.

2.2 Overlap Between Alzheimer's Disease and Physiological Aging Signatures

Similar to the work in oncology, genome-wide gene expression profiling of large cohorts of brain tissue have been used to reconstruct the development of Alzheimer's disease (AD), and these profiles have been compared with the process of normal aging (Podtelezhnikov et al. 2011). Gene expression variation in a cohort of brains from normal and AD-affected individuals could be almost completely explained by a few transcriptional biomarkers that capture the top principle components of variation. These include genes statistically associated with neuronal loss, glial activation, lipid metabolism, and inflammation. Among the key patterns contributing to disease progression, the small but exceptionally tightly correlated metagene, called Inflame, contains about 250 genes upregulated with AD, including many inflammation markers, such as IL1β, IL10, IL16, IL18, multiple HLA genes, as well as markers of macrophages, such as VSIG4, SLC11A1, and apoptosis, such as CASP1/4, TNFRSF1B (p75 death receptor) (Podtelezhnikov et al. 2011).

Together, these biomarkers provide a detailed description of the aging process and its contribution to AD progression. The results of such analysis can be summarized in the form of a state transition model shown in Fig. 1. Aging starts with up-regulation of APOE and other lipid metabolic genes, signifying the transition from N0 to N1. The following upregulation of the Inflame biomarker, composed of inflammation genes, is associated with transition from N1 to N2. The brains in these states (i.e. N1, N2) were diagnosed as normal, because the subjects did not yet exhibit any cognitive impairment associated with AD. The next transition, from N2 to A1, is associated with massive disruptions in metabolic pathways, and an observed marked acceleration of aging then follows. Some brains, however, avoid transitioning to A1 and continue to age into N3. Another transition to AD state A2 can happen later. This transition may appear later than A1 in a particular brain region, and happen much earlier in some other brain regions.

The proposed model is most consistent with an age-based hypothesis of AD that postulates three fundamental steps: initial injury aggravated by age, chronic

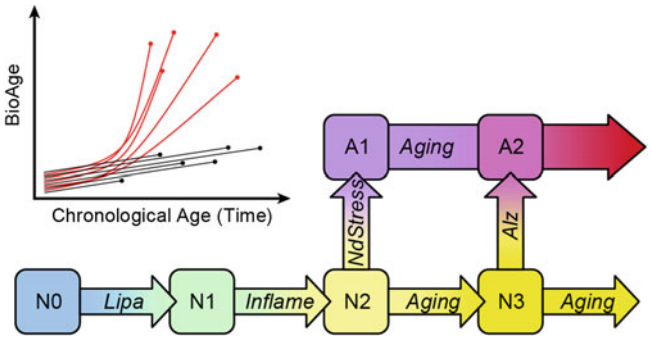


Fig. 1 Alzheimer's disease progression model. The trajectories of biological age (BioAge) changes as a function of time, reflect the relatively constant rate of aging in nondemented subjects (*black*), and acceleration of the rate of aging in AD (*red*). The *dots* represent the postmortem state of the brain captured by gene-expression profiling. The state transition model defines several broad categories for normal brains N0–N3, and for diseased states A1 and A2. The sequence of transitions and associated gene expression biomarkers are shown by *arrows*: lipid metamolism (Lipa), inflammation (Inflame), neurogenenerative stress (NdStress), and EMT signaling specific to Alzheimer's disease (Alz)

neuroinflammation, and subsequent transition of most brain cells to a new state (Herrup 2010). These key stages of the disease were independently observed and associated with transcriptional changes in our analysis of the brain transcriptome. We also identified a striking resemblance of the biological processes behind the disease progression biomarkers to an epithelial-to-mesenchymal transition (EMT) (Kalluri and Weinberg 2009). The AD processes are most similar to EMT type 2, which is dependent on inflammation-inducing injuries for initiation and propagation. Associated with tissue regeneration and organ fibrosis in kidney, lung, and liver, EMT type 2 generates mesenchymal cells that produce excessive amounts of extracellular matrix (ECM). Similarly, a transition of AD brain into a tissue enriched with mesenchymal cells produces a large amount of ECM containing beta-amyloid. This model of the disease implies that multiple independent genetic factors, as well as infections and/or injuries may accelerate consecutive transitions leading to disease. It also suggests different therapeutic strategies for early and late disease stages.

3 The Systems Biology Approach Applied to Human Genetics

We have been witnessing great advances in the generation and analysis of genetic data used to identify the molecular underpinnings of biological traits. With the publication of the first draft of the human genome in 2001, we embarked on a journey of high-throughput data acquisition. Since genetics is founded on the study of variation, it was clear then that a single human sequence would do very little for

geneticists in terms of defining the genetics of complex diseases. Lander and Schork (1994) argued strongly for association studies to unravel the genetics of complex traits in human populations. From that seminal paper, entire companies have been formed (e.g. Affymetrix and Illumina) with the sole purpose of producing SNP chips for interrogating the entire genome at more than 10^6 loci. Additionally, the International HapMap Consortium (Frazer et al. 2007) set out to catalog human variation, and has been embraced by the biomedical community in hopes that identifying variation will be useful for understanding the genetics of complex traits. In 2005, one of the first genome-wide association scans (Klein et al. 2005) was published, illustrating the power of an unbiased view of the genome. However, this study and its design quickly became outdated, and the number of patients being recruited into GWAS studies expanded to more than 10,000 individuals (McGregor et al. 2007). Simultaneously, it was recognized that variation in gene expression also plays an important role in complex traits. Most importantly, gene expression variation has been observed in almost all natural populations studied thus far (Genissel et al. 2008; Gilad et al. 2006; Oleksiak et al. 2002). The impact of phenotype has been profound with respect to gene expression variation. Much like the more obvious structural polymorphisms caused by nonsynonymous mutations and/or indels, variation in gene expression can lead to equally extraordinary changes in phenotype (Bergland et al. 2008; Gompel et al. 2005; McGregor et al. 2007; Oleksiak et al. 2002; Stern 1998).

Although studies focusing on human disease have benefited from the high-throughput revolution, in general, these advances, thus far have lead to somewhat disappointing results. Specifically, genomic studies have found that complex traits are affected by a large number of loci, with any single locus having only a minor explanatory power, leaving most of the variation unaccounted for. This pattern holds for numerous diverse and unrelated traits across many different organisms (Gilad et al. 2008).

In recent years, the integration of genetics and gene expression data to unravel complex traits has taken off, although the heritability of the genetics of gene expression has been met with some debate. To investigate this phenomenon, two groups independently used the Centre d' Etude du Polymorphisme Humain (CEPH families) database derived from transformed lymphoblasts. Using either Affymetrix short-oligonucleotide (Morley et al. 2004) (http://www.affymetrix.com/index.affx) or Agilent long-oligonucleotide (Cheung et al. 2003; Monks et al. 2004) arrays (http://www.agilent.com), it was concluded that a large proportion of gene expression varied between individuals (Cheung et al. 2003; Monks et al. 2004) and with a surprisingly low heritability. It has been suggested (Schadt, EE, personal communication) that these results were driven in part by the fact that the cell lines from which the gene expression data was acquired were "long-term transformed", and, therefore, far from the state they might have been if gene expression had been determined directly without transformation. Indeed, it does appear that the heritability of gene expression in these studies with the CEPH transformed cell lines was surprisingly low (Cheung et al. 2003; Monks et al. 2004; Morley et al. 2004). Relatively, high heritability of gene expression has

otherwise been observed in many different taxonomic groups (Brem et al. 2002; Schadt et al. 2003), and most importantly, in humans (Emilsson et al. 2008). Emilsson et al. (2008) have shown that heritability in gene expression can be captured simultaneously in different tissue compartments (blood and adipose), even if different compartments have compartment-specific gene expression patterns. Although the initial studies of gene expression from lymphoblasts were not informative in terms of measuring the heritability of expression patterns, we now know it is possible. For example, Bill Cookson's lab at the National Heart and Lung Institute, Imperial College, London used lymphoblast gene expression with GWAS to identify "genes" associated with childhood asthma (Dixon et al. 2007; Moffatt et al. 2007). One of the major differences between the lymphoblasts used in this study and the study using the CEPH families is the age of the transformed cell. The CEPH cell lines were relatively old, and thus affected by the vagaries of cell line maintenance and artificial selection.

In view of recent studies, enhanced understanding of the genetics of gene expression does seem to facilitate the dissection of complex traits. For example, the identification of expression quantitative trait loci (eQTLs) (eSNPs) allows for the orthogonal mapping of SNPs that are significant in genome-wide association studies allowing one to point toward the mechanisms of disease (Moffatt et al. 2007; Schadt et al. 2008) (see Sect. 3.2 below). SNPs associated with gene expression in lymphoblasts can also be used to determine sensitivity to chemotherapeutic agents (Huang et al. 2007) given the caveats of above. It has become clear that to derive meaningful and actionable biology from these high-throughput studies, we will need to map together as many orthogonal data sets as possible to squelch the false positives.

3.1 Value of Traditional Genetics

Within the pharmaceutical industry, the field of genetics is primarily utilized in two areas: drug target identification and pharmacogenetics. While the interest in pharmacogenetics has increased in recent years, and can only be expected to grow, the pharmaceutical industry may be slower to adopt systems biology methodologies in this realm, due to the characteristics which make a pharmacogenetic marker useful. Pharmacogenetics markers, or the use of genetic variants to predict individual response to drugs, usually focus on either compound efficacy or AEs predictions. In the case of genetic predictors of compound efficacy, the identified variants need to be both relatively common in the general population and additionally have a "large" effect size; those not fulfilling both criteria are considered unlikely to be used in common clinical practice. For example, even if a genetic variant perfectly predicts "non-response" in patients, if it is only carried in 1 % of the population, physicians are unlikely to use it; similarly a marker that is common, but only increases by 20 % the chance of classifying a patient as being a "responder", would be unlikely to gain wide acceptance. Complex systems

biology-based pharmacogenetic markers with large numbers of SNPs and patients falling along continuums in terms of response would be extremely challenging to utilize in clinical practice, leaving the focus at this time on identifying simple (1–2 genetic variants or SNPs) markers which explain a large portion of variance in selected responses.

In the area of drug target identification, the field of genetics is far more likely to be utilized in the near future. From a recent analysis of the roughly 500 human genes that have been successfully utilized as drug targets, approximately 50 % have been shown to be linked to human diseases (Wang et al. 2012b). This is substantially higher than the proportion of human genes in the genome as a whole that have been linked to disease (roughly 11 %). While it would be difficult to definitively prove the reason for this association, it is likely that in many cases the gene in question (being causal for disease) is unequivocally in the correct pathway for phenotypic modification to ameliorate the disease in question; the remaining factors determining its utility as a drug target would be tied to ease of drugability. In cases where a single mutation in a gene causes an extreme phenotype or disease, selection of the target is more or less straightforward (again assuming availability of druggable domains in the protein). However, in complex diseases which have a large number of genetic risk factors with small odds ratios (ORs), the choice of which gene/s to pursue as targets is not clear. Therefore, methodologies (such as systems genetics) which can illuminate critical pathways involved in disease etiology may also prove to be useful for drug target identification in complex diseases.

While GWAS have been successful in identifying loci associated with well defined diseases, most of the "hits" have had small ORs, frequently less than 1.5 (Hindorff et al. 2009).Considering the small size of the individual effects, the number of associations has been smaller than expected, leading to the question of "missing heritability" in many diseases, where the observed heritability based on GWAS results is significantly lower than the predicted heritability (Manolio et al. 2009). While the reason for this discrepancy is not clear at this time, it seems unlikely that it would be caused by additional common loci which have not yet been discovered. Many GWAS consortia have genotyped patients and controls in the tens of thousands (Estrada et al. 2012; Saxena et al. 2012) and several possible (and not necessarily contradictory/exclusive) explanations have been suggested. For example, interactions between genes/variants could explain a large proportion of the missing heritability if the genes are members of rate-limiting pathways (where a trait depends on multiple pathways) each of which may be a strictly additive trait, dependent upon multiple genes (Zuk et al. 2012). Correct gene assignment into biological pathways depends upon the associated SNPs identified in GWAS being assigned to the correct gene. While this explanation appears easy on the surface, the fact that GWAS arrays have been designed to tag haplotypes and *not* individual genes is not always appreciated, and has undoubtedly led to incorrect assumptions concerning which genes have been implicated by certain GWAS. Upon completion of a GWAS, SNPs identified as "hits" are usually assigned to the nearest gene, without consideration for the fact that the SNP in

question may be in tight linkage with another SNP several to hundreds of kb away (Christoforou et al. 2012). A clear example of the issue can be seen in recent AD GWAS. The APOE ε4 allele has unequivocally been linked to AD in over 100 studies (Bertram et al. 2007); however in large GWAS the most significant SNPs (while in tight linkage with the APOE ε4 allele) have been located *physically* closer to APOC1 (Naj et al. 2011), TOMM40, and PVRL2 (Harold et al. 2009). Although imputation lessens this effect to some extent, the correct assignment of SNPs to genes remains an issue if linkage is not taken into account during GWAS analysis (Christoforou et al. 2012).

3.2 The Genetics of Gene Expression

The systems biology approach developed by Eric Schadt (Emilsson et al. 2008; Schadt et al. 2003, 2008) and colleagues in the Genetics Department within Merck MRL focuses on coalescing data from all levels along the central dogma chain (DNA–RNA–Protein–Metabolite) in an attempt to pull the true positive SNPs from the plethora of false positives. Based on this approach, one place to start in determining SNPs of interest is the intersection of SNPs associated with traits and SNPs associated with gene expression: we term the latter type expression SNPs (eSNPs), and they form the foundation of this integrative genomics approach to understanding the molecular underpinnings of complex traits.

There are two commonly used methods to identify genes that associate with disease. The first method relies on the measurement of gene expression such that genes that associate with trait/phenotype/disease are found to be either up- or down-regulated, with respect to disease state. Although these regulated genes are interesting with respect to trait/phenotype/disease, they do not provide us with an understanding of the causal relationship between disease and gene. That is to say, a proportion of those correlated genes will be reactive to the trait, but not causal. The second method involves the use of genome-wide association studies (GWAS). These studies identify, in an unbiased fashion, loci that associate with disease. The problem that typically arises in these studies is that linkage disequilibrium around the marker can include many genes. The question naturally arises: which gene plays a role in the disease or trait of interest? One method that Merck pioneered and currently uses in an effort to create rank order lists of genes according to biological relevance combines gene expression data with SNP data, assuming that if variable expression of a gene correlates with a disease-associated SNP, then the gene is more likely to be driving the variability of the disease trait. However, we should remain cautious even when a disease SNP correlates with an eSNP. Conversely, correlation between a gene and trait may indicate that the gene is worthy of further investigation and investing of more resources to secure a validation even when no association between the gene and a SNP is found.

Within MRL, the eSNPs discovery program is fairly straightforward. A GWAS is performed with gene expression from a genome-wide expression panel

Table 1 The number of eSNPs discovered in MRL on a tissue-by-tissue basis

Tissue	Tissue Specific	Shared	Total (%)
Blood	410	1940	21
Brain	1,282	3,584	36
Adipose	3,545	9,163	39
Liver	3,602	8,827	41
Total	8,839	2,3514	38

Many eSNPs discovered appear to be tissue specific while others are shared between two or more tissues. These numbers are an absolute minimum—thus the actual number may be larger

functioning as the trait. That is, *marker-by-marker* association tests are conducted with gene expression treated like a standard outcome (e.g. *height*). We typically genotype 500–1,000 K SNPs from a population of ~1,000 individuals and then query the data set to find genes whose expression correlates with the loci being genotyped. For example, with 1,000 K SNPs and a 20,000 gene expression pattern, the marker-by-marker analyses will result in 20 billion linear models (not including permutations required for false discovery rate (FDR) estimates); not an easy undertaking with typical computational resources. However, Merck has taken a brute-force approach, using massively parallel computing, which has involved the systematic compilation of lists of eSNPs from many tissues. As of today, Merck has identified eSNPS from blood, brain, liver, and adipose (Table 1) and is actively working on lung and HCV infected livers.

3.3 Removing the Noise with Orthogonal Data: An Example

The Harold et al. (2009) laboratory kindly provided all SNP and *p* value pairs for AD from their GWAS performed on 16,000 individuals for Merck to conduct an additional analysis. With such a large number of individuals participating in this study, the researchers still only found modest odds ratios for any significant loci, even though there has been a very high estimate of heritability of AD (Gatz et al. 2006). Using previously published or publically available data, we set out to capture the molecular pathways that drive the disease. First, we translated all significant SNPs from Harold et al. into genes by asking which SNPs from the GWAS are members of an eQTL complex, and then we filtered these SNPs further by asking which subset of those SNPs also occurs in the brain as an eQTL. For the next step/layer, we expanded the corresponding gene list by intersecting our GWAS brain eQTL list with published interaction data to generate a network where we deem the edges of the network to be high quality (Proteome, NetPro, BIND, Reactome, KEGG, BioGRID, IntAct, Ingenuity, HPRD, DIP & MINT) (Fig. 2), giving rise to a richer and more complex view of AD. The process allows for highlighting the key genes capturing the complex molecular underpinnings of Alzheimer's disease. Figure 3 depicts a network view of AD with every single

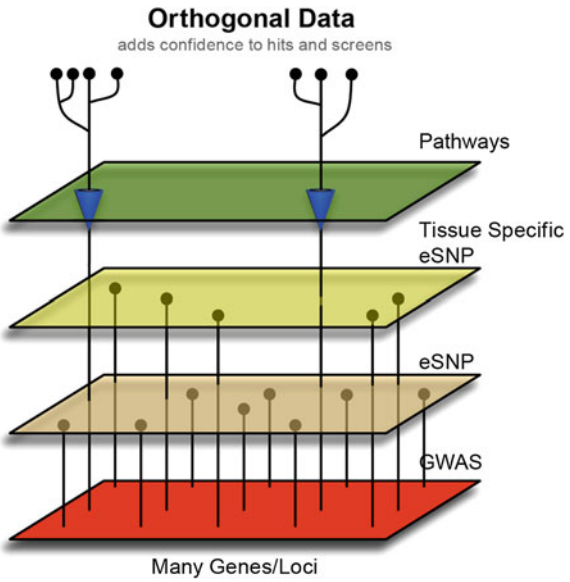


Fig. 2 The workflow to arrive at the network depicted in Fig. 3

gene being part of an eSNP complex in which the SNPs themselves were captured in the Harold (Harold et al. 2009) GWAS at a p value less than 1.0E − 03. The large red nodes are frequently discussed in the context of AD and they are part of an eSNPs complex in the human brain. One hypothesis for the surprisingly low odds ratio in the findings from Harold et al. is that there may be a large amount of epispastic interactions among the loci in the human genome, and the network in Fig. 3 captures it. The fact that greater than 90 of the SNPs that fall into eQTL complex form a single coherent network suggests that we are capturing key genetic players in Alzheimer's disease.

3.4 NGS: Technologies

The use of NGS for whole genome sequencing (WGS) and whole exome sequencing (WES) in large diseased cohorts should circumvent some of the issues previously mentioned, and will hopefully point to genes that will be the next cohort of targets to be followed for drug development. Genes containing frame shifts, large in/dels, missense, or nonsense mutations that are either associated with extreme phenotypes or are shown to be causal for disease development will be obvious candidates; however, data from WGS and WES studies may also enable pathway analysis for the understanding of complex diseases. Data from the 1,000 Genomes Project suggest that individuals carry roughly 250–300 loss-of-function variants in their genome (2010; Buchanan et al. 2012), although subsequent analysis has lowered this estimate to roughly 100 loss-of-function variants per

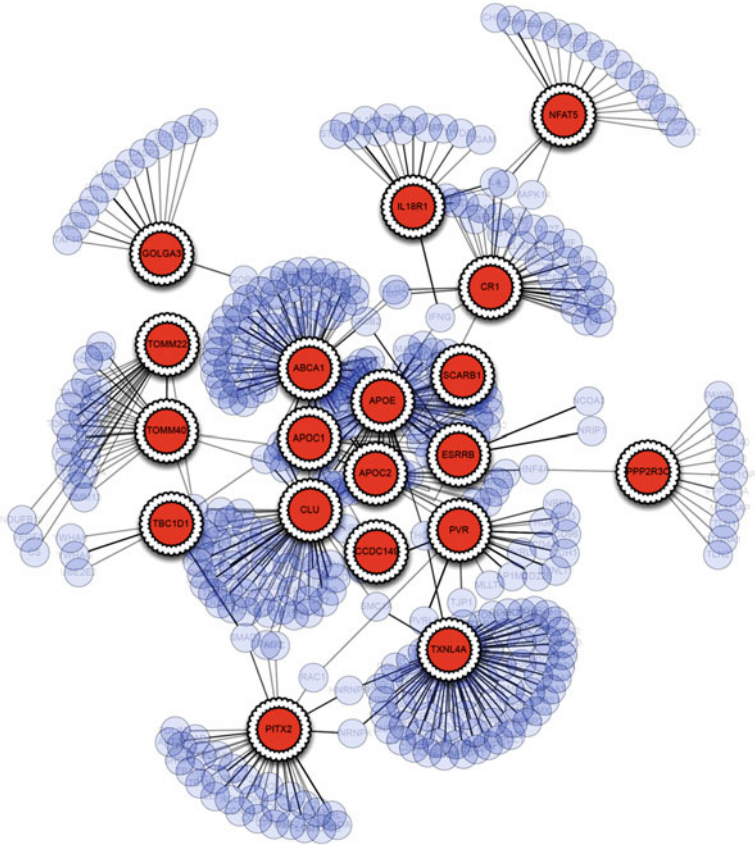


Fig. 3 The network of genes that come up in Harold et al. (2009) forming a coherent network, where the network is defined by orthogonal data and the genes that come from those SNPs that are members of a previously defined eSNP complex. *Red* nodes are genes that are also part of an eSNP complex from the brain

individual (MacArthur et al. 2012). Analysis of large cohorts via WGS or WES combined with a focus on mutations affecting protein coding will bypass the question of linkage, directly implicating genes with diseases. In the case of complex diseases, this should enable pathway analysis beyond that possible with GWAS results alone.

An example of this methodology has been applied in autism. Autism spectrum disorders have a strong genetic component, but for the majority of cases, the genetic cause is unknown. Recently, WES was performed in 209 families with autism (for 677 total exomes) to identify autism candidate genes; 126 truncating or severe missense mutations were detected (O'Roak et al. 2012). This number of candidate drug targets would be impossible to follow up on individually in a meaningful way. Much like the AD example above, these authors mapped their

results onto a protein–protein interaction map showing roughly 40 % of the hits mapped to a highly interconnected network largely implicating the β-catenin signaling pathway (a developmental regulator involved in neuronal development).

4 Identification of an "Inflammatome" for Drug Target and Biomarker Discovery

Recent surveys have indicated that lack of efficacy, and toxicity, are two of the major causes of failure in drug development (Kola 2008; Kola and Hazuda 2005; Kola and Landis 2004). Since efficacy is usually established using pre-clinical models, it is critical to identify and validate drug targets based on reliable animal models and connect the resulting data to efficacy proof of concept (POC) in humans early in the development process. In addition, robust biomarkers capable of reporting such efficacy need to be in place before engaging large clinical trials.

Disease target gene identification is a complicated process without a standard protocol. Efforts have been made to identify disease-specific targets directly from human patient populations based on high-throughput genetic associations (Roses et al. 2005) and many computational methodologies employing pathway- or network-related approaches to mine publically available databases have been proposed and executed (Dezso et al. 2009; Kim et al. 2011b; Ortutay and Vihinen 2009; Roses et al. 2005; Tiffin et al. 2008). However, most pharmaceutical companies still rely on a literature-based approach to find new targets for their drug pipelines. We describe in this section, our effort to identify a reference gene set for future drug target and disease-specific biomarker consideration by an integrated analysis of comprehensive lists of both animal and human genomics data sets available to us. This approach is unique in that it is the first systematic investigation of multiple tissues derived from multiple disease models combining gene expression profiling data with statistically causal genetic networks across rodents and humans, which could have a higher translational value in delivering better targets and disease biomarkers.

4.1 Inflammatome: A Representative Inflammatory Gene Signature

It is well-established that most common diseases not previously thought to be associated with inflammation, such as atherosclerosis (Weber and Noels 2011), cancer (Walczak 2011), diabetes (Hess and Grant 2011; Wen et al. 2012), obesity (Stienstra et al. 2012), osteoarthritis (Kapoor et al. 2011), sarcopenia (Peake et al. 2010), and stroke (Iadecola and Anrather 2011), all have a significant component of inflammation. To ensure our effort would result in a broad coverage of major

Table 2 Rodent disease models included in the inflammatome analysis

	Tissue	FDR[a] (%)	# Genes	Up-regulated	Down-regulated	Overlap with inflammatome	
						Up-regulated	Down-regulated
Inflammatome (mouse)[a]			2,505	1,511	994	1,511	994
Inflammatome (rat)[b]			2,486	1,397	1,089	1,397	1,089
OVA	Lung	1	3,989	2,037	1,952	790	354
IL-1b Tg	Lung	15	3,681	1,851	1,830	732	256
TGFb-Tg	Lung	2	3,483	1,799	1,684	505	265
ApoE KO HFD	Aorta	12	3,995	1,983	2,012	744	263
ob/ob	Adipose	10	3,314	1,696	1,618	358	218
db/db	Adipose	15	3,626	1,514	2,112	454	287
db/db	Islet	1	3,861	1,983	1,878	400	250
CGN inflammatory pain	Skin	10	4,273	1,853	2,420	538	318
Chung neuro pain	DRG	1	4,353	1,844	2,509	579	404
Stroke	Brain	5	4,240	1,990	2,250	598	347
Sarcopenia	Muscle	5	3,790	2,053	1,737	379	235
LPS	Liver	2	3,717	1,790	1,927	467	203

[a] False discovery rate (FDR) was calculated for each individual data set
[b] The inflammatome signature (2,505 mouse genes) was identified by a two-way ANOVA approach (p <=1.0E − 9, Benjamini-Hochberg corrected)
[c] The 2,505 mouse inflammatome genes map to 2,486 rat genes

Table 3 Additional published disease gene signatures significantly overlapping with the inflammatome

Species	Disease/tissue	Common genes	p Value	Reference
Mouse	Arthritis/synovium	1,339	2.3E–180	(Geurts et al. 2009)
	Cancer/bladder	1,180	2.0E–148	(Kim et al. 2011a)
	Glomerulonephritis/kidney	606	1.3E–145	GSE969
	Cancer/prostate	582	9.0E–144	(Bacac et al. 2006)
	Colitis/colon	946	2.4E–131	(Schmidt et al. 2010)
	Cancer/breast	736	4.8E–128	(Liu et al. 2009)
	S. aureus infection/blood	953	2.3E–76	GSE19668
Human	Cancer/brain	1,341	1.2E–86	(de Tayrac et al. 2009)
	Psoriasis/skin	1,347	1.7E–75	(Yao et al. 2008)
	Colitis/colon	1,486	3.4E–72	(Arijs et al. 2009)
	Cancer/breast	867	5.2E–71	(Pedraza et al. 2010)
	Cancer/ovary	1,312	3.3E–67	GSE12172
	Cancer/kidney	1,180	7.5E–63	GSE14762
	HIV infection/lymph node	785	7.4E–59	(Li et al. 2009)

disease areas, we started our gene expression analysis on rodent inflammatory disease models of 12 data sets derived from 11 models including three respiratory diseases (asthma, emphysema, and pulmonary fibrosis), two metabolic diseases (obesity and diabetes), two pain-related diseases (CGN-induced inflammation pain and Chung neuropathic pain), atherosclerosis, LPS-treated liver injury, age-related sarcopenia, and stroke. As listed in Table 2, only the most disease-relevant tissue from each model was profiled with a total of nine tissues on the list. Using the two-way ANOVA approach, we selected a representative signature of 2,505 genes in mouse (which map to 2,483 rat genes) across 12 disease model-tissue combinations as well as disease-specific signatures. Among the 2,505 genes, there are 1,026 genes consistently up- or down-regulated in at least 10 data sets. An annotation of this representative signature indicated that it is highly enriched in macrophage genes and genes associated with inflammation and immune response, thus was termed the "inflammatome" (Wang et al. 2012c)

As expected, the inflammatome signature significantly overlaps with disease signatures derived from each individual model included in the analysis (Table 2). Furthermore, a comprehensive comparison with data available in the public domain showed that it is also significantly overlapping with many disease signatures from both mouse [e.g. arthritis (synovium) (Geurts et al. 2009), glomerulonephritis (kidney), colitis (colon) (Schmidt et al. 2010), bacteria-infected blood, and several types of cancer (bladder, breast, and prostate) (Bacac et al. 2006; Kim et al. 2011a; Liu et al. 2009)] and human [e.g. psoriasis (skin) (Yao et al. 2008), colitis (colon) (Arijs et al. 2009), HIV-infected lymph node (Li et al. 2009), and several types of cancer (brain, breast, kidney, and ovary) (de Tayrac et al. 2009; Pedraza et al. 2010)] (Table 3) which broadens the association of the inflammatome to additional major disease areas such as cancer, autoimmune, and infectious diseases.

A comparison among the inflammatome, a list of drug target genes (currently on market and under investigation) according to GeneGo (http://www.genego.com/), and a list of genes based on cataloged GWAS studies (www.genome.gov/gwastudies, accessed on May 8, 2012) was conducted to assess the potential usefulness of the inflammatome. The results showed that the inflammatome includes 178 out of 803 drug target genes and 545 out of 3,886 GWAS candidate genes; both of these overlaps are significant. It is interesting to point out that two genes, Ppara and Prkaa2 (Ampk) with agonists on the market or under development, are down-regulated in all 12 inflammatome data sets; whereas two other genes, Syk and Jak2 (Mocsai et al. 2010; Quintas-Cardama et al. 2011), both with inhibitors under development for multiple disease areas, are up-regulated in at least 11 data sets. This preliminary analysis provided confidence that the inflammatome signature is highly enriched in current drug targets and GWAS genes for common diseases, and could be potentially utilized as a gene set for selecting new drug targets.

4.2 Macrophage-Enriched Metabolic Network Module

Gene expression alone can not distinguish whether variations in mRNA are causal or reactive to the associated phenotype; whereas analysis including genotype, along with gene expression, and other complex trait data in segregated populations could allow inferring causal relationship and construction of genetic networks which have more predictive power. MRL was among the first research institutes to apply this type of integrated systems biology approach both in rodents (Chen et al. 2008b; Derry et al. 2010; Mehrabian et al. 2005; Schadt et al. 2003, 2005; Yang et al. 2009; Zhu et al. 2004) and in humans (Dobrin et al. 2011; Emilsson et al. 2008; Schadt et al. 2008; Zhong et al. 2010a, b), to identify gene networks perturbed by susceptible loci which then lead to disease. One interesting module identified this way was called the macrophage-enriched metabolic network (MEMN) module (Chen et al. 2008b; Emilsson et al. 2008). The MEMN module is composed of ~1,200 genes in mice and ~2,500 genes in humans, with a highly significant overlap. Expression of these genes is coregulated in liver and adipose, and many MEMN genes have a causal relationship with disease traits associated with metabolic syndrome. The presence of multiple inflammatory genes and macrophage activation pathways suggest that the MEMN module plays a role in macrophage infiltration of liver and adipose tissue. Coexpressed modules such as the MEMN are found to be enriched for defined biological pathways with genes associated with disease traits, and for genes linked to common genetic loci (Lum et al. 2006).

It was subsequently demonstrated that when nine MEMN module genes were individually perturbed by a transgenic or knockout approach, eight of them exhibited a phenotype of abdominal obesity (Yang et al. 2009). Atherosclerosis plaque formation and rupture are directly related to macrophage dysfunction and it was shown that an atherosclerotic plaque signature which can separate inflamed from noninflamed plaque shares common features with the MEMN module (Puig et al. 2011). Recently, Min et al. (Min et al. 2012) performed coexpression network analysis for adipose and blood in humans with metabolic syndrome found a significant overlap between the metabolic syndrome networks and the MEMN module. These results confirm the central role of the MEMN module in metabolic disease, and provide examples of the value of networks in dissecting disease complexity. It is, perhaps, not a surprise that more than 30 and 20 % of inflammatome genes are in common with human and mouse MEMN module genes, respectively.

4.3 Inflammatome Genes are Enriched in Multiple Tissue Gene Networks in Both Mouse and Human

An effort was made to further explore how inflammatome genes perform in other tissue-derived networks by first looking into a Bayesian network (BN) built from

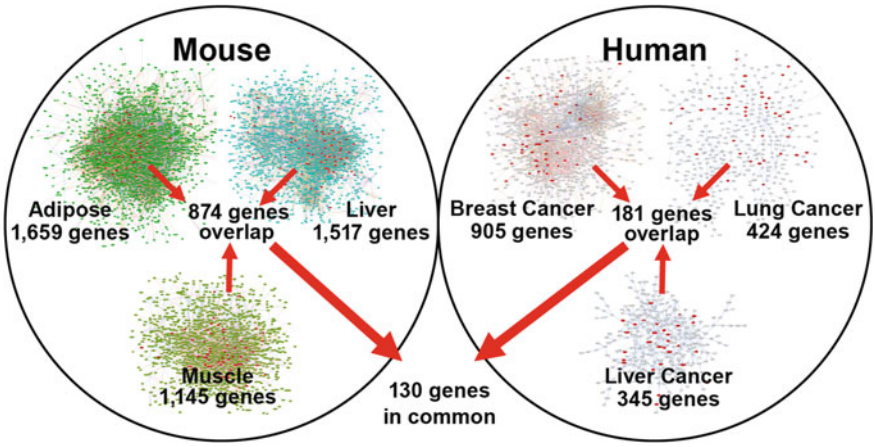


Fig. 4 Inflammatome genes are highly enriched in multiple mouse and human tissue or disease gene networks. Among ~2,500 inflammatome genes, 130 common genes were overlapping among the six Bayesian networks shown above

adipose tissue of a B6 × C3H mouse F2 cross (Cervino et al. 2005). It has been shown that Bayesian networks can be used to extract complex information (e.g. gene expression, genotype, and disease-related traits) from noisy data to derive causal relationships among genes of interest (Zhu et al. 2004). Of the ~2,500 inflammatome genes, 854 are present in the B6 × C3H adipose BN and 406 genes are directly connected in a subnetwork. When the analysis was expanded to an adipose network built from 12 independent mouse F2 crosses, more than 65 % of the inflammatome genes (1,659) were found to be present and directly connected; many supported by data from more than one F2 cross. Similar analyses were performed in liver and muscle networks reconstructed from 12 and 8 F2 crosses with 1,517 and 1,145 inflammatome genes present in the resulting BN, respectively. A comparison among all three BNs identified 874 genes in common (Fig. 4), suggesting strong causal relationships among members of this gene set across multiple tissues.

Similar integrated network analyses were subsequently conducted using large data sets acquired from various human cohorts, with 181 common inflammatome genes present in one analysis including three human cancer BNs from breast, liver, and lung (Dai et al. 2005; Lamb et al. 2011) (Fig. 4), of which 130 are overlapping with the 874 common mouse network genes, indicating a translational value of this gene signature.

4.4 Potential Applications of the Inflammatome Signature

The inflammatome signature, therefore, represents a list of disease-associated genes with many members directly connected in causal genetic networks reconstructed from multiple tissues in both human and mouse. In a sense, it extends the coverage

of MEMN (i.e. adipose and liver in the metabolic syndrome disease) to multiple diseases including many affected tissue types. Since inflammatome signature overlaps significantly with gene signatures derived from additional disease and tissue combinations, such as cancer (bladder, brain, breast, kidney, ovary, and prostate) and infectious diseases (*Staphylococcus aureus*-infected blood and HIV-infected lymph node), its importance could go beyond diseases not covered by the contributing 11 models. Further investigation and validation of key driver genes [described in details in (Zhu et al. 2008)] from this list could help in identifying targets, such as Syk and Jak2 (Mocsai et al. 2010; Quintas-Cardama and Verstovsek 2011), which can be used for development of therapeutics in multiple disease areas.

In addition to identifying a representative inflammation-related gene set, the approach used for identification of the inflammatome could be modified to derive disease-specific genes for biomarker discovery and to study distinct disease-specific pathways and mechanisms. For example, when combining a cartilage data set obtained from human osteoarthritis (OA) patients with the seven mouse data sets used in the inflammatome analysis, we were able to identify asporin (ASPN) as a potential OA-specific disease marker. Expression of asporin is highly regulated in chondrocytes (Duval et al. 2011), and an aspartic acid repeat polymorphism in asporin was found to inhibit chondrogenesis and increase susceptibility to osteoarthritis in multiple Asian populations (Kizawa et al. 2005; Shi et al. 2007; Song et al. 2008).

5 Blood Gene Profiling as a Powerful Tool for Clinical Biomarker Identification

Genome-wide transcriptomic analysis of disease can identify disease- and treatment-specific gene signatures which could then be translated into biomarkers for use in clinical practice (Allantaz et al. 2007; Deng et al. 2006; Scherer et al. 2003; van't Veer and Bernards 2008). Diseased samples from target tissues, however, are usually difficult to acquire and it is even more challenging to obtain control samples from healthy donors. For the pharmaceutical industry, this limitation is particularly daunting when conducting large-scale clinical trials. The high cost and low throughput of direct tissue biopsies simply makes transcriptomic analysis inaccessible and impractical in most late phase clinical trials, and surrogate samples, such as peripheral blood, need to be in place when appropriate, for biomarker research and development.

5.1 Enabling Technological Advancements for Blood mRNA Profiling

Profiling blood samples is not without issues. Blood mRNAs tend to be degraded and some transcripts are induced during sample preparation. These sample collection and processing issues were mostly resolved by the commercialization

of two reagents, PAXgene™ (Rainen et al. 2002) and Tempus™ (Prezeau et al. 2006). A direct comparison of the two systems had been reported which indicated that the method of blood collection and RNA purification could impact gene expression profiles (Asare et al. 2008). The other problem associated with whole blood transcriptomic studies is significant profiling artifacts due to the over-abundance of globin mRNAs (Tian et al. 2009). Most earlier blood profiling studies employed peripheral blood mononuclear cells (PBMC) to circumvent the globin mRNA issue (Burczynski and Dorner 2006; Mohr and Liew 2007). However, preparing PBMC is still tedious enough to prevent that approach from being adopted during large clinical trials; in addition, the preparation does not include potentially critical disease-related blood cell types such as neutrophils and eosinophils. At least, four RNA preparation and labeling methods which remove or block globin mRNAs during the microarray assay have been developed and utilized by investigators to effectively mitigate the negative impact of excessive globin transcripts (Parrish et al. 2010; Vartanian et al. 2009). These technological advancements have greatly improved the ability to reliably identify mRNA-based biomarkers from whole blood in a feasible way during clinical trials (Chaussabel et al. 2010; Julia et al. 2009; Marshall et al. 2010; Mendrick 2011; Pankla et al. 2009; Pascual et al. 2010; Quartier et al. 2011; Tattermusch et al. 2012).

5.2 Baseline Blood Profiling as a Reference

To use blood profiling as a tool for disease or treatment biomarker discovery, it is important to first establish an understanding of variation in gene expression patterns among healthy individuals. To achieve this goal, blood samples from 75 normal volunteers were surveyed by using cDNA microarrays and the main variation in gene expression was found to be associated with relative proportions of specific blood cell subsets. Other contributing factors identified in the study included age, gender, and the time of day when samples were collected (Whitney et al. 2003). In the same study, it was found that immunoglobulin (Ig) gene expression was negatively correlated to donor age, and the expression of a subset of IFN-inducible genes was highly variable among donors. A subsequent blood profiling study of 15 normal individuals with samples collected at multiple time points also showed high variability of IFN-inducible genes and those expressing higher baseline levels had lower response to IFN in vitro (Radich et al. 2004). Recent gene expression and pharmacogenomic studies demonstrated that genes associated with IFN pathways could determine susceptibility to disease or pathogen and response to certain treatments (Assassi et al. 2010; Everitt et al. 2012; Hambleton et al. 2011; Reif et al. 2008; van Baarsen et al. 2008; Zaas et al. 2009).

5.3 Analyzing Blood Profiling Data

Significantly modulated signatures were usually selected based on the standard statistical methods such as *t* test, and analysis of variance (ANOVA) for gene expression profiling data analysis. In some cases, gene signatures associated with clinical endpoints were identified by correlation analysis and a statistical method, Significance Analysis of Microarrays (SAM), based on an adjustable FDR, which has been adapted specifically for genome-wide transcriptomic analysis (Tusher et al. 2001). More sophisticated analyses such as the 'metagene' approach as mentioned earlier, were developed to connect microarray data with biological annotations or clinical phenotypes (Huang et al. 2003) for hypothesis generation; and another gene expression 'deconvolution' method was used to precisely measure the proportions of immune cell types in blood (Abbas et al. 2005, 2009). In a recent study, Zaas et al. (2009) performed a sparse latent factor regression analysis on human peripheral blood gene expression from patients infected with closely related rhinovirus, respiratory syncytial virus, and influenza A and derived a classifier gene set which could distinguish individuals with symptomatic acute respiratory infections (ARIs) from uninfected individuals, and viral from bacterial ARIs with >95 and 93 % accuracy, respectively.

A module-based algorithm developed specifically for blood transcriptome analysis was reported by Chaussabel et al. who employed gene expression profiles from 241 PBMC patient samples with eight different diseases. Twenty-eight co-expressed gene modules were identified and genes within the majority of modules were associated with a particular cell type, biological pathway, or process. The algorithm uses a color intensity-scoring system based on the percentage of probe sets within the module with significant *p* values (Chaussabel et al. 2008). More recently, the same team updated their modular analysis using 410 whole-blood samples from nine disease data sets and came up with 260 modules which could potentially provide more detailed annotations for blood profiling data (Banchereau et al. 2012). In addition, a score dubbed 'Molecular Distance to Health' (MDTH) that measures genome-wide expression perturbation in patients in comparison to healthy individuals was constructed in combination with the revised modules to facilitate the correlation with clinical parameters.

5.4 Integrated Blood Gene Module and Metagene Approach

We integrated the experience learned (Radich et al. 2004; Whitney et al. 2003) and analytical methodologies developed (Abbas et al. 2005, 2009; Chaussabel et al. 2008; Huang et al. 2003) from previous studies with our in-house proprietary data sets and operations to perform blood-related microarray data analysis. For example, instead of the original reported color intensity-scoring system (Chaussabel et al. 2008), we developed a module-scoring algorithm which took the average of all

gene expression levels within the module and represented it as one number to facilitate correlation analysis. Additional 'metagenes' associated with aging (Hong et al. 2008) and other clinical endpoints based on published as well as proprietary data sets were constructed to enrich our capacity for data interpretation and hypothesis generation in several infectious disease and vaccine preclinical and clinical studies.

Most of the reported vaccine-related blood profiling studies focused on identifying genes which predict vaccine efficacy (Bucasas et al. 2011; Gaucher et al. 2008; Nakaya et al. 2011; Palermo et al. 2011; Querec et al. 2009; Vahey et al. 2010). Expression levels of two genes, eukaryotic translation initiation factor 2 α kinase 4 (EIF2AK4) and solute carrier family 2, member 6 (SLC2A6) were associated with antibody titers and antigen-specific CD8 + T cell responses in yellow fever YF-17D vaccine trials (Querec et al. 2009), whereas expression levels of two other genes, TNF receptor superfamily, receptor 17 (TNFRSF17) and CD38, were able to predict antibody titers after immunization with trivalent influenza vaccine (TIV) or yellow fever vaccine (YF-17D) (Nakaya et al. 2011; Querec et al. 2009). Furthermore, it was shown that expression of the CAMK4 kinase was negatively correlated with later antibody titers, and vaccination of TIV in Camk4-deficient mice-induced higher antigen-specific antibody titers than in wild type mice (Nakaya et al. 2011).

The initial focus of our NHP vaccine study was to identify blood gene expression biomarkers associated with the AEs including both systemic (e.g. myalgia, headache, fever, and fatigue) and local (e.g. pain, redness, and swelling) AEs. We ranked several marketed vaccines including Adacel, Menactra, Havrix, Prevnar, and RabAvert along with Merck's experimental flu vaccine V512 and HIV vaccine MRKAd5gag, according to the severity of the AEs they induced in humans and then correlated gene expression signatures induced by these vaccines in NHP with the AEs. By using the data analysis approaches described above, we were able to identify gene modules and metagenes associated with AEs and validated the findings using a second set of six additional vaccines (Wang et al. manuscript in preparation) not included in the training set. We recently reviewed publically available transcriptomic data associated with vaccinated human and NHP whole blood or PBMC, and found some consistent results to our nonhuman primate (NHP) data sets (Wang et al. 2012a), suggesting a potential translatable value of the preclinical model used in our vaccine development programs.

5.5 Blood Gene Profiling in Clinical Practice

Blood-based mRNA biomarkers have been investigated in almost all disease areas with promising results for detecting and monitoring disease and treatment outcomes (Chaussabel et al. 2010; Fang 2007; Han et al. 2008; Hanash et al. 2011; Julia et al. 2009; Marshall et al. 2010; Pankla et al. 2009; Pascual et al. 2010; Quartier et al. 2011; Tattermusch et al. 2012). Efforts were made even in areas not

obviously or traditionally connected with blood, such as in neurological diseases (Kurian et al. 2011; Le-Niculescu et al. 2009; Runne et al. 2007; Scherzer et al. 2007) and drug toxicology (Bushel et al. 2007; Fannin et al. 2010; Huang et al. 2010; Lobenhofer et al. 2008). The biomarker discovery process is usually initiated with genome-wide expression profiling to identify significantly modulated gene sets, followed by employing more stringent criteria to down-select a small subset of genes for assay development (Barth and Hare 2006; van't Veer and Bernards 2008). Although concerns remain about the consistency of transcription profiling (Tang et al. 2010), progress made in the past few years has resulted in tremendous improvement in accuracy, sensitivity, and reproducibility of blood gene assays.

Commercially available blood mRNA assays are now available for detecting multiple diseases (Novak et al. 2012), and an 11 blood gene mRNA signature for predicting rejection following cardiac transplant has been approved by the FDA (Yamani et al. 2007a, b). More communication and collaboration among pharmaceutical industry, academia, diagnostic companies, and regulatory agencies will be needed to standardize profiling study design, data analysis/interpretation, and assay development to utilize this easily accessible tissue to its full capacity in the clinical setting.

6 Future Directions and Conclusions

Due to its high stability, reproducibility, and consistency among individuals (Chen et al. 2008a), plasma miRNA has recently attracted a lot of interest as potential biomarkers for physiological conditions such as drug-induced tissue injury (Laterza et al. 2009) and for disease diagnosis (Cortez and Calin 2009), especially cancer detection (Schrauder et al. 2012). The role other types of noncoding RNA (ncRNA) plays in human disease is being actively pursued (Esteller 2011) and should be closely monitored.

Advancement in next-generation sequencing (NGS) technologies in the past few years (Metzker 2010) has dramatically reduced the cost and time of data acquisition, opening up new areas of systems biology research such as whole genome (Cirulli and Goldstein 2010; Pleasance et al. 2010), whole exome (Clark et al. 2011), whole transcriptome (Ozsolak and Milos 2011) analysis of human patient cohorts which have already impacted medical research in an unprecedented way (Chin et al. 2011; Meyerson et al. 2010; Pleasance et al. 2010). In the infectious disease area, researchers could now expand their understanding of pathogens by performing WGS (Forgetta et al. 2011; Relman 2011) during disease progression. A recent study tracked mutations accumulated in *S. aureus* genome over a 13-month period in an infected host who progressed from carriage to disease, and identified a cluster of mutations that caused truncation of bacterial proteins which could be the cause of pathogenicity (Young et al. 2012). This type of information, when validated in a large cohort, could be utilized for designing

new therapeutics or integrated with host response to generate testable hypotheses of why certain subpopulations are more susceptible to disease.

Metagenomic sequencing of microbiome (Qin et al. 2010; Virgin and Todd 2011) and immune repertoire sequencing (Benichou et al. 2012; Boyd et al. 2009; Klarenbeek et al. 2010; Logan et al. 2011) represent two additional promising research areas enabled by the NGS technology. Sequencing the human gut microbiome, for example, could gain more detailed understanding of the microbial–human interaction which plays an important role in energy metabolism and immunity in the host. It has been increasingly appreciated that an imbalanced gut microbiota could partially result in many diseases, such as *Clostridium difficile* infection (CDI), inflammatory bowel disease (IBD), and the metabolic syndrome. Fecal microbiota transplantation (FMT) (Borody and Khoruts 2011) could become a well-established therapeutic option, once we have a more thorough understanding of the gut microbiome. Immune repertoire sequencing (Rep-seq) (Benichou et al. 2012) has been used in (1) monitoring residual disease and immune reconstitution in chronic lymphocytic leukemia (Boyd et al. 2009; Logan et al. 2011); (2) understanding diversity of T and B cell repertoires (Boyd et al. 2010; Robins et al. 2010; Venturi et al. 2011; Wang et al. 2010); and (3) producing antibodies targeting specific antigen (Reddy et al. 2010). Further investigation of Rep-seq data obtained under infection and vaccination conditions could facilitate understanding of the protective immune response and shed light in future vaccine or therapeutic antibody development.

New sequencing technologies are being developed in a rapid pace and at least two important novel platforms, one using single-molecule nanopore technology (Oxford Nanopore) (Clarke et al. 2009) and the other applying a complementary metal-oxide semiconductor (CMOS) process (Ion Torrent) (Rothberg et al. 2011), became commercially available recently. These platforms could further reduce costs, increase sequencing speeds, and enable systems-based genomics analysis of large patient cohorts. Applying innovative technologies in well-coordinated studies of appropriate preclinical models and human subjects could result in the identification of efficacious targets and robust biomarkers for predicting and reporting responses, which will increase the successful rate of drug development.

Acknowledgement The authors would like to thank Dr. Tessie McNeely (Merck Vaccines Research, West Point, PA) for carefully reading the manuscript and for her comments

References

(2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073

Abbas AR, Baldwin D, Ma Y et al (2005) Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes Immun 6:319–331

Abbas AR, Wolslegel K, Seshasayee D et al (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One 4:e6098

Allantaz F, Chaussabel D, Stichweh D et al (2007) Blood leukocyte microarrays to diagnose systemic onset juvenile idiopathic arthritis and follow the response to IL-1 blockade. J Exp Med 204:2131–2144

Arijs I, De Hertogh G, Lemaire K et al (2009) Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. PLoS One 4:e7984

Asare AL, Kolchinsky SA, Gao Z et al (2008) Differential gene expression profiles are dependent upon method of peripheral blood collection and RNA isolation. BMC Genomics 9:474

Assassi S, Mayes MD, Arnett FC et al (2010) Systemic sclerosis and lupus: points in an interferon-mediated continuum. Arthritis Rheum 62:589–598

Bacac M, Provero P, Mayran N et al (2006) A mouse stromal response to tumor invasion predicts prostate and breast cancer patient survival. PLoS One 1:e32

Banchereau R, Jordan-Villegas A, Ardura M et al (2012) Host immune transcriptional profiles reflect the variability in clinical disease manifestations in patients with *Staphylococcus aureus* infections. PLoS One 7:e34390

Barth AS, Hare JM (2006) The potential for the transcriptome to serve as a clinical biomarker for cardiovascular diseases. Circ Res 98:1459–1461

Benichou J, Ben-Hamo R, Louzoun Y et al (2012) Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology 135:183–191

Bergland AO, Genissel A, Nuzhdin SV et al (2008) Quantitative trait loci affecting phenotypic plasticity and the allometric relationship of ovariole number and thorax length in *Drosophila melanogaster*. Genetics 180:567–582

Bertos NR, Park M (2011) Breast cancer—one term, many entities? J Clin Invest 121:3789–3796

Bertram L, McQueen MB, Mullin K et al (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet 39:17–23

Borody TJ, Khoruts A (2011) Fecal microbiota transplantation and emerging applications. Nat Rev Gastroenterol Hepatol 9:88–96

Boyd SD, Marshall EL, Merker JD et al (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med 1: 12ra23

Boyd SD, Gaeta BA, Jackson KJ et al (2010) Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol 184:6986–6992

Brem RB, Yvert G, Clinton R et al (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–755

Bucasas KL, Franco LM, Shaw CA et al (2011) Early patterns of gene expression correlate with the humoral immune response to influenza vaccination in humans. J Infect Dis 203:921–929

Buchanan CC, Torstenson ES, Bush WS et al (2012) A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. J Am Med Inform Assoc 19:289–294

Burczynski ME, Dorner AJ (2006) Transcriptional profiling of peripheral blood cells in clinical pharmacogenomic studies. Pharmacogenomics 7:187–202

Bushel PR, Heinloth AN, Li J et al (2007) Blood gene expression signatures predict exposure levels. Proc Natl Acad Sci U S A 104:18211–18216

Cervino AC, Li G, Edwards S et al (2005) Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. Genomics 86:505–517

Chaussabel D, Quinn C, Shen J et al (2008) A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity 29:150–164

Chaussabel D, Pascual V, Banchereau J (2010) Assessing the human immune system through blood transcriptomics. BMC Biol 8:84

Chen X, Ba Y, Ma L et al (2008a) Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res 18:997–1006

Chen Y, Zhu J, Lum PY et al (2008b) Variations in DNA elucidate molecular networks that cause disease. Nature 452:429–435

Cheung VG, Jen KY, Weber T et al (2003) Genetics of quantitative variation in human gene expression. Cold Spring Harb Symp Quant Biol 68:403–407

Chin L, Andersen JN, Futreal PA (2011) Cancer genomics: from discovery science to personalized medicine. Nat Med 17:297–303
Christoforou A, Dondrup M, Mattingsdal M et al (2012) Linkage-disequilibrium-based binning affects the interpretation of GWASs. Am J Hum Genet 90:727–733
Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11:415–425
Clark MJ, Chen R, Lam HY et al (2011) Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29:908–914
Clarke J, Wu HC, Jayasinghe L et al (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4:265–270
Cortez MA, Calin GA (2009) MicroRNA identification in plasma and serum: a new tool to diagnose and monitor diseases. Expert Opin Biol Ther 9:703–711
Dai H, van't Veer L, Lamb J et al (2005) A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. Cancer Res 65:4059–4066
de Tayrac M, Etcheverry A, Aubry M et al (2009) Integrative genome-wide analysis reveals a robust genomic glioblastoma signature associated with copy number driving changes in gene expression. Genes Chromosom Cancer 48:55–68
Deng MC, Eisen HJ, Mehra MR et al (2006) Noninvasive discrimination of rejection in cardiac allograft recipients using gene expression profiling. Am J Transplant 6:150–160
Derry JM, Zhong H, Molony C et al (2010) Identification of genes and networks driving cardiovascular and metabolic phenotypes in a mouse F2 intercross. PLoS One 5:e14319
Dezso Z, Nikolsky Y, Nikolskaya T et al (2009) Identifying disease-specific genes based on their topological significance in protein networks. BMC Syst Biol 3:36
Dixon AL, Liang L, Moffatt MF et al (2007) A genome-wide association study of global gene expression. Nat Genet 39:1202–1207
Dobrin R, Greenawalt DM, Hu G et al (2011) Dissecting cis regulation of gene expression in human metabolic tissues. PLoS One 6:e23480
Duval E, Bigot N, Hervieu M et al (2011) Asporin expression is highly regulated in human chondrocytes. Mol Med 17:816–823
Emilsson V, Thorleifsson G, Zhang B et al (2008) Genetics of gene expression and its effect on disease. Nature 452:423–428
Esteller M (2011) Non-coding RNAs in human disease. Nat Rev Genet 12:861–874
Estrada K, Styrkarsdottir U, Evangelou E et al (2012) Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nat Genet 44:491–501
Everitt AR, Clare S, Pertel T et al (2012) IFITM3 restricts the morbidity and mortality associated with influenza. Nature 484:519–523
Fang KC (2007) Clinical utilities of peripheral blood gene expression profiling in the management of cardiac transplant patients. J Immunotoxicol 4:209–217
Fannin RD, Russo M, O'Connell TM et al (2010) Acetaminophen dosing of humans results in blood transcriptome and metabolome changes consistent with impaired oxidative phosphorylation. Hepatology 51:227–236
Forgetta V, Oughton MT, Marquis P et al (2011) Fourteen-genome comparison identifies DNA markers for severe-disease-associated strains of *Clostridium difficile*. J Clin Microbiol 49:2230–2238
Frazer KA, Ballinger DG, Cox DR et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861
Gatz M, Reynolds CA, Fratiglioni L et al (2006) Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry 63:168–174
Gaucher D, Therrien R, Kettaf N et al (2008) Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. J Exp Med 205:3119–3131
Genissel A, McIntyre LM, Wayne ML et al (2008) Cis and trans regulatory effects contribute to natural variation in transcriptome of *Drosophila melanogaster*. Mol Biol Evol 25:101–110

Geurts J, Joosten LA, Takahashi N et al (2009) Computational design and application of endogenous promoters for transcriptionally targeted gene therapy for rheumatoid arthritis. Mol Ther 17:1877–1887
Gilad Y, Oshlack A, Rifkin SA (2006) Natural selection on gene expression. Trends Genet 22:456–461
Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet 24:408–415
Gompel N, Prud'homme B, Wittkopp PJ et al (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature 433:481–487
Hambleton S, Salem S, Bustamante J et al (2011) IRF8 mutations and human dendritic-cell immunodeficiency. N Engl J Med 365:127–138
Han M, Liew CT, Zhang HW et al (2008) Novel blood-based, five-gene biomarker set for the detection of colorectal cancer. Clin Cancer Res 14:455–460
Hanash SM, Baik CS, Kallioniemi O (2011) Emerging molecular biomarkers–blood-based strategies to detect and monitor cancer. Nat Rev Clin Oncol 8:142–150
Harold D, Abraham R, Hollingworth P et al (2009) Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet 41:1088–1093
Harrison C (2011) Patent watch: the patent cliff steepens. Nat Rev Drug Discov 10:12–13
Herrup K (2010) Reimagining Alzheimer's disease—an age-based hypothesis. J Neurosci 30:16755–16762
Hess K, Grant PJ (2011) Inflammation and thrombosis in diabetes. Thromb Haemost 105 (Suppl 1):S43–S54
Hindorff LA, Sethupathy P, Junkins HA et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106:9362–9367
Hong MG, Myers AJ, Magnusson PK et al (2008) Transcriptome-wide assessment of human brain and lymphocyte senescence. PLoS One 3:e3024
Huang E, Ishida S, Pittman J et al (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. Nat Genet 34:226–230
Huang RS, Duan S, Bleibel WK et al (2007) A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proc Natl Acad Sci U S A 104:9758–9763
Huang J, Shi W, Zhang J et al (2010) Genomic indicators in the blood predict drug-induced liver injury. Pharmacogenomics J 10:267–277
Iadecola C, Anrather J (2011) The immunology of stroke: from mechanisms to translation. Nat Med 17:796–808
Julia A, Erra A, Palacio C et al (2009) An eight-gene blood expression profile predicts the response to infliximab in rheumatoid arthritis. PLoS One 4:e7556
Kalluri R, Weinberg RA (2009) The basics of epithelial-mesenchymal transition. J Clin Invest 119:1420–1428
Kapoor M, Martel-Pelletier J, Lajeunesse D et al (2011) Role of proinflammatory cytokines in the pathophysiology of osteoarthritis. Nat Rev Rheumatol 7:33–42
Keller MP, Attie AD (2010) Physiological insights gained from gene expression analysis in obesity and diabetes. Annu Rev Nutr 30:341–364
Kim SK, Yun SJ, Kim J et al (2011a) Identification of gene expression signature modulated by nicotinamide in a mouse bladder cancer model. PLoS One 6:e26131
Kim YA, Wuchty S, Przytycka TM (2011b) Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput Biol 7:e1001095
Kizawa H, Kou I, Iida A et al (2005) An aspartic acid repeat polymorphism in asporin inhibits chondrogenesis and increases susceptibility to osteoarthritis. Nat Genet 37:138–144
Klarenbeek PL, Tak PP, van Schaik BD et al (2010) Human T-cell memory consists mainly of unexpanded clones. Immunol Lett 133:42–48
Klein RJ, Zeiss C, Chew EY et al (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389
Kola I (2008) The state of innovation in drug development. Clin Pharmacol Ther 83:227–230

Kola I, Hazuda D (2005) Innovation and greater probability of success in drug discovery and development—from target to biomarkers. Curr Opin Biotechnol 16:644–646
Kola I, Landis J (2004) Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 3:711–715
Koscielny S (2008) Critical review of microarray-based prognostic tests and trials in breast cancer. Curr Opin Obstet Gynecol 20:47–50
Kurian SM, Le-Niculescu H, Patel SD et al (2011) Identification of blood biomarkers for psychosis using convergent functional genomics. Mol Psychiatry 16:37–58
Lamb JR, Zhang C, Xie T et al (2011) Predictive genes in adjacent normal tissue are preferentially altered by sCNV during tumorigenesis in liver cancer and may rate limiting. PLoS One 6:e20090
Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048
Laterza OF, Lim L, Garrett-Engele PW et al (2009) Plasma MicroRNAs as sensitive and specific biomarkers of tissue injury. Clin Chem 55:1977–1983
Le-Niculescu H, Kurian SM, Yehyawi N et al (2009) Identifying blood biomarkers for mood disorders using convergent functional genomics. Mol Psychiatry 14:156–174
Li Q, Smith AJ, Schacker TW et al (2009) Microarray analysis of lymphatic tissue reveals stage-specific, gene expression signatures in HIV-1 infection. J Immunol 183:1975–1982
Liu S, Umezu-Goto M, Murph M et al (2009) Expression of autotaxin and lysophosphatidic acid receptors increases mammary tumorigenesis, invasion, and metastases. Cancer Cell 15:539–550
Lobenhofer EK, Auman JT, Blackshear PE et al (2008) Gene expression response in target organ and whole blood varies as a function of target organ injury phenotype. Genome Biol 9:R100
Loboda A, Nebozhyn M, Klinghoffer R et al (2010) A gene expression signature of RAS pathway dependence predicts response to PI3 K and RAS pathway inhibitors and expands the population of RAS pathway activated tumors. BMC Med Genomics 3:26
Loboda A, Nebozhyn MV, Watters JW et al (2011) EMT is the dominant program in human colon cancer. BMC Med Genomics 4:9
Logan AC, Gao H, Wang C et al (2011) High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. Proc Natl Acad Sci U S A 108:21194–21199
Lum PY, Chen Y, Zhu J et al (2006) Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. J Neurochem 97(Suppl 1):50–62
MacArthur DG, Balasubramanian S, Frankish A et al (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828
Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–753
Marshall KW, Mohr S, Khettabi FE et al (2010) A blood-based biomarker panel for stratifying current risk for colorectal cancer. Int J Cancer 126:1177–1186
McGregor AP, Orgogozo V, Delon I et al (2007) Morphological evolution through multiple cis-regulatory mutations at a single gene. Nature 448:587–590
Mehrabian M, Allayee H, Stockton J et al (2005) Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. Nat Genet 37:1224–1233
Mendrick DL (2011) Transcriptional profiling to identify biomarkers of disease and drug response. Pharmacogenomics 12:235–249
Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46
Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11:685–696
Min JL, Nicholson G, Halgrimsdottir I et al (2012) Coexpression network analysis in abdominal and gluteal adipose tissue reveals regulatory genetic loci for metabolic syndrome and related phenotypes. PLoS Genet 8:e1002505
Mocsai A, Ruland J, Tybulewicz VL (2010) The SYK tyrosine kinase: a crucial player in diverse biological functions. Nat Rev Immunol 10:387–402

Moffatt MF, Kabesch M, Liang L et al (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448:470–473
Mohr S, Liew CC (2007) The peripheral-blood transcriptome: new insights into disease and risk assessment. Trends Mol Med 13:422–432
Monks SA, Leonardson A, Zhu H et al (2004) Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75:1094–1105
Mook S, Van't Veer LJ, Rutgers EJ et al (2007) Individualization of therapy using Mammaprint: from development to the MINDACT Trial. Cancer Genomics Proteomics 4:147–155
Morley M, Molony CM, Weber TM et al (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743–747
Naj AC, Jun G, Beecham GW et al (2011) Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet 43:436–441
Nakaya HI, Wrammert J, Lee EK et al (2011) Systems biology of vaccination for seasonal influenza in humans. Nat Immunol 12:786–795
Novak DJ, Liew GJ, Liew CC (2012) GeneNews limited: bringing the blood transcriptome to personalized medicine. Pharmacogenomics 13:381–385
Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. Nat Genet 32:261–266
O'Roak BJ, Vives L, Girirajan S et al (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485:246–250
Ortutay C, Vihinen M (2009) Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. Nucleic Acids Res 37:622–628
Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12:87–98
Palermo RE, Patterson LJ, Aicher LD et al (2011) Genomic analysis reveals pre- and postchallenge differences in a rhesus macaque AIDS vaccine trial: insights into mechanisms of vaccine efficacy. J Virol 85:1099–1116
Pankla R, Buddhisa S, Berry M et al (2009) Genomic transcriptional profiling identifies a candidate blood biomarker signature for the diagnosis of septicemic melioidosis. Genome Biol 10:R127
Parrish ML, Wright C, Rivers Y et al (2010) cDNA targets improve whole blood gene expression profiling and enhance detection of pharmocodynamic biomarkers: a quantitative platform analysis. J Transl Med 8:87
Pascual V, Chaussabel D, Banchereau J (2010) A genomic approach to human autoimmune diseases. Annu Rev Immunol 28:535–571
Peake J, Della Gatta P, Cameron-Smith D (2010) Aging and its effects on inflammation in skeletal muscle at rest and following exercise-induced muscle injury. Am J Physiol Regul Integr Comp Physiol 298:R1485–R1495
Pedraza V, Gomez-Capilla JA, Escaramis G et al (2010) Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness. Cancer 116:486–496
Pleasance ED, Cheetham RK, Stephens PJ et al (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463:191–196
Podtelezhnikov AA, Tanis KQ, Nebozhyn M et al (2011) Molecular insights into the pathogenesis of Alzheimer's disease and its relationship to normal aging. PLoS One 6:e29610
Prezeau N, Silvy M, Gabert J et al (2006) Assessment of a new RNA stabilizing reagent (Tempus Blood RNA) for minimal residual disease in onco-hematology using the EAC protocol. Leuk Res 30:569–574
Puig O, Yuan J, Stepaniants S et al (2011) A gene expression signature that classifies human atherosclerotic plaque by relative inflammation status. Circ Cardiovasc Genet 4:595–604
Qin J, Li R, Raes J et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65

Quartier P, Allantaz F, Cimaz R et al (2011) A multicentre, randomised, double-blind, placebo-controlled trial with the interleukin-1 receptor antagonist anakinra in patients with systemic-onset juvenile idiopathic arthritis (ANAJIS trial). Ann Rheum Dis 70:747–754
Querec TD, Akondy RS, Lee EK et al (2009) Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol 10:116–125
Quintas-Cardama A, Verstovsek S (2011) New JAK2 inhibitors for myeloproliferative neoplasms. Expert Opin Investig Drugs 20:961–972
Quintas-Cardama A, Kantarjian H, Cortes J et al (2011) Janus kinase inhibitors for the treatment of myeloproliferative neoplasias and beyond. Nat Rev Drug Discov 10:127–140
Radich JP, Mao M, Stepaniants S et al (2004) Individual-specific variation of gene expression in peripheral blood leukocytes. Genomics 83:980–988
Rainen L, Oelmueller U, Jurgensen S et al (2002) Stabilization of mRNA expression in whole blood samples. Clin Chem 48:1883–1890
Reddy ST, Ge X, Miklos AE et al (2010) Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. Nat Biotechnol 28:965–969
Reif DM, McKinney BA, Motsinger AA et al (2008) Genetic basis for adverse events after smallpox vaccination. J Infect Dis 198:16–22
Relman DA (2011) Microbial genomics and infectious diseases. N Engl J Med 365:347–357
Robins HS, Srivastava SK, Campregher PV et al (2010) Overlap and effective size of the human CD8 + T cell receptor repertoire. Sci Transl Med 2:47ra64
Roses AD, Burns DK, Chissoe S et al (2005) Disease-specific target selection: a critical first step down the right road. Drug Discov Today 10:177–189
Rothberg JM, Hinz W, Rearick TM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352
Runne H, Kuhn A, Wild EJ et al (2007) Analysis of potential transcriptomic biomarkers for Huntington's disease in peripheral blood. Proc Natl Acad Sci U S A 104:14424–14429
Saxena R, Elbers CC, Guo Y et al (2012) Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. Am J Hum Genet 90:410–425
Schadt EE, Monks SA, Drake TA et al (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297–302
Schadt EE, Lamb J, Yang X et al (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710–717
Schadt EE, Molony C, Chudin E et al (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6:e107
Schadt EE, Friend SH, Shaywitz DA (2009) A network view of disease and compound screening. Nat Rev Drug Discov 8:286–295
Scherer A, Krause A, Walker JR et al (2003) Early prognosis of the development of renal chronic allograft rejection by gene expression profiling of human protocol biopsies. Transplantation 75:1323–1330
Scherzer CR, Eklund AC, Morse LJ et al (2007) Molecular markers of early Parkinson's disease based on gene expression in blood. Proc Natl Acad Sci U S A 104:955–960
Schmidt N, Gonzalez E, Visekruna A et al (2010) Targeting the proteasome: partial inhibition of the proteasome by bortezomib or deletion of the immunosubunit LMP7 attenuates experimental colitis. Gut 59:896–906
Schrauder MG, Strick R, Schulz-Wendtland R et al (2012) Circulating micro-RNAs as potential blood-based markers for early stage breast cancer detection. PLoS One 7:e29770
Shi D, Nakamura T, Dai J et al (2007) Association of the aspartic acid-repeat polymorphism in the asporin gene with age at onset of knee osteoarthritis in Han Chinese population. J Hum Genet 52:664–667
Song JH, Lee HS, Kim CJ et al (2008) Aspartic acid repeat polymorphism of the asporin gene with susceptibility to osteoarthritis of the knee in a Korean population. Knee 15:191–195
Sorlie T, Tibshirani R, Parker J et al (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 100:8418–8423

Stern DL (1998) A role of Ultrabithorax in morphological differences between Drosophila species. Nature 396:463–466
Stienstra R, Tack CJ, Kanneganti TD et al (2012) The inflammasome puts obesity in the danger zone. Cell Metab 15:10–18
Tang BM, Huang SJ, McLean AS (2010) Genome-wide transcription profiling of human sepsis: a systematic review. Crit Care 14:R237
Tattermusch S, Skinner JA, Chaussabel D et al (2012) Systems biology approaches reveal a specific interferon-inducible signature in HTLV-1 associated myelopathy. PLoS Pathog 8:e1002480
Tian Z, Palmer N, Schmid P et al (2009) A practical platform for blood biomarker study by using global gene expression profiling of peripheral whole blood. PLoS One 4:e5157
Tiffin N, Okpechi I, Perez-Iratxeta C et al (2008) Prioritization of candidate disease genes for metabolic syndrome by computational analysis of its defining phenotypes. Physiol Genomics 35:55–64
Trusheim MR, Burgess B, Hu SX et al (2011) Quantifying factors for the success of stratified medicine. Nat Rev Drug Discov 10:817–833
Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98:5116–5121
Vahey MT, Wang Z, Kester KE et al (2010) Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS, S malaria vaccine. J Infect Dis 201:580–589
van Baarsen LG, Vosslamber S, Tijssen M et al (2008) Pharmacogenomics of interferon-beta therapy in multiple sclerosis: baseline IFN signature determines pharmacological differences between patients. PLoS One 3:e1927
van de Vijver MJ, He YD, van't Veer LJ et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999–2009
van 't Veer LJ, Dai H, van de Vijver MJ et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536
van 't Veer LJ, Dai H, vvan de Vijver MJ et al (2003) Expression profiling predicts outcome in breast cancer. Breast Cancer Res 5:57–58
van't Veer LJ, Bernards R (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 452:564–570
Vartanian K, Slottke R, Johnstone T et al (2009) Gene expression profiling of whole blood: comparison of target preparation methods for accurate and reproducible microarray analysis. BMC Genomics 10:2
Venturi V, Quigley MF, Greenaway HY et al (2011) A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. J Immunol 186:4285–4294
Virgin HW, Todd JA (2011) Metagenomics and personalized medicine. Cell 147:44–56
Walczak H (2011) TNF and ubiquitin at the crossroads of gene activation, cell death, inflammation, and cancer. Immunol Rev 244:9–28
Wang C, Sanders CM, Yang Q et al (2010) High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. Proc Natl Acad Sci U S A 107:1518–1523
Wang IM, Bett AJ, Cristescu R et al (2012a) Transcriptional profiling of vaccine-induced immune responses in humans and non-human primates. Microb Biotechnol 5:177–187
Wang ZY, Fu LY, Zhang HY (2012b) Can medical genetics and evolutionary biology inspire drug target identification? Trends Mol Med 18:69–71
Wang IM, Zhang B, Yang X et al (2012c) Systems analysis of eleven rodent disease models reveals an inflammatome signature and key drivers. Mol Syst Biol 8:594
Weber C, Noels H (2011) Atherosclerosis: current pathogenesis and therapeutic options. Nat Med 17:1410–1422
Wen H, Ting JP, O'Neill LA (2012) A role for the NLRP3 inflammasome in metabolic diseases—did Warburg miss inflammation? Nat Immunol 13:352–357

Whitney AR, Diehn M, Popper SJ et al (2003) Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A 100:1896–1901

Yamani MH, Taylor DO, Haire C et al (2007a) Post-transplant ischemic injury is associated with up-regulated AlloMap gene expression. Clin Transplant 21:523–525

Yamani MH, Taylor DO, Rodriguez ER et al (2007b) Transplant vasculopathy is associated with increased AlloMap gene expression score. J Heart Lung Transplant 26:403–406

Yang X, Deignan JL, Qi H et al (2009) Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. Nat Genet 41:415–423

Yao Y, Richman L, Morehouse C et al (2008) Type I interferon: potential therapeutic target for psoriasis? PLoS One 3:e2737

Young BC, Golubchik T, Batty EM et al (2012) Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proc Natl Acad Sci U S A 109:4550–4555

Zaas AK, Chen M, Varkey J et al (2009) Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. Cell Host Microbe 6:207–217

Zhong H, Beaulaurier J, Lum PY et al (2010a) Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. PLoS Genet 6:e1000932

Zhong H, Yang X, Kaplan LM et al (2010b) Integrating pathway analysis and genetics of gene expression for genome-wide association studies. Am J Hum Genet 86:581–591

Zhu J, Lum PY, Lamb J et al (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet Genome Res 105:363–374

Zhu J, Zhang B, Schadt EE (2008) A systems biology approach to drug discovery. Adv Genet 60:603–635

Zuk O, Hechter E, Sunyaev SR et al (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109:1193–1198

Insights into Proteomic Immune Cell Signaling and Communication via Data-Driven Modeling

Kelly F. Benedict and Douglas A. Lauffenburger

Abstract Over the past decade, studies applying data-driven modeling approaches have demonstrated significant contributions toward the integrative understanding of multivariate cell regulatory system operation. Here we review applications of several of these approaches, including principal component analysis, partial least squares regression, partial least squares discriminant analysis, decision trees, and Bayesian networks, and describe the advances they have offered in systems-level understanding of immune cell signaling and communication. We show how these approaches generate novel insights from high-throughput proteomic data, from classification to association to influence to mechanisms. Looking forward, new experimental technologies involving single-cell measurements of cytokine expression beckon extension of these modeling techniques to inference of immune cell–cell communication networks, with a goal of aiding development of improved vaccine therapeutics.

Contents

K. F. Benedict · D. A. Lauffenburger (✉)
Department of Biological Engineering, Massachusetts Institute of Technology,
Room: 16-343, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
e-mail: lauffen@mit.edu

Current Topics in Microbiology and Immunology (2013) 363: 201–233
DOI: 10.1007/82_2012_249

Published Online: 10 August 2012

1 Introduction

The human immune system is complex at all scales, spanning the molecular level of translational and transcriptional events within cells, intracellular protein signaling interactions, cell–cell communication via soluble cytokines, and organ function. In recent years, systems biology approaches have yielded new kinds of insight into different components of the immune system through findings that emphasize a multivariate integrative perspective on systems-level properties and function.

Many of these new insights have been generated using theory-driven, also called knowledge-based approaches, where mathematical models are constructed based on theorized hypotheses (Fig. 1). In this approach, published literature and previous experimental results are used to guide construction of the mathematical model. Choice of model boundaries, important input/output relationships, key species, and parameter values are all selected based on prior knowledge and hypotheses governing the system of interest. Interactions between species are described with mathematical relationships, often in the form of differential equations. These theory-driven approaches enable crucial hypothesis-testing of system-level properties and evaluation of parameter values in the broad context of an entire signaling network (Benedict et al. 2011), both of which are difficult to obtain with experimental work alone. In the field of immunology, a deep body of the experimental literature and broad range of experimental assays for measuring transcriptional and intracellular protein signaling events have allowed for valuable use of theory-driven approaches, including understanding how various IκB proteins affect NFκB signaling dynamics (Hoffmann et al. 2002), evaluating APOBEC3G- and Vif-based therapeutic strategies for HIV infection (Hosseini and Gabhann 2012), elucidating the role of shared receptor components, ligand competition, and feedback loops in IL-7 signaling (Palmer et al. 2008), and understanding how Th and Treg IL-2 feedback loops and signaling dynamics shape different cellular microenvironments (Busse et al. 2010).

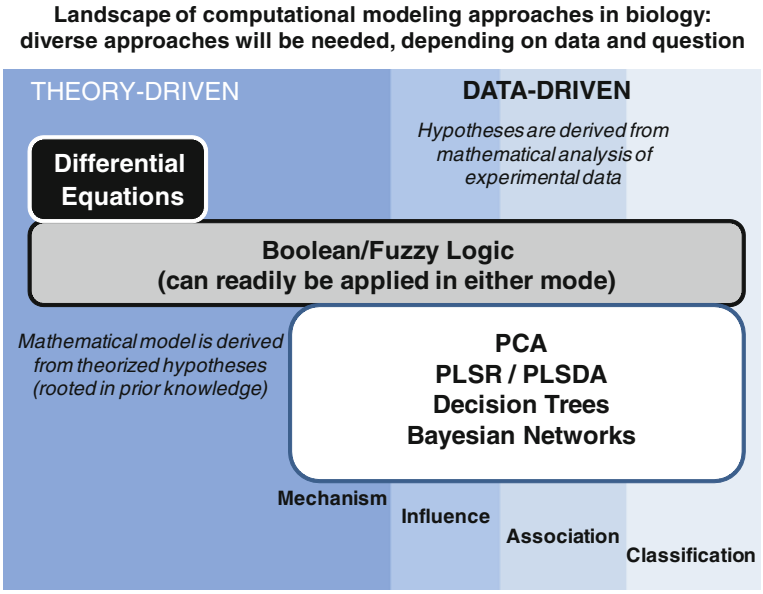


Fig. 1 Data-driven modeling techniques can give a wide range of insight into biological events. Type of insight gained is not dependent on specific technique used, but rather on questions asked and type of data used, as illustrated by recent work that has used a number of techniques to gain different types of insight

As we move forward in understanding the human immune system, it will be crucial to evaluate the immune system at multiple scales, beyond intracellular signaling to cell phenotypes, cell–cell interactions, and in vivo tissue function. In contrast to intracellular cell signaling pathways which have been comparatively well-characterized, in order to understand immune system function at multiple scales we will need to quantitatively identify key elements of systems that we have little current knowledge of, including relevant systems-level interactions, appropriate boundaries, and important input/output relationships. Data-driven modeling approaches offer a valuable complementary approach to theory-driven models and enable identification and study of poorly characterized systems using data obtained directly from a given system to quantitatively characterize it. In data-driven approaches, hypotheses are derived directly from mathematical analysis of experimental data in contrast to theory-driven approaches that utilize prior knowledge (Fig. 1). Mathematical relationships are delineated to link system components to each other as well as to important system input or output parameters. This is done based on data alone and without prior knowledge of system function. Data-driven approaches also offer the ability to integrate data obtained from different sources and across different physiologic scales, which will be critical for moving toward in vivo study of immune system function.

Insights gained from data-driven approaches applied to immunological systems can span a broad spectrum of specificity, from broad, high-level classification and prediction to association, influence, and even new mechanistic insight (Fig. 1). Previous work (we review here) using data-driven techniques such as principal component analysis (PCA), partial least squares regression (PLSR), partial least squares discriminant analysis (PLSDA), decision trees, and Bayesian inference networks (Table 1) has shown that the type of insight gained from data-driven modeling approaches is not dependent on the specific type of analysis used, but rather on the questions asked, the nature and quantity of available data, and the contextual settings in which the approach is utilized.

One useful application of data-driven techniques is as a methodology for classifying or predicting important biological events. This is especially useful in clinical settings where it is necessary to rapidly and efficiently differentiate between different immune states or cell types, such as the examples we review here, including differentiation between different types of infections (Prilutsky et al. 2011), or screening high quality cells for use in cell therapy (Rivet et al. 2011). In these settings, distinguishing between biological states is of greater importance than knowledge of specific mechanisms governing these differences. Since, accuracy is of higher importance than detailed mechanisms, these approaches often require larger amounts of data but less biologically meaningful parameters.

Identifying new system boundaries requires association of unknown system components with an important system behavior or output of interest. Data-driven approaches can be used to identify important associations across different physiologic scales and with different types of experimental data. For instance, given a large complex data set comprising measurements of molecular regulatory activities, data-driven approaches can extract groups of molecular activities that are statistically associated with a given cell phenotype or behavior. They can also associate signaling events with tissue- and patient-level characteristics, such as disease state, prognosis, or likelihood of response to treatment. Even if associations between events and states are not different enough to be used as a robust predictive tool, they can still generate new hypotheses for follow-up studies, especially when biologically meaningful data is used.

Maps of molecular regulatory activities, such as proteomic signaling or transcriptional processes, have been traditionally created based on the intuitive aggregation of results from separate experimental studies, each of which focused on different parts of the map. In contrast, data-driven techniques can systematically generate influence maps based on only high-throughput experimental data, usually with data obtained before and after some perturbation. For example, given a set of protein signaling measurements made in a resting state that are also measured after perturbation of different signaling nodes, data-driven techniques can generate a predicted connectivity map of influence for all molecular species measured in the data. Though the approach does not require prior knowledge, prior knowledge is useful to guide the selection of biological measurements to include in the high-throughput data used to create the influence map. The more biologically meaningful the data is, the more biologically meaningful the influence map will be.

Table 1 Summary of data-driven modeling approaches

Approach	Data	Method	References
PCA	A set of observations, each with measured features (X)	Identifies systematic orthogonal patterns of features (principal components) that account for variation of X	• Janes and Yaffe (2006) • Martens and Martens (2001)
PLSDA	A set of observations, each with measured features (X) and a known class label (Y)	Identifies systematic orthogonal patterns of features (latent variables) that best discriminate between classes Y	• Janes and Yaffe (2006) • Martens and Martens (2001)
PLSR	A set of observations, each with measured features (X) and some other quantitative variable(s) of interest (Y)	Identifies systematic orthogonal patterns of features (latent variables) that account for covariance of Y with X	• Janes and Yaffe (2006) • Martens and Martens (2001)
Decision trees	A set of observations, each with measured features (X) and a qualitative or quantitative class label (Y)	Classifies the data by class labels (Y) after iterative separation based on measured features (X). Result is a tree-like diagram illustrating the hierarchy of importance of various features (X) in classifying the data based on class labels (Y)	• Kingsford and Salzberg (2008) • Geurts et al. (2009)
Bayesian inference	Activity measurements of biological species, before and after some perturbation(s)	Determines probabilistic relationships between species, based on pre and post-perturbation activity levels. Result is an influence map illustrating dependence relationships among species	• Pe'er (2005) • Friedman (2000)

Data-driven techniques can also identify important new systems-level mechanisms, especially when coupled with careful follow-up experiments. Such insight is best gained when the data that are used for the model is biologically meaningful, and model predictions are confirmed with interventional follow-up experiments. For example, this type of approach can identify new combinatorial protein signaling events relevant to a phenotype or behavior. Combinatorial events identified can then be used to generate new hypotheses for system-level mechanisms, such as feedback loops and other pathway crosstalk, that can be verified with specific stimulators or inhibitors.

The value of data-driven approaches in clinically relevant immunological settings has been demonstrated at the transcriptional level with the use of human immune cell microarray data for prediction of patient responses to yellow fever and seasonal flu vaccines (Querec et al. 2009; Nakaya et al. 2011). These studies demonstrate the value of data-driven approaches in predicting patient responses and hint at the potential for identifying critical new mechanisms of action. Recent advances in experimental technology have allowed for the acquisition of high-throughput proteomic signaling measurements from biological specimens and the subsequent assembly of large complex data sets that are similar in size to microarray data. In this review, we focus on the use of data-driven approaches in evaluation of proteomic data, especially as it becomes more broadly available in high-throughput form. Evaluation of proteomic data will be especially important as we continue to study the immune system at different scales.

Here, we review recent studies that demonstrate applications of data-driven modeling using a wide range of techniques, including PCA, PLSR, PLSDA, decision trees, and Bayesian inference (Table 1). We highlight how the questions asked, the contextual setting, and the data available help determine the type of insight that can be gained using different techniques.

2 Classification/Prediction Insights

The ability to distinguish between biological events or states can be a valuable diagnostic tool, even without accompanying knowledge of associated mechanistic differences. Data-driven modeling approaches have the capacity to differentiate events or classes based on linear combinations of multiple features. This is often more powerful than traditional approaches in immunology that identify differences based on one or two features. Below we highlight several examples from the recent literature, where data-driven modeling of proteomic events was employed to enhance either classification or prediction in immunology applications.

2.1 Assessment of Cytotoxic T Cell Age for Adoptive T Cell Therapy of Cancer

The power of a multivariate approach has recently been demonstrated for applications in adoptive T cell therapy used to treat melanoma, non-Hodgkin's lymphoma, chronic lymphocytic leukemia, and neuroblastoma. In adoptive T cell therapy, tumor-specific CD8+ T cells harvested from a given patient are clonally expanded ex vivo and transferred back to the patient to promote an anti-tumor response. One major difficulty of the approach is loss of in vivo tumor specificity during clonal expansion of the T cells, and the transition of expanded cells to senescent states.

A recent study (Rivet et al. 2011) employed multivariate analysis to determine T cell senescence based on cell surface markers and intracellular protein signaling events in four healthy donors. A newly developed microfluidic device allowed for flow cytometric measurement of CD28, CD27, cell shape, and cell size in parallel with the dynamic phosphorylation of six proteins (CD3, CREB, ERK, LAT, Lck, and Zap70) downstream of T cell activation signaling at eight time points, from 0.5 to 7 min. After stimulation with IL-2 and bead-based anti-CD3+, CD8+ T cells reached replicative senescence after 12 population doublings and this was associated with changes in some individual measurements, including a decrease in the proportion of cells in S/G2 phases, decreased cell size, increased variance in cell shape, a decreased number of cells expressing CD28, decreased mean fluorescent intensity (MFI) of CD27 expression, and a global decrease in the magnitude of peak activation levels of all proteins.

Though there were changes in individual measurements, none of the signaling or surface expression markers were individually sufficient to distinguish populations based on days in culture. Though the expression of surface markers CD28 and CD27 generally decreased with age, they could not robustly differentiate senescent T cell populations because their quantitative expression level varied greatly between donors. Clusters generated by hierarchical clustering methods were also donor-specific and unable to identify groups of markers related to age that would be predictive across all donors.

Though individual measurements were unable to identify factors related to T cell age across all donors, a PLSR model suggested that multivariate combinations of protein expression measurements were able to identify robust differences associated with T cell senescence across all donors. In the PLSR model, 48 signaling measurements, CD28, CD27, and CD3 expression levels, and flow cytometry measurements of cell shape and size were regressed against T cell days in culture and number of population doublings. Multivariate combinations of these measurements were sufficient for prediction of both T cell days in culture and population doublings metrics, with a goodness of fit of $R^2 = 0.96$ and variance captured with $Q^2 = 0.78$. Together, the first two principal components were most informative for separating data based on T cell age (Fig. 2a). A plot of the loadings of the first principal component revealed that a combination of heterogeneity in cell shape, CD57

expression, and basal level of phosphorylated-ERK were associated with increased T cell age (Fig. 2b).

Cross validation was used to assess the robustness of the model in predicting unknown data. In this approach, data from each donor were iteratively omitted before the PLSR model was trained to data from the remaining three donors. Regression was performed on the mean of the four different model predictions resulting in an R^2 value of 0.84 for model predictions of days in culture and an R^2 value of 0.94 for model predictions of population doublings, suggesting the model would be useful for prediction of unknown data.

Signaling information alone was sufficient to make predictions, as a model with only signaling information was able to predict days in culture and population doublings with regression coefficients of 0.84 and 0.94, respectively. Early instant derivatives of ERK and Zap70 around 1–1.5 min were most informative for predictions made with the signaling model. Likewise, the model with cell surface marker data was also able to make good predictions, with regression coefficients of 0.78 for predicting days in culture and 0.98 for predicting population doublings.

A PLSR model was also used to explore the relationship between surface marker expression and signaling information. Two models were generated to determine if (1) surface marker expression could predict signaling information and (2) signaling information could predict surface marker expression. Results indicated that surface marker expression was not sufficient to predict signaling information ($R^2 = 0.27$, $Q^2 = 0.1$). In contrast, signaling information was sufficient for predicting surface marker expression (correlation coefficient ranging from 0.75 to 0.91). The instant derivative of ERK phosphorylation was most associated with CD27 MFI and there were strong correlations between Lck phosphorylation and CD28 expression. Early signaling dynamics of ERK, Lck, and LAT were also highly related to CD28 expression.

Overall, this approach illustrated how multivariate analysis of high-throughput proteomic data can overcome inter-donor variability to distinguish populations of CD8+ T cells that are most useful for T cell therapy, where previous univariate analysis of surface markers was not able to do so. PLSR allowed for the integration of data from different assays and across different scales, including population-level dynamic phosphor-signaling events and single-cell flow cytometry measurements related to shape and size. New combinatorial boundaries and interactions were discovered that may be associated with T cell senescence, including the signaling proteins Lck and ERK, and the surface markers CD28, and CD27. Though the key focus of this work was multivariate analysis for classification and prediction of T cell senescent states, it also provided new hypotheses for systems level mechanisms involving cell surface markers (CD28 and CD27) and early signaling dynamics of ERK, Lck, and LAT. It demonstrated how PLSR and high-throughput proteomic data can be used together to discover new associations between immune cell surface marker expression and downstream signaling events. Follow-up experimental work could use these hypotheses to identify new systems-level mechanisms.

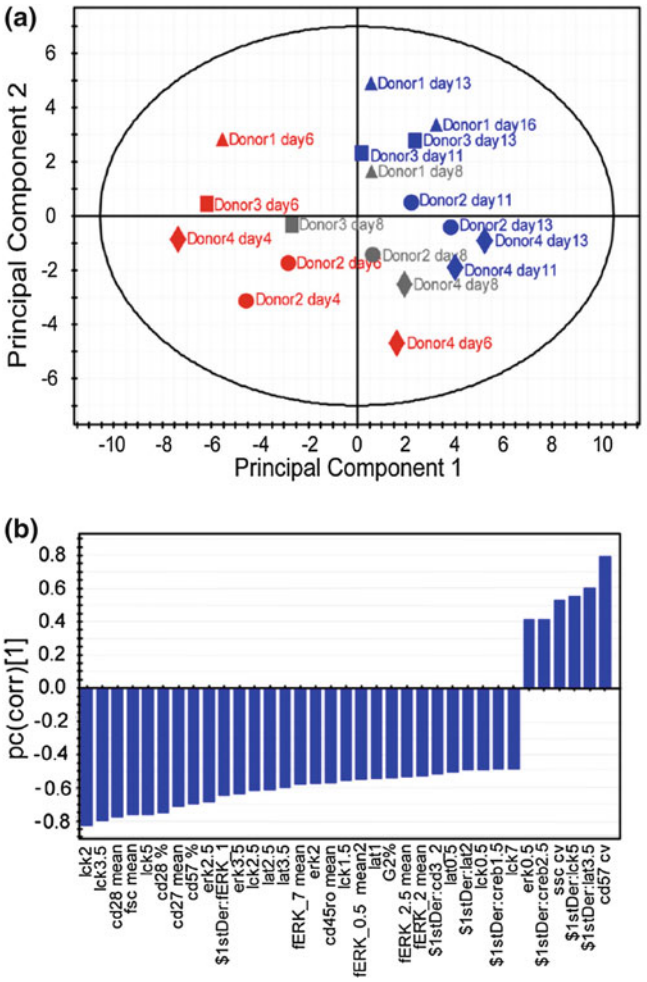


Fig. 2 **a** Using 48 dynamic cell signaling measurements, cell surface markers, and flow cytometry measurements of cell shape and size, PLSR was able to differentiate CD8+ T cells based on time in cell culture. **b** The first principal component indicated that heterogeneity in cell shape, CD57 expression, and basal level of phosphorylated-ERK was most associated with a longer number of days in culture (Rivet et al. 2011). Reprinted with permission from The American Society for Biochemistry and Molecular Biology

2.2 Differentiation Between Bacterial and Viral Infection with Chemiluminescent Signatures of Circulating Phagocytes

Rapid and sensitive differentiation between bacterial and viral infections in patients is crucial for limiting adverse side effects, controlling antibiotic resistance, and reducing healthcare costs. Current diagnoses rely on time consuming methods

that are not always sensitive or specific, such as bacterial culture, X-ray scans, PCR for viral antigens, and white blood cell counts (Prilutsky et al. 2011). A promising new diagnostic tool to aid clinicians in differentiating between bacterial and viral infections involves the use of multivariate analysis of whole-blood measurements of reactive oxygen species (ROS) production by phagocytes, immune cells that play an important role in defense against bacterial and viral infections. After encountering a pathogen, phagocytes increase their production of ROS, which can be measured by light emission, or chemiluminescence (CL), upon the addition of luminol. Since previous work had indicated that the metabolic activity of phagocytes was different in bacterial and viral infection, Prilutsky et al. (2011) examined whether multivariate analysis of CL measurement of ROS generation by phagocytes might be a rapid, sensitive method for distinguishing between bacterial and viral infections.

The approach was tested on 69 infected patients: 33 with diagnosed bacterial infections from X-rays and positive blood culture findings, and 36 with probable viral infections (nonbacterial). Six healthy patients were used as controls. Luminol and zymosan were added to fresh whole blood and CL was measured for three systems of ROS measurement, termed standard, priming, and aging. Experimental CL curves were recorded and kinetic parameters were calculated from each curve including: (1) extracellular ROS generation connected to phagocytosis, (2) intracellular ROS generation connected to phagocytosis, and (3) intracellular ROS generation not connected to phagocytosis. Parameters (1) and (3) were each split further into two components. Overall data for each patient included 82 different parameters derived from the three standard, priming, and aging systems. The CL curves indicated an increased intracellular ROS generation in bacterial infections compared to viral infections. In contrast, CL curves from viral infection indicated an increase in extracellular ROS generation.

Among all infected donors, 51 were selected for training the model and 18 were used to test the predictive power of the model. A C4.5 decision tree algorithm was used to classify the patients as viral, bacterial, or control cases based on the 82 kinetic parameters derived from the CL curves. Tenfold, stratified cross-validation was applied in ten iterations to determine the model's predictive ability in the training data. Overall, the C4.5 decision tree algorithm was able to accurately classify 94.7 % of the data in the training set and 69.2 % of the data in cross validation, making it a better predictive model when compared to other machine learning methods for classification that were tested, including Support Vector Machines and Naïve Bayes classification.

The decision tree (Fig. 3) identified CapSA (the capacity of the aging system) to be the most important parameter for differentiating bacterial infection from healthy control. RelEff_SA (the relative effectiveness of the aging system compared to the standard system) appeared in two different decision nodes (Fig. 3) in the decision tree and was most important for distinguishing bacterial from viral infections, as it was higher in most bacterial infections. RelEff_SP (the relative effectiveness of the primed system when compared with the standard system) was

also important and found to be higher in some of bacterial cases when compared with the viral cases.

Three parameters were most important in differentiating viral cases from others: Time_nonPhago1_SA (the time to peak of the last portion of non-phagocytosis-related CL of the aging system), SlopeSP (the ratio of the peak magnitude to time required to reach this point), and RelEff_SA (the relative effectiveness of primed and aged systems), both of which were lower in viral infections compared to bacterial. Using the C4.5 decision tree algorithm, 88.9 % (16 out of 18 patients) of data in the test set were correctly classified with 75 % prediction accuracy for test bacterial cases and 100 % accuracy for test viral cases.

This work was a superb illustration of how measurements with little biological meaning may be extraordinarily valuable in clinical settings if they enable differentiation between two clinical states. Though the kinetic parameters derived from chemiluminescent curves provided little information regarding metabolic events in phagocytes in response to viral and bacterial pathogens, in combination with a decision tree algorithm they were a powerful diagnostic tool for rapid and sensitive differentiation between viral and bacterial infection in clinical settings.

3 Association Insights

Often the main components of a system of interest in immunological research are largely unknown. Donor-to-donor variability and differences in methods of experimental measurement make it difficult to isolate the most important protein signaling events driving a given immune phenotype, disease state, or behavior. Data-driven modeling approaches used in combination with high-throughput data sets allow for the selection of fundamental systems-level interactions associated with a phenotype of interest. Though new associations may not provide direct mechanistic insight, they indicate new interactions, boundaries, input, and output that may be most relevant to a given behavior. This sets a new framework for future experimental studies and knowledge-based modeling efforts.

3.1 Association of Protein Expression with CD8+ T Cell Phenotype

Cytotoxic CD8+ T cells are key coordinators of the immune system that mediate the killing of pathogen-infected cells. A number of experimental methods have been developed to study the function of CD8+ T cells, but donor-to-donor variability and limitations in experimental technology have made it difficult to determine the breadth of their function in relation to phenotype. Cell surface markers CCR7 and CD45RA have been identified as consistent markers for

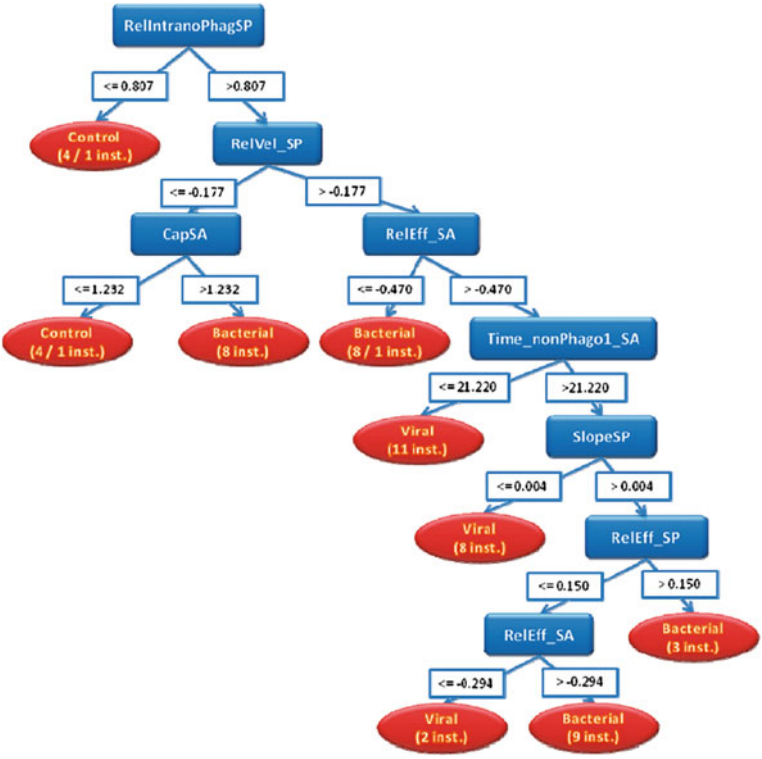


Fig. 3 A decision tree algorithm was able to classify infections in as bacterial, viral, or control based on 82 kinetic parameters derived from chemiluminescent curves measured in whole blood from 75 patients. Reprinted from Pritlusky et al. (2011), with permission. Copyright 2011 American Chemical Society

different subsets of CD8+ T cells that have been exposed to antigen or are antigen naïve, respectively. However, surprisingly little is known about the number and types of cytokines secreted by different CD8+ T cell subsets, and how this function is related to antigen specificity and killing ability.

Newell et al. (2012) developed a new experimental technology to obtain high-throughput protein expression measurements from different CD8+ T cell subsets and used PCA to associate protein expression with various T cell subsets. CD8+ T cells from six healthy donors were exposed to heavy metal isotope-labeled antibodies before processing by high-throughput mass spectrometry. This enabled measurement of the expression of 36 or more proteins, compared to the 10–11 proteins usually monitored by traditional flow cytometry methods. Using this method they were able to measure 17 surface markers (CD3, CCR7, CD11a, CD7, CD8, CD27, CD28, CD29, CD43, CD45RA, CD45RO, CD49d, CD57, CD62L, KLRG1, and HLA-DR), 10 intracellular species (IL-2, GM-CSF, MIP-1α, MIP-1β, granzyme B, CD69, perforin, TNF-α, IFN-γ, and CD107a/b) and three other

parameters (DNA content, cell length, and live/dead) within CD8+ T cells from different subsets of (classified by CCR7 and CD45RA expression) and with different antigen specificity, identified by labeling with MHC tetramers.

Principal component analysis was employed to look for association of 25 of the protein expression measurements with the established CD8+ T cell surface markers CCR7 and CD45RA (to mark na, central memory, and effector subsets) after stimulation with PMA and ionomycin. When plotted in the multivariate, principal component space, these 25 measurements were able to differentiate CD8+ T cells based on lineage across all six healthy donors. PCA of data from six donors indicated that the first two principal components accounted for 50 % of the variance, and the first three accounted for 60 %. Component 1 was most representative of naïve versus memory status of cells, component 2 separated based on differentiation status, and component 3 distinguished variation within the central memory subset. Overall, data from all six donors formed a "Y-shaped" pattern in the principal component space with naïve, central memory, and effector cells occupying distinct regions of the shape consistently across six different donors (Fig. 4). This pattern was not dependent on any one of the 25 parameters, as removing them individually had no effect on overall scores.

In a separate analysis, peptide-MHC tetramers were used to label T cells from these six donors according to antigen specificity for CMV, EBV, and flu (Newell et al. 2012). The same 25 protein expression measurements were made after treatment with PMA and ionomycin. These antigen-specific T cells were able to express 56-106 different combinations of the cytokines measured (compared to 512 possible combinations). In general, flu-specific cells tended to make TNF-α, but not MIP-1β compared to CMV- and EBV-specific cells. Also, CMV-specific cells were less likely to make IL-2 when compared with EBV- and flu-specific cells, while flu-specific cells were more likely to make GM-CSF. Despite the wide range of cytokines secreted, antigen-specific T cells occupied different areas in the principal components space, similar to the manner in T cells lineage subsets occupied different regions of the principal component space. One interesting result from the study of antigen-specific cells was that none of the antigen-specific cells occupied a multivariate space associated with a subset of central memory cells that were also CD49d-negative. Since CD49d has been identified as an integrin involved in cellular trafficking, one new hypothesis generated from this result was that antigen-specific CD8+ T cells all used the CD49d integrin for trafficking purposes, and therefore none were CD49d negative.

This work was crucial for robustly defining new systems of protein interactions that were highly relevant and specific to the function of different CD8+ T cell phenotypes, even without direct mechanistic insight. Past work has illustrated that it is difficult to discern robust functional differences due to donor-to-donor variability. The broad panel of proteins measured and multivariate approach used in this study was able to extract important functional differences despite normal variability between individual healthy donors. These 25 proteins now defined new boundaries of a system relevant to the study of CD8+ T cell phenotypes and can fuel additional study with follow-up experiments or knowledge-based modeling

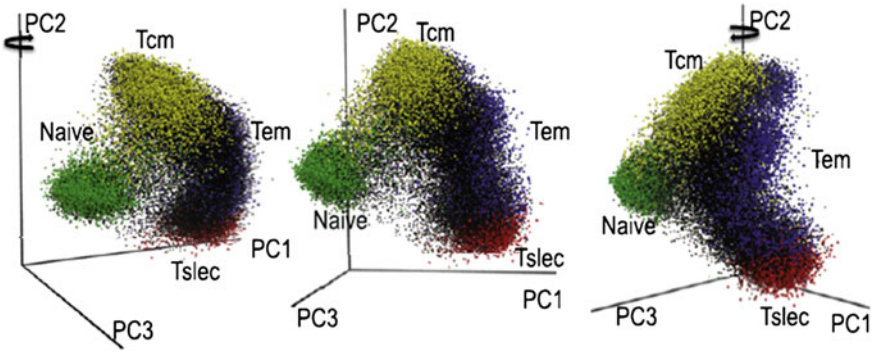


Fig. 4 PCA applied separately to data from three healthy donors indicated that 25 protein expression measurements were able to differentiate various CD8+ T cell subsets. T cell subset location in the multivariate space was similar for all three donors. Reprinted from Newell et al. (2012), with permission. Copyright 2012, Elsevier

approaches. Future work using this data might employ PLSDA to determine specific patterns of protein expression that best differentiate between antigen-specific cells and/or CD8+ T cell subsets. In addition, PLSDA with cross validation would allow for a more robust assessment of the predictive ability of these protein expression markers for distinguishing between different CD8+ T cell subsets.

3.2 Association of Single-Cell Dynamic Cytokine Secretion Events with T Cell Subsets

With experimental collaborators, our group has recently used PCA of high throughput, dynamic cytokine secretion measurements from CD3+ T cells to determine whether various cytokine secretion events can be associated with different T cell subsets (Han et al. 2012). Dynamic cytokine secretion measurements were obtained using a novel assay platform called "microengraving," where single CD3+ T cells were isolated from the peripheral blood of healthy donors into an array of subnanoliter wells such that most wells contained only one cell. Antibody-coated glass slides placed over the arrays were used to capture secreted cytokines from each well, and quantitative secretion rates were calculated from the total amount of cytokine captured over a given time period. Cells in the array were subsequently stained with fluorescent antibodies and imaged to determine the differentiation state based on surface markers CD3+, CD8+, CCR7, and CD45RA, and the viability of the cells. In this study, quantitative secretion rates of IFN-γ, IL-2, and TNF-α from single human CD3+ T cells were measured 2, 4, 6, 8, 10, 12, 14, and 16 h after stimulation with phorbol 12-myristate 13-acetate (PMA) and ionomycin. Results showed that cytokine secretion occurred in stochastic bursts, with the time of

initiation varying between cells, likely due to variation in the expression level of kinases, transcription factors, or slow epigenetic modifications.

Principal component analysis was performed on dynamic cytokine secretion rates at the eight time points for CD3+ CD8- T cells. Overall, analysis of dynamic data from eight time points was more effective for classifying T cell subsets (naïve, central memory, effector, effector memory) than PCA of time-integrated data over 6 h (41 ± 1 % misclassification error compared to 33 ± 1 % misclassification error) which would be similar to a the singular time point used when making traditional flow cytometry measurements. Interestingly, PCA was better able to classify T cell subsets by cytokine secretion when the secretion data were time-aligned (58 ± 4 %) versus unaligned [41 % misclassification error (Fig. 5)].

This work demonstrated that multivariate analysis of quantitative, dynamic single-cell cytokine secretion measurements may be more informative for the study of T cell subsets than traditional flow cytometry methods, since PCA of dynamic measurements was more effective in classifying T cell subsets. This work also indicated that IFN-γ, IL-2, and TNF-α were differentially secreted between T cell subsets, even though direct insight into the mechanisms governing their release were not obtained. The fact that the time-aligned secretion data better distinguished between subsets than the raw data also provided important new perspective on the dynamics of cytokine secretion: namely that it occurs in stochastic, fast bursts rather than sustained secretion that initiates at the same time among stimulated cells. This information on single-cell secretion activity would have been masked in a study of populations of cells and difficult to extract from this complex data set without multivariate analysis. Though it was not explored in this study, an interesting addendum would be to identify major differences in loadings of the principal components in the models used for classifying different T cell subsets. For example, which time points and cytokine secretion events were most important for discriminating between the different subsets? This information could provide new hypotheses for mechanistic differences between CD3+ T cell subsets and ideas for future data-driven and knowledge-based modeling. PLSDA would also provide insight into which linear combinations of secretion events/time points best discriminated between subsets.

3.3 Association of Immune Cell Protein Signaling with Donor Age and Race

Longo et al. (2012) used multivariate analysis to identify protein signaling events in various immune cell types that may be associated with age or race. Whole PBMCs from 60 healthy adult donors were stimulated for 15 min with one of 12 different immune modulators (IFN-α, IFN-Υ, IL-4, IL-10, B cell activator anti-IgD, IL-2, IL-6, IL-27, CD40L, R848, LPS, and PMA). Multi-parametric flow cytometry was used to quantify eight different phospho-signaling events (p-Stat,

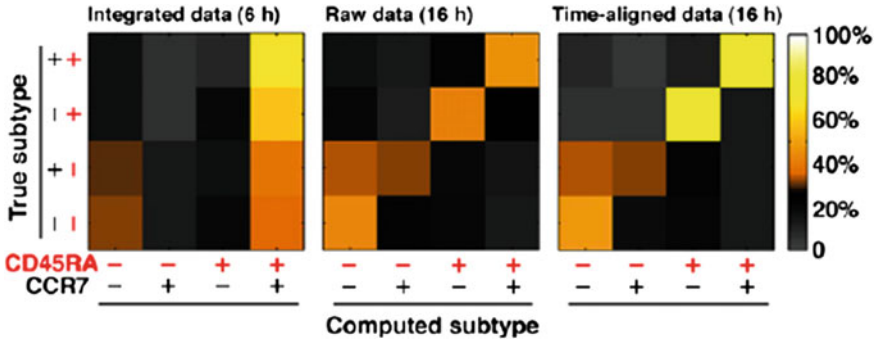


Fig. 5 PCA of quantitative cytokine secretion rates of IL-2, IFN-γ, and TNF-α measured at eight time points (time-aligned) were better able to classify CD3+ T cell subsets than secretion rate information integrated over 6 h (integrated data). PCA of time-aligned data was better at classifying subsets than PCA of raw data. Reprinted from Han et al. (2012), with permission

p-Stat3, p-Stat5, p-Stat6, p-Akt, p-S6, p-NF-κB, and p-Erk) in different subsets of immune cells, including all viable cells, monocytes, lymphocytes, B cells, CD3-CD20- lymphocytes, CD8+ T cells, CD4+ T cells, CD45RA- CD8+ T cells, CD45RA- CD4+ T cells, CD45RA+ CD8+ T cells, and CD45RA+ CD4+ T cells.

Univariate correlations between signaling events in different immune cell types were determined by calculating Pearson correlation coefficients for different signaling events in each cell type to create a basic map of immune function in healthy donors. In general, signals were positively correlated within each immune cell population rather than between different populations.

Principal component analysis was performed to identify signaling nodes that might be associated with age or race. Data from the 60 healthy donors were split evenly into training and test sets. Multi-linear regression was used to identify individual nodes associated with age or race before principal components analysis was performed to identify groups of associated signaling nodes. For age and race, separately PCA models were created using training data and then applied to test data. Inspection of the first principal component of the PCA age model showed many age-associated signaling nodes were within the T cell subset and one node was in the B cell subset. IFN-α activation of pStat5 in CD8+ CD45RA+ cytotoxic T cells, IL-2 activation of p-Stat5 in CD4+ CD45RA+ in T helper cells, IL-27 activation of p-Stat5 in CD8+ CD45RA+ cytotoxic T cells, and IL-4 activation of p-Stat6 were all associated with age in the model of the training data, and subsequently validated in the test data set. In the race-associated model, two signaling nodes were validated in the test set: anti-IgD/LPS stimulation of pAkt in B cells and anti-IgD/LPS stimulation of pS6 in B cells. Both were higher in European Americans compared to African Americans.

Overall, this work was able to identify signaling nodes in diverse cell types that may be associated with age or race, and created a functional map of signaling nodes induced by various stimuli in different immune cell types. The rich data set generated in this study begs for further multivariate analysis and data-driven

modeling. In addition to the PCA used to identify main signaling nodes driving variation within the data set, it would be informative and interesting to perform PLSDA to determine patterns of signaling nodes that best differentiate donors of different races. Likewise, it would be interesting to apply PLSR to determine patterns of cytokines most associated with age. PLSDA could be used to identify patterns of activated signaling nodes that best differentiate between immune cell types or between the different stimuli used. This type of analysis would enable the identification of patterns of signaling nodes (as opposed to individual signaling nodes) that best differentiate classes of immune cell types, and would generate new hypotheses for follow-up mechanistic studies. Also exciting would be potential follow-up experiments suggested by the authors that include measurements made at different time points. Additional experiments at longer time scales would enable the interaction of various immune cell types over time, allowing the exploration of how immune cell–cell interactions may evolve over time in diverse cell types.

4 Influence Insights

Diagrams illustrating influence and connectivity between species have been traditionally created using intuition and prior knowledge from experiments conducted in different settings. Here, we describe how recent work has employed various data-driven modeling techniques to methodically determine network connectivity of intracellular protein signaling and cytokine expression events from high-throughput data.

4.1 New Influence Maps for Intracellular Protein Signaling in Human Primary Naive CD4+ T Cells

Traditional methods for creating connectivity maps of protein signaling pathways have involved intuitive reconstruction of collective results from separate studies. Sachs et al. (2005) illustrate how Bayesian inference algorithms can systematically generate influence maps from high-throughput protein signaling data without prior knowledge of a signaling system. In this study, flow cytometry was used to quantify 11 protein signaling events downstream of the receptors CD3, CD28, and LFA-1 in thousands of primary human CD4+ T cells after treatment with 9 different stimulatory or inhibitory agents (Fig. 6: Measured pRaf S259, pErk1/pErk2 T202/Y204, p38 T180/Y182, pJnk T183/Y185, pAkt S473, pMek1/pMek2 S221/S217, pCREB, pPKA, pCaMKII, cleaved caspase 10/2, pPLC-γ Y783, pPKC S660, PIP_2, and PIP_3 after activation with either α-CD3, α-CD28, ICAM-2, PMA, β2cAMP or inhibition with G06976, a PIP_2 inhibitor, U0126, and LY294002). The large data set generated was analyzed with a Bayesian network inference

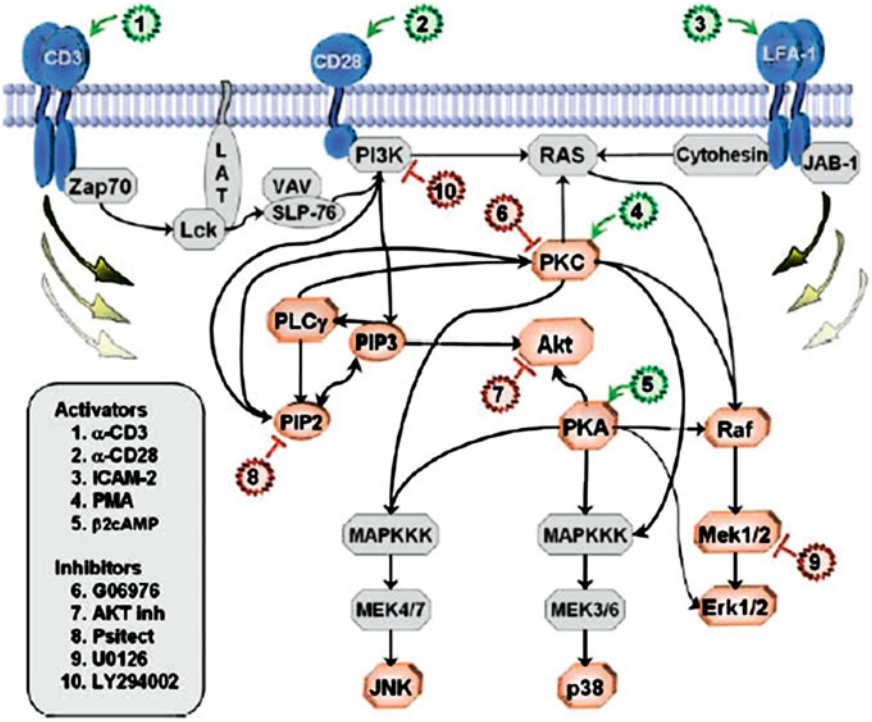


Fig. 6 A high-throughput proteomic data set was created by measuring protein phospho-signaling events (*orange*) after stimulation (*green*) or inhibition (*red*) of signaling nodes downstream of CD3, CD38, or LFA in naive CD4+ T cells. Reprinted from Sachs et al. (2005), with permission

algorithm to create a new signaling inference map (Fig. 7) with 17 high-confidence contributory arcs between the 11 species measured. Of the 17 influence arcs generated by the algorithm, 15 were "expected" (extensively described in the literature), 2 were "reported" (mentioned in some reports but not extensively described), and only 1 of the influence arcs were "reversed" (the opposite of literature reports). Three known influence arcs reported in the literature were missed by the model (Fig. 7).

The Bayesian algorithm was able to identify different types of relationships between protein signaling events. Several links identified by the model were direct enzyme-substrate relationships, including the phosphorylation of Raf by PKA and the phosphorylation of MEK by Raf. The model also allowed for the recognition of influence arcs involving protein species that were not directly perturbed in the study. Raf was not perturbed by any of the activators or inhibitors used, but the model still accurately deduced a causal influence arc from Raf to MEK. The model was also able to identify indirect influence connections involving species not measured by experiments in the study. For example, the model correctly deduced

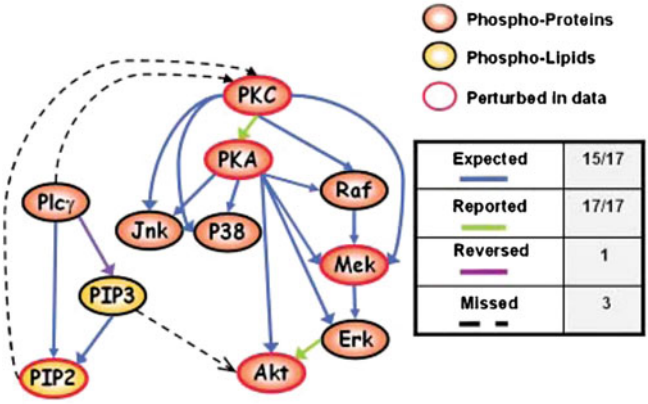


Fig. 7 Given high-throughput proteomic data, a Bayesian influence algorithm was used to create and influence map for all protein species measured in the study. Reprinted from Sachs et al. (2005), with permission

PKA and PKC activation of MAPK p38 and JNK, even though these activation steps utilized MAPK kinases that were not explicitly measured. The Bayesian algorithm was able to ignore network connections that were explained by other influence arcs. For example, although Erk activation is downstream of Raf, the model did not create an arc from Raf to Erk because Erk activation was mediated by the arcs from Raf to MEK and MEK to Erk.

One influence arc identified by the model but not reported in the literature for CD4+ T cells was the activation of Akt by ERK. The post-analysis literature searching revealed that this had been previously reported in colon cancer cells lines. To confirm model findings that this event was also important in CD4+ T cells, the authors inhibited Erk1 or Erk2 with small interfering RNA (siRNA) and measured levels of S473-phosphorylated Akt as well as PKA activity (which the model did not find to be influenced by Erk). As predicted by the model, p-Akt was reduced after siRNA inhibition of Erk1, but the activity of PKA was unchanged after inhibition of Erk.

The authors of this work highlighted how the high-throughput data used in this study were well-suited for the Bayesian inference technique used because of three specific attributes: firstly, the large size of the data set with measurements from thousands of cells permitted high-confidence inference of causal influence between signaling events, despite the inherent noise in biological data; secondly, flow cytometry measurements made in single cells avoided population averaging effects that occur when using methods such at Western blots; and finally, multiple stimulatory and inhibitory perturbations enabled better identification of protein–protein influence relationships. In order to highlight these points, the authors created three additional test data sets including (1) a data set without any interventional steps (1,200 points versus 5,400 in the original data set), (2) a population averaged data set without single-cell events, and 3) a much smaller data set, with

data randomly excluded, that was similar in size to data obtained from Western blots (only 420 points versus 5,400 in the original data set). Bayesian maps constructed from these sets were largely inferior to the influence map constructed from the original data set. For example, the map generated from the set without intervention only identified 10 undirected arcs, with eight arcs expected or reported and 10 arcs missing. The small, 420-point data set failed to infer many known associations, and identified many inexplicable, possibly incorrect connections compared to the full 5,400-point data set. The model generated from the population-averaged data set missed five influence arcs when compared to the model generated from the single-cell data set of the same size.

This work illustrated the ability of Bayesian network analysis to systematically generate new protein signaling influence diagrams from high-throughput protein signaling measurements. Although prior knowledge was used to identify protein signaling measurements that would give the most biological insight, the influence diagram was generated by the Bayesian network inference algorithm from the high-throughput data alone, independent of prior knowledge about the signaling system. Connections and relationships identified by this study were biologically relevant to CD4+ T cell signaling, because the high-throughput data used were itself biologically meaningful. The overall focus of this work was to identify broad influences and cross-talk among signaling pathways downstream of CD4+ T cell activation. Follow-up experiments using siRNA knockdowns confirmed the Bayesian inferences identifying Erk activation of Akt as an important new signaling relationship in CD4+ T cells.

4.2 Influence Maps of Cytokine Expression in CD4+ T Cells

Recently, decision tree analysis has been employed to explore the relationships between different cytokine expression events in activated CD4+ T cells (Simon et al. 2012). Mice were immunized with recombinant glucose-6-phosphate isomerase (G6PI) with Freund's complete adjuvant and CD4+ T cells were harvested from lymph nodes 21 days after immunization. Harvested cells were stimulated with G6PI for 6 h and cytokine secretion was blocked for the last 4 h. The expression of six cytokines (GM-CSF, TNF-α, RANKL, IL-2, IL-17, and IFN-γ) was measured with flow cytometry and a correlation matrix was generated to illustrate co-expression relationships between the different cytokines. A correlation matrix was not sufficient to accurately describe the combinatorial, hierarchical relationships between three or more cytokines.

To obtain combinatorial, hierarchical relationships between the cytokines, a decision tree algorithm was employed to explore relationships between MFI of each cytokine with those of the other five. A separate decision tree was generated for each cytokine with each node in the tree representing other cytokine expression events that were highly associated with expression of the cytokine of interest. The tree generated for IFN-γ expression indicated that TNF-α was the most important

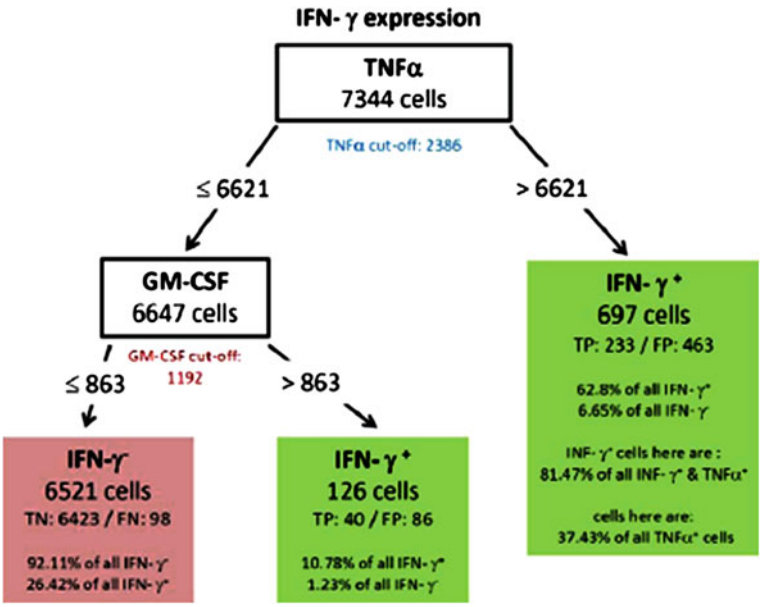


Fig. 8 A decision tree indicates the hierarchy of importance of different cytokine expression events in expression of IFN-γ. Of the five cytokines measured, TNF-α and GMCSF were most important for predicting expression of IFN-γ. Reprinted from Simon et al. (2012), with permission

of the five other cytokines measured for distinguishing IFN-γ-expressing cells because 62.8 % of these cells had a TNF-α MFI greater than 6621, while only 6.65 % of IFN-γ negative cells had a TNF-α MFI greater than 6621. GM-CSF was the next most important cytokine for distinguishing IFN-γ expression: 92 % of all IFN-γ-expressing cells also had a GM-CSF MFI less than 863 (Fig. 8). Similar analyses were done for all six cytokines. Other results for each cytokine are illustrated by decision trees found in the published work (Simon et al. 2012).

This work illustrated how decision trees can be used to map influence relationships for cytokine communication networks in immune cells. We are currently employing this methodology for understanding the response of different patient cohorts to viral infections in terms of the dynamics of multiple cytokine effects among various immune cell subpopulations applied to data sets similar to those described in a recent publication (Ndhlovu et al. 2012).

5 Insight into Mechanism

Insight gained from data-driven modeling techniques can go beyond association, yielding new systems-level mechanistic insight and identifying combinatorial events involved in immune system phenotype and function. Results can then be

used to guide focused literature searches and follow-up experiments with specific pharmacological stimulators and inhibitors to confirm model findings and identify new systems-level mechanisms involved in immune processes. Data-driven techniques are especially promising for identifying new mechanisms relevant to in vivo settings, where compensatory mechanisms may mask the importance of certain network signaling events. The approach also offers the opportunity to associate molecular-level protein signaling mechanisms with cell and tissue-level phenotype and function, which is often difficult to achieve using knowledge-based modeling.

5.1 Identification of a New Autocrine Cascade Involved in TNF-Induced Apoptosis

Tumor necrosis factor (TNF) is a cytokine capable of initiating both cell death and survival via different pathways. Experimental work has identified crosstalk between death and survival pathways, but systems-level interactions have not been well-characterized, and are especially complex when considering the presence of paracrine and autocrine signals. Using PLSR, Janes et al. (2006) were able to identify a novel TNF-induced autocrine cascade associated with TNF-induced cell death and survival.

HT-29 human colonic adenocarcinoma cells were treated with 10 combinations of TNF-α, insulin, or EGF at sub-saturating (low) or saturating (high) concentrations. Cell extracts were harvested at 13 time points, seven from 0 to 2 h and six from 4 to 24 h. Using Luminex-based assay, 19 quantitative protein activity measurements were made at each time point including IKK, JNK1, MK2, pY1068 EGF receptor, total EGF receptor, pS217/221 MEK, ERK, pIRS-1, pS473 Akt, Akt kinase activity, total Akt, phospho Forkhead transcription factor, and cleaved caspase 8. Discriminant partial least squares regression (DPLSR) was performed to identify signaling events that were quantitatively associated with different stimuli and the results were visualized in the principal component space (Fig. 10a). The first principal component differentiated all stimuli from mock conditions; the second and third principal components best differentiated between stimuli. Mapping the signals on the multivariate space revealed several unexpected interactions. Notably, EGFR and downstream signaling events (ERK, MEK) mapped equidistantly from TNF-α and EGF stimulation (Fig. 9a). A literature search revealed that TNF can induce the shedding of EGF ligands from the cell membrane into the surrounding media. ELISA measurements identified a TNF-induced increase in the production of the EGFR ligand TGF-α with fast kinetics, suggestive of a post-translational mechanism for production. Other experiments showed that TNF-induced TGF-α production was able to stimulate downstream EGFR signaling, since cells treated with TNF-α and an EGFR-blocking antibody reduced MEK ERK signaling.

Another unexpected interaction highlighted by the principal component score plot was that IKK activity was surprisingly distant from TNF-α, a previously identified inducer (Fig. 9a). Closer inspection of the signaling data revealed that, unlike the other known TNF-induced signals JNK1 and MK2, IKK was unique in that it had a second phase of activation that occurred 4–24 h after TNF stimulation. Careful literature searching revealed reports of late-phase NF$\kappa\beta$ activation in keratinocytes that was stimulated by autocrine discharge of IL-1α. Based on this knowledge, the authors designed additional experiments to confirm TNF-induced IL-α release, IL-1α activation of IKK, and the dependency of late-phase IKK activation on IL-1α. They were also able to use careful experimental design to show that the release of IL-1α was dependent on autocrine shedding of TGF-α. All together the work resulted in identification of a new autocrine cascade, critical to TNF-induced cell death (Fig. 9b).

Overall, this study illustrated how multivariate analysis can be coupled with focused literature searching and follow-up experiments to provide new insight into systems-level mechanisms governing the functions of a signaling network. In this study, the approach led to the discovery of a new autocrine cascade where TNF-α induces rapid TGF-α shedding and subsequent stimulation of EGFR and IL-1α (Fig. 9b). This work illustrates the power of the combination of data-driven modeling with careful experimental design. Experiments alone can be sufficient to identify and confirm new interactions; however, in this study the identification of broadened associations and linkages to network function were only possible with visualization of the complex signaling data set in a multivariate space.

5.2 Identification of New Mechanisms for Regional and Temporal Variation of the Apoptotic Response to Inflammatory Cytokines in the Small Intestine

Data-driven modeling techniques can be especially useful for discerning important multivariate correlations in vivo systems where compensatory mechanisms may mask important mechanistic events.

To gain systems-level insight relevant to the pathogenesis of inflammatory bowel disease, Lau et al. (2011) used a multivariate approach to identify new protein signaling mechanisms that accounted for regional differences in the apoptotic response of the mouse small intestine to inflammatory cytokines. Antibody-based TNF-α inhibition has been used clinically to treat chronic inflammation associated with inflammatory bowel disease, supporting the concept that TNF signaling is critical to the function of the small intestine and the disease phenotype. However, specific relationships between complex TNF-α signaling and tissue inflammatory phenotype are not completely understood, and may vary in different cell types and tissue microenvironments within the small intestine.

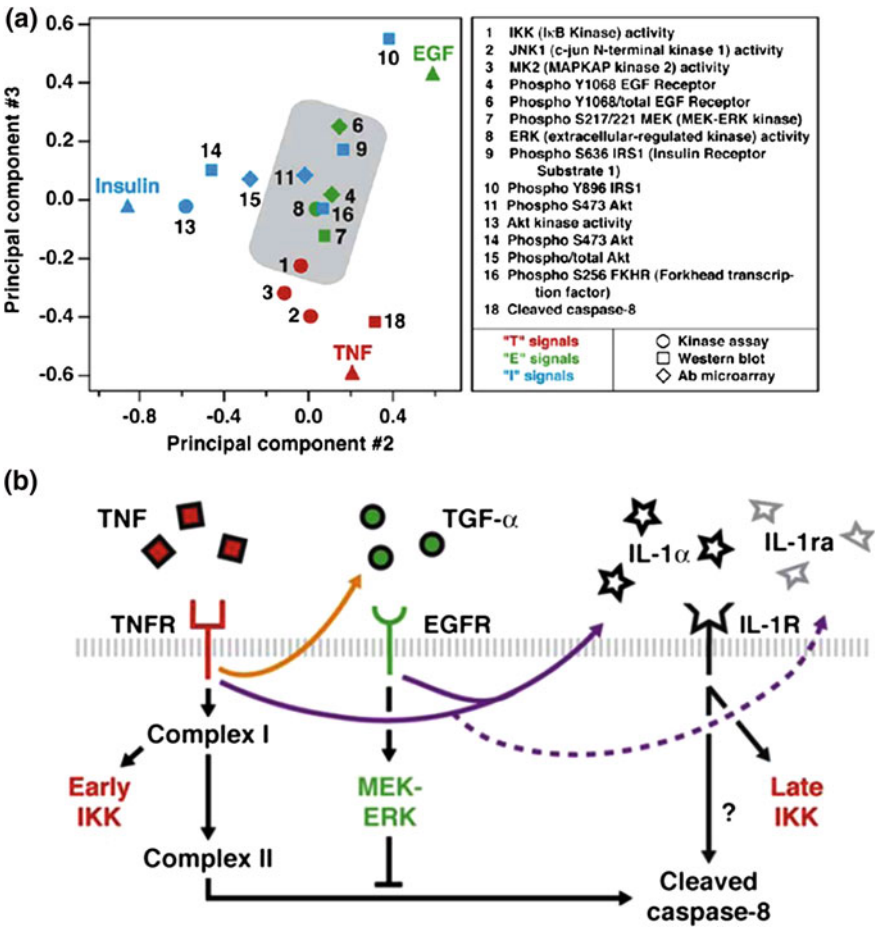


Fig. 9 **a** DPLSR of protein signaling events after stimulation with insulin, EGF, and TNF-α revealed that surprisingly, EGFR ligands ERK, and MEK (labeled *7* and *8*) mapped equidistant from TNF-α and EGF stimulation and IKK (labeled *1*) mapped surprisingly distant from its known inducer TNF-α. **b** Focused literature searching and follow-up experiments resulted in discovery of a novel autocrine cascade mechanism involving TNF-induced shedding of TGF-α ligand, TGF-α stimulation of EGFR, and subsequent activation of both IL-1α and IL-1ra to promote or inhibit cell death, respectively. Reprinted from Janes et al. (2006), with permission

In this study, healthy mice were injected intravenously with a 5 μg bolus of recombinant TNF-α to induce phenotypic changes in the small intestine that were similar to inflammatory bowel disease. Immunohistochemical staining of cleaved caspase 3 and phosphorylated histone 3 (pH3) revealed that the response that was induced varied along the length of the intestine, with apoptosis being more prevalent in the duodenum and proliferation more prevalent in the ileum. A time course of TNF-induced apoptosis indicated that the ileum was inherently unaffected by TNF-α, whereas the time course of TNF-induced apoptotic effects in

the duodenum indicated a dose-dependent response. In the duodenum, a high TNF dose resulted in an abrupt, early peak in apoptosis, whereas a low dose resulted in a delayed and more gradual peak. Western blot and immunohistochemical analysis suggested that increased expression of TNFR1 in the duodenum could be responsible for the differential apoptotic responses. However, work using TNFR knock-out animals revealed that there was no direct relationship between TNFR1, TNFR2, and the different TNF-induced apoptotic effects in the duodenum and the ileum.

To determine whether differences in protein signaling could be associated with phenotypic changes in the small intestine, a Luminex-based assay was used to quantitatively measure 14 phospho-protein signaling events (including I$\kappa\beta\alpha$, RSK, Stat3 s, JNK, MEK, p38, Akt, c-Jun, Stat3Y, S6, ERK1/2, and ATF2) in the duodenum and ileum of 55 healthy mice at different time points after infusion of low (5 μg) and high (10 μg) doses of TNF. Hierarchical clustering organized the data set into groupings by the spatial regions of the small intestine (duodenum and ileum). PLSDA was used to determine whether linear combinations of different protein signaling events were able to predict the TNF-induced apoptotic phenotype of tissues. Samples were classified into three groups: (1) no apoptosis and proliferation (ileum after systemic TNF dose), (2) late apoptosis and arrest (duodenum after a systemic TNF dose of 5 μg/ml), and (3) early apoptosis (duodenum after TNF dose of 10 μg/ml). PLSDA was used to determine which of the measured protein signaling events best classified the data based on phenotypic responses of the small intestine. PLSDA successfully classified these groups, with latent variable 1 best separating the data based on spatial location within the small intestine (ileum versus duodenum) and latent variable 2 best separating the early/late apoptosis data from the duodenum in response to low versus high TNF dose (Fig. 10a). Key loadings in the first latent variable were transient phosphorylation of MEK, ERK, and p38 which were associated with resistance to proliferative arrest in the ileum; ATF2 and c-Jun were associated with apoptosis in the duodenum (Fig. 10b). In contrast, late phosphorylation of ERK, MEK, and RSK were associated with the early apoptotic peak in the duodenum. Altogether, this information suggested the presence of two different stages of MAPK activation that occur in the small intestine after systemic TNF stimulation, with early MEK and ERK responsible for the maintenance of proliferation and resistance to apoptotic arrest in the ileum. This early activation did not occur in the duodenum, resulting in apoptosis. After a higher dose of TNF in the duodenum, MEK-ERK-RSK signaling was necessary for preserving homeostasis. Thus, MAPK signaling may play different roles in the small intestine in different contexts. To test these hypotheses, the authors performed additional experiments treating animals with a MEK inhibitor (PD325901) 2 h before TNF injection. When the PLSDA model was applied to this data set, MEK-inhibited duodenal samples that received the lower dose of TNF were classified as duodenal samples that received the higher dose, suggesting that MEK inhibition shifted the apoptotic peak of these samples from late to early (Fig. 10c and d). Altogether, multivariate analysis indicated that ERK signaling was a key mediator in the resistance of the ileum to proliferative

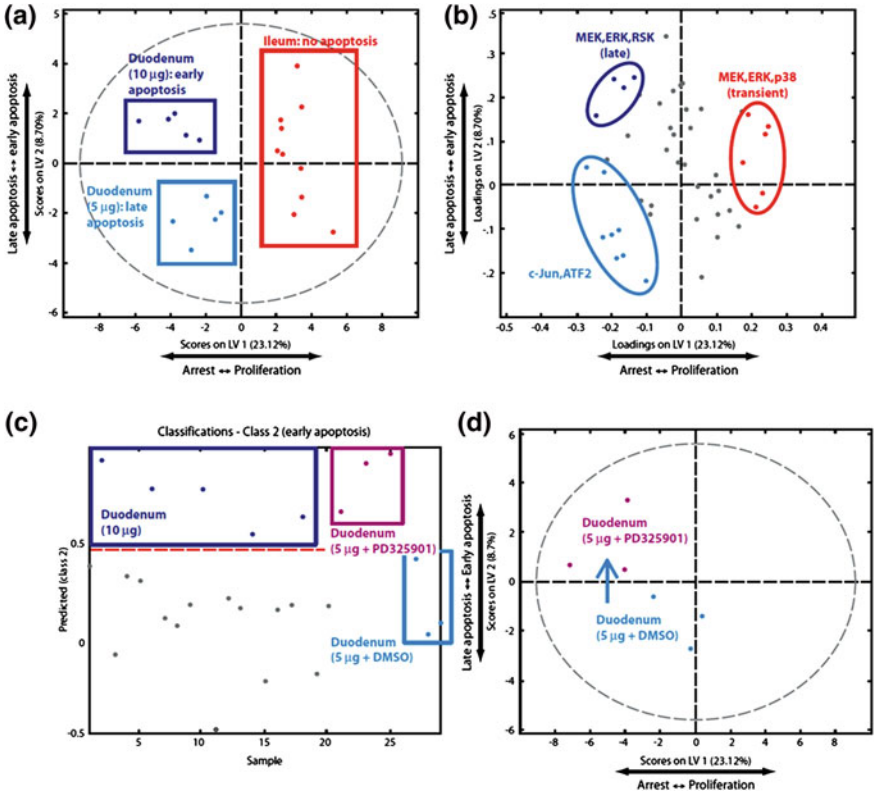


Fig. 10 **a** PLSDA of protein signaling in the mouse small intestine after systemic TNF infusion revealed spatial (duodenum versus ileum) and dose-dependent (10 versus 5 µg) differences in apoptotic effects that were associated with MEK signaling (**b**). Follow-up experiments with a MEK inhibitor (PD325901) confirmed the finding, as low-dose TNF administered with a MEK inhibitor induced duodenum responses similar to those induced by high-dose TNF (**c** and **d**). Reprinted from Lau et al. (2011), with permission

arrest, and the duodenum's early apoptotic response to low doses of TNF-α. Treatment with a MEK inhibitor confirmed this finding, shifting ileum responses of the 14 signaling events toward the early apoptotic duodenum response region in the multivariate space. Likewise, MEK inhibition resulted in a shift of the early apoptotic response of the duodenum to low dose TNF-α to the late apoptosis region of the control duodenum response to high-dose TNF-α (Fig. 10c and d).

To illustrate the power of the multivariate approach, Lau et al. generated an artificial MEK inhibition data set that would be representative of a reductionist, univariate approach. In this data set, all signaling measurements were kept the same as they were measured in the data generated without MEK inhibition, with the exception of Erk phosphorylation which was set to zero. PLSDA of this artificial data set, and visualization in the multivariate space, was not able to capture the shifts in duodenal responses that were captured in the actual data set.

This work demonstrated the value of a multivariate approach by comparing multivariate results to results that would have been obtained from a more traditional, univariate reductionist approach. When compared with the multivariate approach, the univariate approach was unable to capture all differences in small intestine TNF-α induced, MEK inhibited signaling responses, and small intestine TNF-α induced responses. This highlights the importance of data-driven modeling approaches for the discovery of new, systems-level signaling mechanisms relevant to in vivo tissue phenotype and function. As illustrated in this study, data-driven approaches may be especially relevant to signaling studies in vivo settings where compensatory signaling mechanisms often preclude the detection of network perturbations induced by disease or during pharmacologic intervention in univariate data sets. A multivariate approach offers the additional advantage of allowing for the association of molecular-level protein signaling events with cell phenotype and tissue function, something that is often difficult or impossible to do with knowledge-driven and/or experimental approaches alone.

6 Combining Data-Driven and Theory-Driven Approaches

Finally, we present an unusual example where data-driven modeling is used in combination with knowledge-based modeling to take synergistic advantage of the broad organizational and statistical power of the former and the mechanistic specificity of the latter.

6.1 Decision Tree Analysis for Evaluating the Effects of Initial Conditions on an ODE Model of FasL Induced Apoptosis

The combination of data-driven and knowledge-driven modeling approaches can be yield a powerful method for determining the importance of initial conditions in the behavior of a signaling pathway. Hua et al. (2006) used combined mechanistic and data-driven modeling approaches to study the Fas pathway and its regulation of cell death, which has important implications in cancer and autoimmune diseases. Here, data-driven approaches were employed to determine the effect of the initial conditions of species on the output of an ODE-based model. A simplified, mechanistic knowledge-based ODE model from a previous study (Hua et al. 2005) was used to predict caspase 3 activation resulting from signaling events downstream of FasL stimulation (Fig. 11). In order to determine the effect of initial conditions on FasL signaling pathways, the values of nine key initial conditions were shifted either 10-fold higher or lower than baseline using a Monte Carlo algorithm. One million different modeling simulations were performed with the various initial conditions and the output of each simulation expressed as fold-change in activated caspase 3 over baseline conditions.

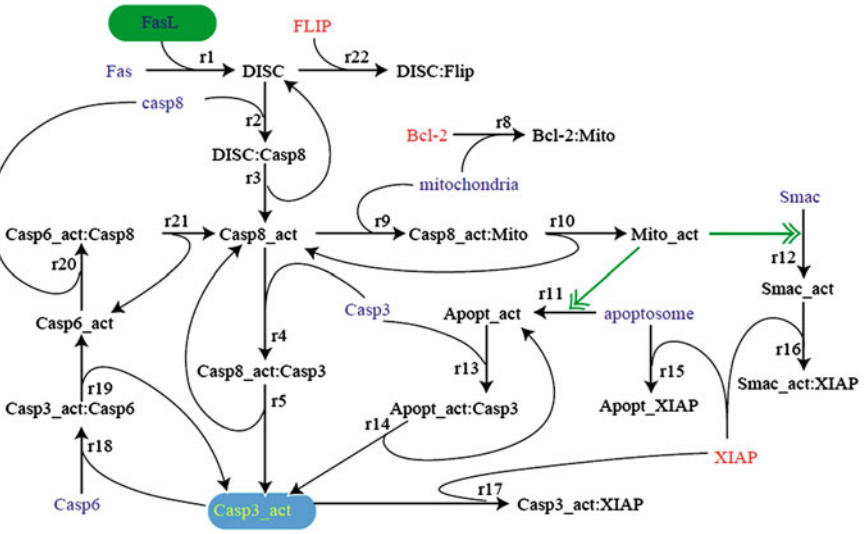


Fig. 11 A FasL signaling modeled with ODEs, in which caspase 3 activation is induced by Fas ligand. Reprinted from Hua et al. (2005), with permission

The simulation results varied considerably and depended on the initial conditions. In order to simplify the analysis, a *k*-means clustering algorithm classified the output as: (1) high FasL sensitivity (a fast increase in cleaved caspase-3 with addition of FasL), (2) medium FasL sensitivity, and (3) low FasL sensitivity (slow or no increase in cleaved caspace-3 output with addition of FasL). A decision tree algorithm determined the role of initial conditions (nodes) in high (III), medium (II), or low (I) FasL sensitivity (clusters) and the resulting tree gave a hierarchy of importance of nodes for each of the FasL sensitivity clusters (Fig. 12).

The decision tree identified XIAP and Fas initial conditions as the most important for determining the sensitivity of caspase 3 activation to FasL, and demonstrated that the importance of the initial condition of one species is highly dependent on the initial concentrations of other species. For example, the tree specified a range of 3.1to 5.2-fold increase from baseline of XIAP initial conditions, over which the Smac initial concentration affected the FasL sensitivity of the system (Fig. 12). When the initial concentration of XIAP was greater than 5.2-fold higher than baseline, the system was always insensitive to Fas regardless of the initial conditions of other species or Smac, because there was always excess XIAP to bind cleaved caspase-3. When XIAP was below 3.1-fold over baseline, the initial concentration of Smac also did not matter because XIAP levels were so low that they were not able to sequester caspase-3 and caspace-9, regardless of Smac concentration (Fig. 12). A univariate sensitivity analysis was applied to the ODE model and identified Fas and Flip as the two species that the model output was most sensitive to. Interestingly, Fas and Flip also appeared important for many of the outcomes of the decision tree, confirming the concept that their initial

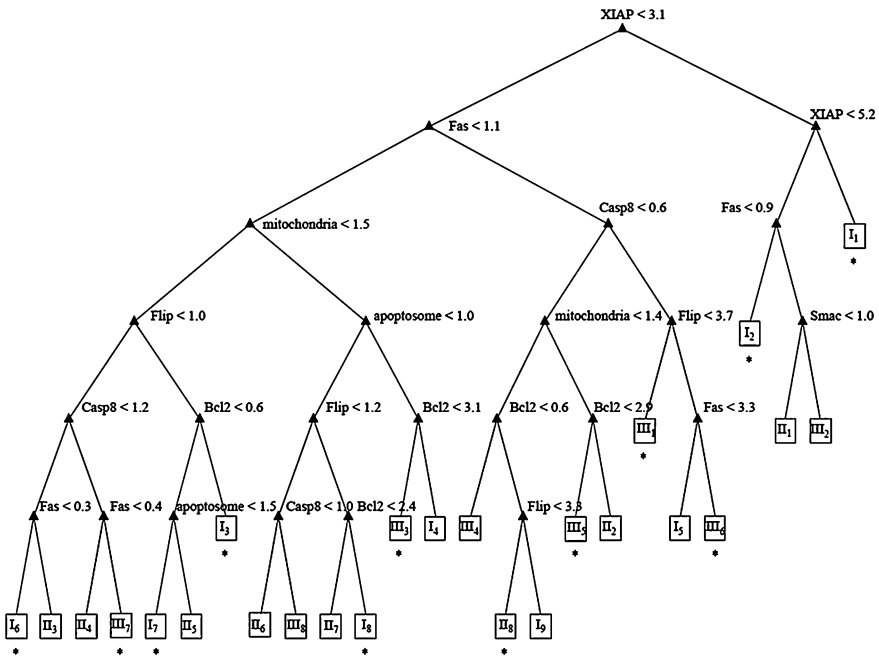


Fig. 12 A decision tree indicates the hierarchy of importance of various initial conditions in network low (*I*), medium (*II*) and high (*III*) FasL sensitivity (*square clusters*) based on fold change from baseline from initial conditions of various model species (represented at nodes), including XIAP, Fas, Flip, Smac, Bcl2, and Casp 8. To read the tree: if statement at a node is true, proceed right on the tree. Reprinted from Hua et al. (2005), with permission

set-points were important for the sensitivity of the output to initial conditions of many of the other species.

The decision tree in this study was validated with a test set that was created by running the ODE-simulation 1,000 times more, then tested against the generated decision tree to determine how well it was able to predict an independent data set. Overall, the decision tree was able to correctly predict the sensitivity outcome based on initial conditions for 71 % of the test data, much higher when compared to the 33 % prediction accuracy that would be expected from random chance prediction of three different clusters.

One exciting aspect of this study was an analysis that was done to determine the minimum number of changes in initial conditions that would be necessary to switch the sensitivity of the Fas pathway from one direction to another. The authors defined the total number of species that would need to be modified to change a cellular behavior as the COST. They computed the COST for transition between different sensitivity states (clusters of the tree), and reported them (Table 2). Their findings suggested the most efficient way to switch a cell from one state of Fas sensitivity to another. One example highlighted was the switching of a cell from a Fas-insensitive to a sensitive response. COST analysis suggested the

Table 2 *COST matrix* A COST matrix indicates the most efficient ways to transfer between low and high FasL sensitivity by altering initial conditions

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9
III_1	2.18	2.67	1.92	1.67	1.00	1.77	1.92	1.94	1.95
III_2	2.01	1.52	2.50	2.50	1.52	2.51	2.49	2.50	1.52
III_3	2.98	2.49	2.00	1.00	2.50	1.87	1.92	1.88	2.99
III_4	3.31	3.79	2.65	3.61	2.18	1.96	1.65	3.60	1.00
III_5	3.07	3.55	3.07	2.61	1.95	2.61	2.60	2.48	1.46
III_6	2.98	3.22	1.99	2.22	1.00	2.77	1.99	1.96	1.59
III_7	3.39	3.23	2.16	2.67	2.76	2.00	2.17	3.17	3.00
III_8	3.64	3.15	3.01	2.03	3.30	2.47	2.65	1.54	4.51

For example, the average number of species that would need to be modified to transition from cluster I_1 (low Fas sensitivity) to III_1 (high Fas sensitivity) would be 2.18. Reprinted from Hua et al. (2005), with permission

most efficient way to perform the switch would be to start with leaf I_6, and increase the Fas and caspase-8 initial conditions to transition it to the sensitive response in leaf III_1 with a cost of 1.77 (Table 2 and Fig. 12). The authors also highlighted that overall, the cost for transition from one state to the other was greater than one, indicating that more than one species must be altered in parallel to change the sensitivity of the network.

7 Looking Forward

We hope that the examples we have discussed here offer readers a stimulating foundation for the kinds of problems that can be readily addressed using the various established data-driven modeling approaches. As we move forward in understanding the human immune system at multiple scales, one exciting prospect offered by proteomic data-driven modeling is the ability to broadly characterize cytokine microenvironments and relate them to important immune system phenotypes and disease states. This would represent a "top-down" approach to evaluating immune cell–cell interactions, in contrast to traditional work that typically takes a "bottom-up" approach, focusing on detailed protein signaling events in single immune cells or homogenous populations of a single cell type. In a bottom-up approach, system behavior is intuitively or computationally reconstructed based on knowledge of individual events within cells. Data-driven modeling presents the opportunity to characterize the immune cell–cell interactions with a "top-down" approach by associating patterns of cytokine secretion with an entire system of interacting immune cell types without requiring detailed knowledge of mechanisms governing individual cell signaling events, the specific mechanisms of cell–cell interaction, or even the cytokine secretion events associated with each cell type. This could be applied to cultured whole peripheral blood mononuclear cells (PBMCs) or to environments created in different tissue types. Though the approach lacks specific mechanistic insight into the roles of individual cell types or signaling events, it may identify new, robust systems-level

behavior and provide a unique perspective that is relevant to in vivo immune system function. A top-down approach could be especially useful for the development of new vaccine strategies that alter cytokine microenvironments, in contrast to current strategies that target one immune cell type.

With respect to the development of modeling methodology and implementation in the coming years, two clearly beckoning challenges can be easily identified. Both challenges can be characterized as addressing integration in specific dimensions: in the first case, "horizontal integration", moving from the study of individual components to the study of multiple components concomitantly, and in the second case "vertical integration", that of moving from the study of system operation (phenotype, essentially) from the simplest contexts at the smallest space- and time-scales to more complex contexts involving larger space- and/or time-scales (Lauffenburger 2012).

The first important challenge is to demonstrate computational modeling frameworks for integrating diverse data types—e.g., gene sequence, gene expression, gene knockout/knockdown, protein expression, protein activities, and cell behavior. Two promising methods have been reported recently in application to the integration of proteomic and transcriptomic (Lan et al. 2011) data ['Prize Collecting Steiner Tree' (Huang and Fraenkel 2009)], and of transcriptomic and gene knockout data ['ResponseNet' (Lan et al. 2011)] in yeast. While both are formally data-driven methods, they require substantial prior knowledge in terms of protein–protein and protein–DNA interaction databases for mapping the "omic" measurements in a manner permitting computational optimization of information flow. For mammalian biology applications, such as the immune system, the state of this type of knowledge is quite nascent—although attempts are arising to strengthen this necessary foundation (Kirouac et al. 2012).

A second important challenge is to demonstrate computational modeling frameworks for integrating data from diverse spatial scales—e.g., how properties of multiple molecular components affect cell functions, how properties of multiple cell types affect tissue functions, how properties of multiple tissue types affect organism behavior, and how population behavior arises from multiple animals, subjects, and/or patients. We refer readers to an excellent review of multiscale modeling in the immune system (Chavali et al. 2008) for background, and note that this field has to date emphasized theory-based, or knowledge-driven frameworks in which hypotheses are postulated from the previous literature and implemented into simulation calculations. This emphasis, of course, restricts the scope of problems that can be addressed. Accordingly, we urge here incorporation of data-driven models at one or more of the scales as is likely necessary to comprehend horizontally integrated, multivariate data in the context of vertically integrated multi-scale models.

The third dimension of integration appreciated is "dynamic integration"—how to concomitantly use "static" properties such as sequence and expression levels in concert with "dynamic" information such as kinetics, mechanics, and transport phenomena of molecular and cellular processes. We can envision the use of data-driven modeling to gain insights concerning how dynamic information depends on

static properties. An interesting example is illustrated by a study that elucidates how protein sequence data can be employed to understand molecular transport properties in nuclear pore complexes (Colwell et al. 2010).

Taken together, we anticipate that a central coin-of-the-realm will be "hybrid models" comprised of diverse mathematical frameworks within an overall computational process. This situation would clearly recognize that for the foreseeable future biological knowledge will be sufficient for models constructed on the basis of prior understanding in only certain realms, while in other realms biological knowledge will remain inadequate for relying solely on that more traditional approach and will require formulation in terms of data-driven models.

Acknowledgments Many thanks to Gregory Szeto, Brian Joughin, and Alexandra Hill for assistance in editing the manuscript. This work was partially supported by the Ragon Institute of MGH, MIT, and Harvard, NIH grant U19 AI 089992, and NIH grant TR01- EB010246.

References

Benedict KF, Mac Gabhann F, Amanfu RK, Chavali AK, Gianchandani EP, Glaw LS, Oberhardt MA, Thorne BC, Yang JH, Papin JA, Peirce SM, Saucerman JJ, Skalak TC (2011) Systems analysis of small signaling modules relevant to eight human diseases. Ann Biomed Eng 39:621–635

Busse D, de la Rosa M, Hobiger K, Thurley K, Flossdorf M, Scheffold A, Hofer T (2010) Competing feedback loops shape IL-2 signaling between helper and regulatory T lymphocytes in cellular microenvironments. Proc Natl Acad Sci U S A 107:3058–3063

Chavali AK, Gianchandani EP, Tung KS, Lawrence MB, Peirce SM, Papin JA (2008) Characterizing emergent properties of immunological systems with multi-cellular rule-based computational modeling. Trends Immunol 29:589–599

Colwell LJ, Brenner MP, Ribbeck K (2010) Charge as a selection criterion for translocation through the nuclear pore complex. PLoS Comput Biol 6:e1000747

Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7:601–620

Geurts P, Irrthum A, Wehenkel L (2009) Supervised learning with decision tree-based methods in computational and systems biology. Mol BioSyst 5:1593–1605

Han Q, Bagheri N, Bradshaw EM, Hafler DA, Lauffenburger DA, Love JC (2012) Polyfunctional responses by human T cells result from sequential release of cytokines. Proc Natl Acad Sci U S A 109:1607–1612

Hoffmann A, Levchenko A, Scott ML, Baltimore D (2002) The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. Science 298:1241–1245

Hosseini I, Gabhann FM (2012) Multi-scale modeling of HIV infection in vitro and APOBEC3G-based anti-retroviral therapy. PLoS Comput Biol 8:e1002371

Hua F, Cornejo MG, Cardone MH, Stokes CL, Lauffenburger DA (2005) Effects of Bcl-2 levels on Fas signaling-induced caspase-3 activation: molecular genetic tests of computational model predictions. J Immunol 175:985–995

Hua F, Hautaniemi S, Yokoo R, Lauffenburger DA (2006) Integrated mechanistic and data-driven modelling for multivariate analysis of signalling pathways. J R Soc Interface 3:515–526

Huang SS, Fraenkel E (2009) Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Sci Signal 2:40

Janes KA, Yaffe MB (2006) Data-driven modelling of signal-transduction networks. Nat Rev Mol Cell Biol 7:820–828

Janes KA, Gaudet S, Albeck JG, Nielsen UB, Lauffenburger DA, Sorger PK (2006) The response of human epithelial cells to TNF involves an inducible autocrine cascade. Cell 124: 1225–1239

Kingsford C, Salzberg SL (2008) What are decision trees? Nat Biotechnol 26:1011–1013

Kirouac DC, Saez-Rodriguez J, Swantek J, Burke JM, Lauffenburger DA, Sorger PK (2012) Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. BMC Syst Biol 6:29

Lan A, Smoly IY, Rapaport G, Lindquist S, Fraenkel E, Yeger-Lotem E (2011) ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. Nucleic Acids Res 39:W424–W429

Lau KS, Juchheim AM, Cavaliere KR, Philips SR, Lauffenburger DA, Haigis KM (2011) In vivo systems analysis identifies spatial and temporal aspects of the modulation of TNF-alpha-induced apoptosis and proliferation by MAPKs. Sci Signal 4:ra16

Lauffenburger DA (2012) The multiple dimensions of Integrative Biology. Integr Biol (Camb) 4:9

Longo DM, Louie B, Putta S, Evensen E, Ptacek J, Cordeiro J, Wang E, Pos Z, Hawtin RE, Marincola FM, Cesano A (2012) Single-cell network profiling of peripheral blood mononuclear cells from healthy donors reveals age- and race-associated differences in immune signaling pathway activation. J Immunol 188:1717–1725

Martens H, Martens M (2001) Multivariate analysis of quality: an introduction. Wiley, Chichester

Nakaya HI, Wrammert J, Lee EK, Racioppi L, Marie-Kunze S, Haining WN, Means AR, Kasturi SP, Khan N, Li GM, McCausland M, Kanchan V, Kokko KE, Li S, Elbein R, Mehta AK, Aderem A, Subbarao K, Ahmed R, Pulendran B (2011) Systems biology of vaccination for seasonal influenza in humans. Nat Immunol 12:786–795

Ndhlovu ZM, Proudfoot J, Cesa K, Alvino DM, McMullen A, Vine S, Stampouloglou E, Piechocka-Trocha A, Walker BD, Pereyra F (2012) Elite controllers with low to absent effector CD8+ T cell responses maintain highly functional, broadly directed central memory responses. J Virol

Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM (2012) Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. Immunity 36:142–152

Palmer MJ, Mahajan VS, Trajman LC, Irvine DJ, Lauffenburger DA, Chen J (2008) Interleukin-7 receptor signaling network: an integrated systems perspective. Cell Mol Immunol 5:79–89

Pe'er D (2005) Bayesian network analysis of signaling networks: a primer. Sci STKE 2005:l4

Prilutsky D, Shneider E, Shefer A, Rogachev B, Lobel L, Last M, Marks RS (2011) Differentiation between viral and bacterial acute infections using chemiluminescent signatures of circulating phagocytes. Anal Chem 83:4258–4265

Querec TD, Akondy RS, Lee EK, Cao W, Nakaya HI, Teuwen D, Pirani A, Gernert K, Deng J, Marzolf B, Kennedy K, Wu H, Bennouna S, Oluoch H, Miller J, Vencio RZ, Mulligan M, Aderem A, Ahmed R, Pulendran B (2009) Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol 10:116–125

Rivet CA, Hill AS, Lu H, Kemp ML (2011) Predicting cytotoxic T-cell age from multivariate analysis of static and dynamic biomarkers. Mol Cell Proteomics 10:M110 003921

Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308:523–529

Simon S, Guthke R, Kamradt T, Frey O (2012) Multivariate analysis of flow cytometric data using decision trees. Front Microbiol 3:114

Critical Dynamics in Host–Pathogen Systems

Arndt G. Benecke

Abstract Host–pathogen interactions provide a fascinating example of two or more active genomes directly exerting mutual influence upon each other. These encounters can lead to multiple outcomes from symbiotic homeostasis to mutual annihilation, undergo multiple cycles of latency and lysogeny, and lead to coevolution of the interacting genomes. Such systems pose numerous challenges but also some advantages to modeling, especially in terms of functional, mathematical genome representations. The main challenges for the modeling process start with the conceptual definition of a genome for instance in the case of host-integrated viral genomes. Furthermore, hardly understood influences of the activity of either genome on the other(s) via direct and indirect mechanisms amplify the needs for a coherent description of genome activity. Finally, genetic and local environmental heterogeneities in both the host's cellular and the pathogen populations need to be considered in multiscale modeling efforts. We will review here two prominent examples of host–pathogen interactions at the genome level, discuss the current modeling efforts and their shortcomings, and explore novel ideas of representing active genomes which promise being particularly adapted to dealing with the modeling challenges posed by host–pathogen interactions.

Contents

A. G. Benecke (✉)
Centre National de la Recherche Scientifique, Institut des Hautes Études Scientifiques, 35 route de Chartres, 91440 Bures sur Yvette, France
e-mail: arndt@ihes.fr

Current Topics in Microbiology and Immunology (2013) 363: 235–259
DOI: 10.1007/82_2012_260

Published Online: 12 September 2012

1 A Systems Biology Challenge: Multiscale Integration

After having generated high hopes and even more massif parallel data, *systems biology* is clearly on the verge of entering into a new phase to fulfill on the initial promise of revolutionizing not only the way we do biology but also our understanding of biologic phenomena (Tisoncik and Katze 2010). Success of this new phase will depend on solving some fundamental problems which so far have not, or only superficially been addressed, and will require more than ever a concerted and integrated effort spanning the entire spectrum of exact sciences.

The central problem we need to address is the integration of data and insights over multiple scales as to be able to make meaningful predictions about how complex traits and phenotypes emerge from assemblies of objects and the molecular mechanisms linking these objects on the one hand, and on the other, to be able to decompose phenotypes rapidly to understand the defining dynamics and their molecular basis. The former, inference-based analysis thereby actually encompasses also evolutionary questions, as most of the biologic systems we try to understand and describe are remarkably robust despite stochasticity being present, if not integral part of the mechanisms at multiple levels. The latter challenge of decomposition is still the main bottleneck on the road of designing therapeutical and vaccination strategies in biomedical research.

> Decomposition and inference across time and space scales define the ultimate paradigm of systems biology research in as much as, if achieved and abstracted, the combination of both would lead to meaningful mapping functions from the object space to the phenotype space (Φ) and back (f) (Fig. 1).

The problem of integration over multiple scales is not unique to biology but also a major issue in physics and chemistry or social and economic sciences (Lesne and Lagües 2012). The problem, however, is particularly hard in biology, as the

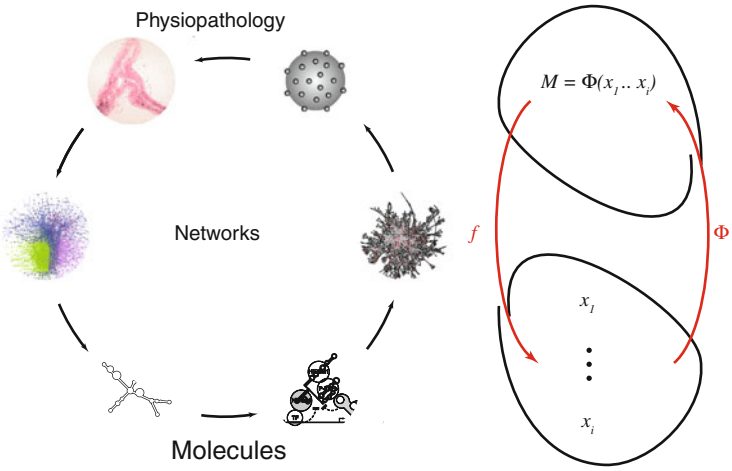


Fig. 1 Systems Biology Life Cycle: Decomposition of complex traits and phenotypes to understand the systems dynamics and the defining molecular objects and the mechanisms by which they interact; inference to make meaningful predictions as to how different objects interact to give rise to phenotypes and traits. Both processes will heavily rely on the identification and analysis of different biologic networks at different scales. The integration of information, objects, and their dynamics across scales represents the main challenge of systems biology today. Successful integration is the *sine qua non* requirement to identify and formulate the mapping functions ϕ and f from object space to phenotype space and back. Having a full set of these transforming functions would elevate the need to measure all objects and describe all possible phenotypes, and thus represent understanding of the system

integration has to be bi- rather than unidirectional. Consider a dune, thus a physical object—the dune's physical properties depend entirely on the physical properties of the sand-corn. Using renormalization techniques, it is possible to mathematically describe a dune and investigate its properties under changing conditions (wind, humidity), without considering each sand-corn individually with simple equations such as the original Bagnold formula (Bagnold 1936). In biology, the physical properties of the molecular assembly such as a chromatin fiber will not only depend on the physical properties of the histones and the DNA, but in addition the histones and thus their the physicochemical properties have evolved under selective pressure acting on the chromatin fiber and its function (Benecke 2003, 2006; Bécavin et al. 2010). This symmetry established by the retrograde action of evolution is something which currently can not be captured by techniques such as renormalization (Lesne 1998, 2011), but will need to be accounted for in multiscale integration efforts. We have defined the term *function-dependent self-scaling* for models which describe for instance chromatin structure as a function of activity at the scale relevant to this activity (Lavelle and Benecke 2006).

Multiscale integration in biology is a fundamental problem for which currently little ideas exist how it could be solved. There are a few other problems of similar fundamental nature such as the role of stochasticity in biologic mechanisms and

how robustness of these mechanisms across changing environmental and systems-internal conditions can be maintained (Kaern et al. 2005). Interestingly, stochasticity here might be a solution more than a problem in many respects, but again a formal framework to describe, quantify, and predict such mechanisms is lacking. In what will follow, we will discuss some recent insights into functional genome representations to add a novel layer of investigation to the problem of gene expression regulation, chromatin structural dynamics, and genome structure–function relationships. These representations are thought to be particularly useful to compare genomes from closely related species and more importantly to provide new ideas of how to treat the case of two or more genomes operating together in a single cell such as is the case in infectious settings (Aderem et al. 2011; Tisoncik et al. 2009). To this end, we will first discuss two recent examples of successful network structure inference and dynamics analysis in systems virology, analyze the implications these results have for our thinking of genome function, and finally provide some ideas how to further investigate these systems using functional genome representations as a first step for a multiscale modeling effort.

2 SIV Infection in Natural Hosts

The definition of an effective HIV vaccine has only made modest progress despite prodigious efforts, as HIV successfully evades efficient and durable recognition by the human immune system (Ross et al. 2010; Belisle et al. 2011). Similarly, AIDS resistance in SIV natural host primates has been formerly believed to be caused by a lack of innate and adaptive immune recognition. This view is currently changing as four independent systems biology driven efforts have investigated in a comparative manner, the transcriptome dynamics in PBMCs and CD4+ cells of natural hosts for SIV as compared to Asian/New World primates that develop AIDS following SIV infection. Indeed, natural hosts just as AIDS progressor species display a rapid and strong innate immune response to SIV infection, and display all signs of successful immune activation (IA). The changes in the gene expression profiles are not only remarkably concordant between different natural hosts such as African Green Monkeys (*Chlorocebus sabaeus*) and Sooty Mangabeys (*Cercocebus atys*), but also comparable in composition and strength to Rhesus Macaques (*Macaca mulatta*) and Pigtail macaques (*Macaca nemestrina*), the latter two being both AIDS progressors (Jacquelin et al. 2009; Bosinger et al. 2009; Favre et al. 2009; Lederer et al. 2009; Rotger et al. 2011). By systematic comparison of the gene networks indicative of IA between AIDS progressors and non-progressors not only common themes were identified, but also remarkable differences as to the duration of the innate immune response to SIV have been observed (Fig. 2). Indeed, IA in natural hosts ceases after the acute infection stage, typically after 2–4 weeks, whereas the gene networks driving the IA in AIDS progressors are still

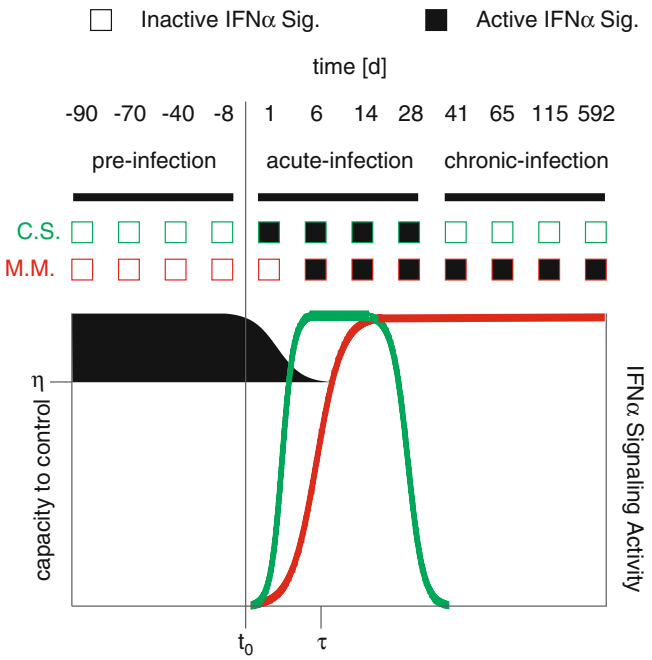


Fig. 2 Immune Activation in a Natural Host versus an AIDS Progressor—the West Coast Model. PBMCs from six African Green Monkeys (SIV Natural Host, *Chlorocebus sabaeus*, here: "C.S.") and six Rhesus Macaques (AIDS Progressor, *Macaca mulatta*, here: "M.M.") were analyzed pre- and post-SIV infection at the indicated time points using transcriptome profiling and the activity of the Interferon α signaling pathway was inferred using ontology enrichment analysis (□ = predicted inactive, ■ = predicted active, both at $p < 10E{-}3$) (Jacquelin et al. 2009). Two significant differences are observable: (i) C.S. control IA during the chronic phase of infection as opposed to M.M., (ii) C.S. seems more rapid in activating innate immunity than M.M. (Jacquelin et al. 2009). Similar differences are found in CD4+ cells from lymph nodes (Jacquelin et al. 2009), as well as other, independent studies involving a similar collection of data and different combinations of natural hosts and AIDS progressors (Bosinger et al. 2009; Favre et al. 2009; Lederer et al. 2009; Rotger et al. 2011). The recently proposed *West Coast Model* (Benecke et al. 2012) postulates that control of IA in natural hosts is a function of a mechanism reminiscent of kinetic proofreading (Hopfield 1974). Thereby, the capacity to control IA requires IA to cross threshold η before time τ. In the case of AIDS progressors, η is only reached after time τ, and thus the attenuation signal is not generated (a surfer missing the right moment to get on the board)

found active after the acute phase, and remain so until onset of symptoms of immunodeficiency (Bosinger et al. 2011, 2012; Manches and Bhardwaj 2009; Mir et al. 2011; Brenchley et al. 2010; Harris et al. 2010). Thus, it is the control of chronic IA, rather than absence thereof, which protects natural hosts from developing AIDS.

2.1 Control of Chronic Immune Activation in Natural Hosts

How can control, or absence of control in progressors, respectively, be thought to occur? Different hypotheses have been put forward, some of which can be disregarded or are unlikely to provide conclusive answers. SIV natural hosts do not display significantly altered infection or viral amplification rates and viral set-point titers. Moreover, chronically infected natural hosts maintain comparably high viral titers and can propagate virus. Viral particles isolated from natural hosts can be used to infect other animals (Jacquelin et al. 2009; Bosinger et al. 2009; Favre et al. 2009; Lederer et al. 2009; Estes et al. 2008). Thus, control of IA is neither directly connected to viral load nor is viral pathogenicity significantly altered during the course of infection.

The current hypothesis of how IA is attenuated in natural hosts is the presence of active signaling cascades which, upon a yet unidentified signal either attenuate IA in natural hosts or keep IA active in AIDS progressors. A logic table summarizes the four possible hypotheses depending on whether activators or repressors of attenuation or activation are considered (Table 1) (Bosinger et al. 2011; Harris et al. 2010). Currently, a specific search is underway in the different time resolved transcriptome profiles to identify such activators or repressors of either immune attenuation or IA, and which are differentially expressed/regulated in progressors and non-progressors. It will be of general, beyond the HIV field, interest to identify and characterize such activators and repressors which can promote or control chronic IA with obvious impacts for organ transplantation and autoimmune disorders (Rotger et al. 2011; Bosinger et al. 2011; Harris et al. 2010; Ye and Maniatis 2011; Lepelley et al. 2011).

The current generally accepted ideas on the control of IA in natural hosts, thus, postulate a necessary regulatory event (whether positive or negative) specific to either progressors or non-progressors. Thus, a dedicated signaling cascade composed of at least a sensor for a specific attenuation/activation signal, a transcriptional regulator, and a relay unit linking the sensor to the effector. Not only the molecules that are required specifically in either class of species, but also the nature of the specific signal pose a challenge in terms of evolution as an entire signaling pathway is required. Recall also that the signal for instance does not likely originate from the virus. Facing these dilemmas, we have recently formulated an alternate hypothesis for the absence of chronic IA in natural hosts which is based on a dynamic interpretation of the earliest innate sensing events following viral infection (Benecke et al. 2012). For the time being, this hypothesis is only modestly carried by direct experimental observations, as the time resolution with which early signaling events are usually studied is at least an order of magnitude above what would be required to directly assess the merits of the proposition. On the other hand, if this hypothesis, which appeals through its simplicity, would turn out to lead to the identification of a novel mechanism controlling long-term IA through early events, it would also define novel possible avenues for HIV vaccine development.

Table 1 Logic table for current hypotheses regarding control of IA in natural hosts

	Immune attenuation		Immune activation	
	N.H.	A.P.	N.H.	A.P.
Activator	+	−	−	+
Repressor	−	+	+	−

N.H. Natural Host, *A.P.* AIDS Progressor, + present, − absent

2.2 Kinetic Proofreading as a Possible Mechanism for IA Control

Kinetic proofreading is a potent mechanism known in molecular discrimination (Hopfield 1974). Kinetic proofreading is a process in which, through expenditure of additional energy, ligand recognition is split into two or more individual events in order to increase specificity and discriminatory capacity between closely related ligands or interaction partners with modestly different free energies of binding. In a first step, usually coupled to a conformational change in the receptor achieved through the hydrolysis of ATP, a candidate ligand is bound and presented to an independent interaction surface. Only if this second, independent interaction occurs rapidly enough, the recognition is conclusive, otherwise the ligand is released as the receptor snaps back into its original conformation. This mechanism has been studied in great detail theoretically and shown to drastically increase recognition of a *bonafide* ligand over analog molecules with very similar free energies. The error thereby is reduced beyond the thermodynamic bound—sometimes referred to as the specificity paradox upon which Hopfield based his predictions that ribosomes match codons and amino-acid-loaded tRNA anticodons using a kinetic proofreading mechanism. This has later been proved experimentally also for the way that aminoacyl tRNA synthetase operates (Hopfield 1974; Hopfield et al. 1976). Furthermore, and more relevant to this discussion, T-cell receptors use kinetic proofreading to enhance discrimination of *bonafide* ligands from closely related molecules to ensure correct signaling (McKeithan 1995). Finally, some evidence suggests that kinetic proofreading could also be found at the basis of RIG-I or TLR mediated recognition of foreign in innate immunity (Loo and Gale 2011; Liu and Gale 2010; Suthar et al. 2010).

For the sake of argument, let us assume that a strong and immediate innate immune response is not only a first line of defense to gain the required time for setting the stage for adaptive immunity, but that it is also a mechanism to proofread the adaptive immune response. In this scenario, some of the mechanisms of innate immunity would be required to be activated in order to maintain sustained, general IA beyond acute infection. Absence of innate proofreading would then lead to total inactivation of immune function. However, also the exact opposite effect might be at work—innate proofreading is required to attenuate continued IA. We believe that this latter scenario is more likely, and better reflects the general observations made about immunity. A typical pathogen will trigger

(many) different innate sensors simultaneously. The multitude of signals acts synergistically to mount the immediate innate IA which in turns triggers adaptive immunity. Maintaining this early response over prolonged periods of time, as observed in AIDS progressors, does not add any advantage to the system, however, is costly in terms of energy expenditure and precludes specific activation of downstream processes. If one of the different innate sensing mechanisms serves as proofreading mechanism, it makes more sense to propose that the proofreading is meant to attenuate the early innate response rather than sustaining or driving it as the latter would be redundant with the other mechanisms. In other words, the proofreading would simply signal that innate IA has been successfully triggered and thus needs to be attenuated in the near future in order to set the stage for adaptive immunity, avoid exhaustion of resources, and redundant signaling without added benefit.

Therefore, an innate sensing mechanism that triggers attenuation of IA would represent a simple feedforward control which does not require any additional specific signaling pathways or additional signals in order to be functional (see Goodman et al. 2011) for an interesting example of a feedforward mechanism in viral replication). This appears to be one strong argument in favor of the existence for such a dual purpose innate sensing that acts in one of those two aspects reminiscent of kinetic proofreading.

The second interesting argument can be formulated in favor of this hypothesis which is the dynamics of proofreading. As discussed above, through the addition of irreversible (energy consuming) steps prior to and integral part of faithful recognition a delay function is implemented. In other words, every one of the independent irreversible prerecognition steps needs time to complete; and thus the increase in specificity of recognition is not only 'bought' through energy consumption but also accompanied with varying delays between the initial encounter and positive recognition, which are a function of the number of successive prerecognition steps and physical proximity. In this context, the time delay creates a lag-time for the attenuation signal of innate activation which would prevent early shutdown. In other words, not relying on a specific signaling pathway for attenuation creates the problem that innate IA and its attenuation are triggered at the same time leading to conflicting signals. If, however, the attenuation signal is lagging behind because of its increased specificity, a functional feedforward repression is implemented (Benecke et al. 2012).

Finally, the dynamics of such a proofreading mechanism could potentially also explain the differences observed between natural hosts and AIDS progressors following SIV infection. As a matter of fact, a kinetic proofreading mechanism defines two boundaries on time. First, discussed above, there is a lower bound for the recognition process defined by the delay in time over the one or several irreversible steps. But also a second, upper bound, on time is explicitly part of the mechanism. If the recognition step n is too slow compared to the step $n-1$ the process aborts as unsuccessful. Hence, the execution time for step n is bounded by a function of the off-rate of $n-1$. Practically speaking, the hypothesis presented here suggests that there exists a window of time during which recognition has to

occur in order to trigger attenuation of innate IA. This window of time starts with the earliest prerecognition event at t_0 (infection) and continues up to some upper limit τ which has to be sufficiently close that robust (a significant fraction of a large number of events) recognition can occur. If this recognition occurs to late, the attenuation signal can no longer be released and IA continues chronically. This is a strong hypothesis which should be verifiable experimentally. Indeed, there even seems evidence in the existing transcriptome profiles for early dynamics playing a key role in the attenuation of IA in natural hosts, and why immune attenuation does not occur in AIDS progressors (Fig. 2). Indeed, it appears that innate IA occurs more rapidly in the natural host AGM as compared to Rhesus Macaques if the ontology-based inference of the activity of the interferon α pathway is accepted as a proxy (Jacquelin et al. 2009). The lower schematic illustrated the two main differences in the activation and attenuation kinetics between the AGMs (green) and the Macaques (red) and also schematizes the window of opportunity (black) for a feedforward attenuation mechanism reminiscent of kinetic proofreading. The threshold η needs to be crossed by the early recognition events before τ expires (see above) and too slow IA in the case of AIDS progressors (red), albeit sufficient in amplitude to cross η, fails to do so within the window of opportunity set by the proofreading mechanisms' upper and lower bounds on time. Note that, we assume here that the lower bound is defined by the first encounter with viral particles/components thereof or immediately after, thus is identical for the two species in this experiment, and that the upper bound is a function of the intrinsic lifetime of prerecognition complexes assumed to be identical in both cases as well. Thus, the only variable in the system is the speed with which IA occurs in both species. This can be viewed as analog to the situation of a surfer. If pathogen encounter and innate recognition as foreign is considered a wave at the beach, then IA could be seen as a surfer getting up on his surfboard. If the surfer fails to mount during the window of opportunity (defined by the width of the wave-back, thus intrinsic to the wave), the surfer will sink; thus, the term west coast model used (see Benecke et al. 2012 for a detailed discussion on this argument). Relevance of this model stems from the following observations: SIV-infected NHPs and HIV-infected human AIDS progressors mount their innate immune response too slow or rather too late leading to a non-attenuation and thus chronic IA. This unresolved innate IA wears down the system and leads consequently to decline in CD4+ T-cells, the hallmark of AIDS (Pandrea et al. 2011). Natural hosts for SIV on the other hand, such as sooty mangabeys, African green monkeys, and mandrills display timely responses to infection leading to successful IA and concomitant IA attenuation and, due to absence of specific humoral responses long-term tolerance of the virus (Jacquelin et al. 2009; Bosinger et al. 2009; Favre et al. 2009; Lederer et al. 2009; Rotger et al. 2011).

Comparative transcriptome profiling between an SIV infected natural host (here: *C. sabaeus*) and a progressor (here: *M. mulatta*) shows evidence of a lag-time of IFNα (as proxy for innate IA) signaling in progressors (Jacquelin et al. 2009) (Fig. 2). Note that, this delay of about a week might, however, be due to phenomena not necessarily related to the kinetics of IA, as the amplification

kinetics of the two adapted SIV viruses might be different, or for instance, we do not know whether or not the effective doses might be different between the two species. Still, it seems unlikely that such before mentioned effects would entail such profound changes in the IA kinetics, and thus this experimental finding might be regarded as a potential support of the proposition of kinetic autoattenuation of IA in natural hosts. It will be of outmost interest to better characterize the activation dynamics across the entire spectrum of known natural hosts and progressors in order to contrast possible differences in the activation kinetics with human subjects (or more likely *ex vivo* cellular models) representing the different observed classes [progressor, long-term non-progressor (LTNP), elite controller (EC)] as especially the LTNPs would be candidates of having acquired a similar attenuation mechanism as natural hosts. In this context, particular attention should also be given to the investigation of co-infection schemes with different pathogens (Schreiber et al. 2011). This would then also lead to the proposition that, similarly as to non-human primates, it is not the absence of an effective adaptive immune response to HIV itself but the failure to control the innate immune response which is the main driver of AIDS.

2.3 The Importance of Timing Across Scales

In conclusion, the proposition of mechanisms similar to kinetic proofreading for the coupling between innate and adaptive immunity is appealing as it combines simplicity with fidelity. Thereby, innate IA, with its obvious role of identifying foreign from self, would in the same time serve as a guard against inappropriate initiation of adaptive immunity by automatically attenuating the primary response. In order for this model to work, however, one needs to evoke the concept of a fading capacity to attenuate IA, and postulate that the attenuation threshold η is never reached in AIDS progressors in time τ (Fig. 2). Conclusive insights on the model presented for the coupling between innate and adaptive immunity, and the propositions regarding SIV and possibly HIV infection will require the successful translation of molecular profiles such as the transcriptome profiles obtained in the four cited studies into a dynamic view of the host's cellular immunity. This might sound simpler than it indeed is for several reasons such as experimental limitations imposed by the model systems or the technologies at hand for monitoring molecular events and their proxies (mRNA, signal cascade activation, metabolic activity), but mainly as one will need to overcome the problem of integration over multiple scales from the dynamics of single molecular events (in the micro- to millisecond range) to events at the organ level occurring on the scale of hours to days (please refer to the remarks made in Sect. 1). After having discussed briefly the second example of the importance of the network dynamics in immune responses from respiratory virus infections in Sect. 3, we will develop some ideas of how this general problem might be partially solvable for the particular cases discussed here (Sect. 4).

3 Network Dynamics in Respiratory Virus Infections

Other chapters of this volume discuss in great detail the case of different respiratory viruses and their interactions with their native hosts. We will, therefore, discuss here only a single finding from recent work on a meta-analysis of host transcriptome responses to a compendium of essentially Flu and SARS infection scenarios. As will be seen below, the observation made by Chang et al. (2012) pertains to host response dynamics, similarly as the studies discussed with respect to SIV and the innate IA in different hosts. Distinctively, the respiratory virus example does not compare different hosts for the same of differently adapted viruses, but rather different viruses (or pathogenic states) in a single host.

3.1 Meta-analysis of Mouse Transcriptome Responses to Respiratory Viruses

The threat of a highly lethal viral pandemic remains a major threat; the recent SARS-CoV 2003 and the H5N1 pandemics testify to the uncontrolled potential of emergence of respiratory viruses with possibly devastating characteristics reminiscent of the 1918 Spanish Flu (Donnelly et al. 2003; Beigel et al. 2005). Accordingly, major efforts are directed toward an understanding of the viral determinants of pathogenicity and their possible horizontal drift on the one hand and possible restriction factors or key modulators of pathogenicity on the side of the host on the other.

Deriving robust and unique molecular fingerprints for physiopathologic phenotypes from massive parallel experimental data is not only of extraordinary value for the understanding of pathogenicity but also a serious challenge given the current absence of systematic procedures (Ein-Dor et al. 2005). Biologic variability and insufficient sampling of the relevant state-space at present preclude formal approaches to molecular signature definition. A molecular signature is best defined using the isolation principle (Gregorius 2006) as the minimum number of biologic observables required to (i) discriminate the studied phenotype from some (ideally: any) other existing phenotype (external isolation), (ii) differentiate sufficiently between replicate analyses of the same phenotype thereby capturing biologic variation (internal cohesiveness), (iii) be robust against technical and biologic variability, and (iv) be of biologic relevance by representing the underlying more complex phenotype in its principal characteristics.

In order to advance in the definition of the hallmarks of lethal infection by respiratory viruses, Chang et al. compiled a compendium of published individual transcriptome studies on mouse lungs in order to identify gene signatures which abbey by the definitions set forth above. The compendium of microarray data from the 12 analyzed studies was composed of a total of 733 individual transcriptome profiles, roughly equally distributed over the three physiopathological groups

('high', 'medium', and 'low' pathogenicity) and their corresponding controls. Four different methods of meta-analysis stemming from two different philosophical approaches were used and compared in their absolute and relative performance. Processed data were either converted to logratios to identify genes that show opposite regulation in HPI and LPI, or directly submitted to meta-analysis by direct comparisons. In previous studies, both targeted and genome-wide approaches have been used to identify particular host pathways deregulated during infection. In parallel, a direct comparison of gene expression in 'high' and 'low' pathogenicity groups was performed. Statistically significantly differentially expressed genes were compiled to result in a characteristic gene signature when comparing the initial 'high' and 'low' groups. The fundamental difference between the three earlier, logratio based methods, and the latter direct comparison signature is the implicit choice of reference gene expression levels as well as the subsequent classifier used to choose signature genes. While the former methods will select for those genes that are uniquely/oppositely regulated in 'high' versus 'low' pathogenicity settings, the latter will select for genes that are statistically significantly differentially expressed between both conditions. The logratio meta-analysis derived signatures could be, in accordance with Sonnenschein et al. (2011), referred to as 'digital' and the direct comparison signature which comprises both gene IDs and gene expression values as 'analog'. All of the pathogenicity signatures were then compared among each other and characterized individually toward the objective to characterize responses that were present across high-pathogenic infections (HPI) and low-pathogenic infections (LPI).

The analog pathogenicity signature (aPS), correctly classified test data from the comparison of infection with one of two swine-origin influenza virus A strains, pandemic H1N1 (CA/04), or a mouse-adapted lethal variant (MA1 CA/04) (Bradel-Tretheway et al. 2011) not comprised in the initial compendium used for the competitive meta-analyses. In-depth analysis of the aPS revealed, furthermore, that biologic conditions classified as intermediate between HPI and LPI often belonged in the case of MPI data to late time points after infection, and for HPI data to early time points, leading to an analog immune response model for respiratory virus infection. The aPS derived by comparative meta-analysis of this respiratory virus infection compendium can be, thus, used to correctly classify host transcriptome responses according to clinical pathogenicity. The reason why the aPS outperforms the alternate digital pathogenicity signatures derived through the other three meta-analysis methods is explained by the striking observation of an analog that is continuous and correlated, host gene expression response to pathogenicity. Gene expression of this continuos response can be either positively or negatively correlated with pathogenicity, the latter being only recently recognized to exist (Kash et al. 1918; Cilloniz et al. 2010). This finding has not only technical implications for molecular signature definition strategies, but also for the understanding of the physiopathology of respiratory virus infection: continuous responses of gene networks to pathogenicity rather than different or oppositely regulated networks specific to 'high' or 'low' pathogenicity dominate the immunologic response of the host to viral infection which has major implications

for medical targeting of these networks. On the other hand, the observation of analog immune responses lends hope to the successful identification and boosting of host innate and adaptive immune mechanisms against high pathogenicity infections.

3.2 Dynamic Interpretation of Gene Expression and Pathogenicity Correlation

Important in this context is the possibility that infectious outcome might be encoded by the activation dynamics of host response gene regulation. In other words, one might have a hard time to find genes specifically responding to HPI or LPI, but rather only different activation dynamics for genes regulated in either case. Figure 3 illustrates the possible underlying mechanisms for such an observation.

Comparative meta-analysis of the host transcriptome dynamics following infection with high- or low-pathogenic respiratory viruses identified a gene signature characteristic of the pathogenicity of the virus (Chang et al. 2012). Highly pathogenic viruses such as influenza A subtype H5N1, reconstructed 1918 influenza A virus, and SARS-CoV thus illicit the same immune reaction than low- and medium-pathogenic viruses, however, to a higher degree. The observed strong correlations with pathogenicity could originate from two different, dynamic regimes of the underlying network (Fig. 3).

In conclusion, the meta-analysis of transcriptome profiles from respiratory virus infections reveals again critical dynamics of innate immunity at time-scales below currently investigated scales. The possibility of similar mechanisms at work when comparing the case of SIV infection in natural hosts (Sect. 2) and respiratory virus infections in mice (Fig. 3 right), possibly even further strengthens the general idea of time dynamics being of critical importance to host–pathogen interactions. In the following section, we will ask how such dynamics can be better inferred and analyzed using novel genome representations.

4 Integration Over Time-Scales Using Probability Landscapes Over Genome Sequences

In what follows, we will discuss a recent proposition for a mathematical description of a genome and associated activities. We will first argue for the need of such a structure, then discuss the general outline of the recently proposed structure, and finally discuss how this structure might help to further the concepts

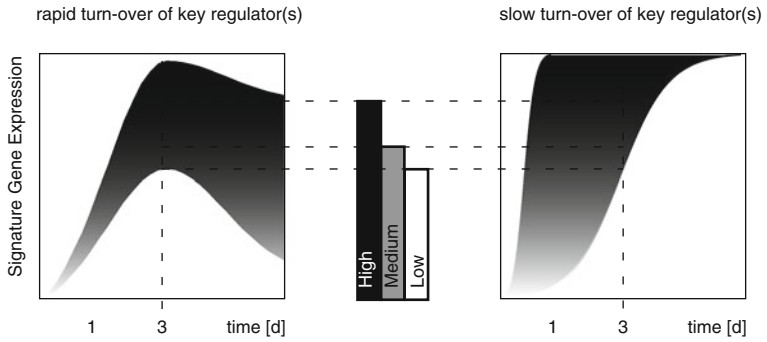


Fig. 3 Two alternate dynamic interpretations of the observed strong correlation between gene expression activity and pathogenicity (Chang et al. 2012). The uncovered positive and negative correlations between mRNA levels produced from a signature set of genes relevant to respiratory virus infection in mice with the corresponding pathogenicity of the virus (viruses or conditions were attributed to one of three discrete categories 'high', 'medium', and 'low', center) have two possible mechanistic origins. First, as initially proposed by Chang et al. (2012), while variable in time, a given gene at any given moment will be expressed as a function of viral pathogenicity (*left*). Second, it is also possible that all the signature genes will share similar expression values independent of the pathogenicity of the virus, in this case, however, at different moments in time (*right*). These regimes are not necessarily exclusive. Note that with the current resolution of the existing data a direct inference of which of the two regimes actually at work is impossible. Note also that the identification of which of the two mechanisms is at work would lead to strong, testable hypotheses, and provide directions for future experiments aiming at dissecting the gene regulatory network(s) relevant to the viral pathogenicity. The identification of the key regulator(s) driving the effective network and its dynamics were greatly facilitated if one could make a prediction as to the turnover of these regulators (which can be estimated from the time-series data for all genes). Note finally that the regime described on the right—disparity in activation (and symmetrically repression, not illustrated)—resembles the observations made in the case of comparative SIV infection in natural hosts versus AIDS progressors (Sect. 2, Fig. 2, opening the exciting possibility of a similar, if not identical, phenomenon taking place in both scenarios)

discussed in the two examples above (Sects. 2, 3) by providing a basis for the decomposition and inference over multiple time-scales (Sect. 1).

4.1 Requirements for a Mathematical Structure for the Object Genome

Today, genome biology is essentially based on (linear) statistical approaches. This is somewhat surprising as the amount of available information and experimental data is not, nor likely will ever be in the near future, sufficient to derive proper statistics on the object 'genome'. The large number of different biologic conditions will not be exploitable and the space of biologic conditions hence will remain extremely sparsely sampled. Furthermore, it will almost nowhere reach sufficient

density (e.g. recordings of many independent biologic replicates) to allow proper statistics. Moreover, simultaneous observation of all relevant determinants at all relevant scales over time is not possible, the experimental data will remain independent observations. Statistics on those will not enable to construct causal links rather than correlations between them. Furthermore, standard statistics is inappropriate for the questions posed since biologic processes are not generic, and arguments of parsimony, typicality, and natural chance of occurrence fail. Finally, statistical descriptions *per se* do not provide causal relationships, and hence do not provide comprehension of the underlying mechanisms. There are no obvious computational remedies to these limitations due to the evolutionary (and possibly other) feedback from the level of the higher, emergent scale down to the molecular scale as discussed in Sect. 1 (Moore 1990; Israeli and Goldenfeld 2006).

The object genome (which includes all of its possible activities) is likely to be 'computationally irreducible' (Moore 1990), meaning that if we aim at computing the behavior generated by genomic information, we have to perform as many operations as there are time steps, elements, and interactions. There is, hence, little possible reduction of the complexity of the biologic system *genome* by computational methods unless a unified, mathematical self-consistent structure can be formulated. *Time* will be one important but not necessarily privileged dimension of such a structure.

4.2 An Emerging Proposition for a Mathematical Genome Structure

In order to go beyond statistical approaches and, thus, to reach a level of understanding of genomes which is sufficient for meaningful inference of regulatory processes the current concept of a letter-based alphabet for genome coding needs to be revisited. Comprehension, or at least the possibility of inference of networks and their dynamics over multiple scales is likely a prerequisite to targeting multifactorial diseases such as cancer, genetic disorders, or pathogen-induced malignancies. The examples discussed above illustrate well the limitations of current methodologies at hand. Let us, thus, first recapitulate the main features which need to be captured by mathematical (or functional) genome representation: a genome (i) codes for a number of molecular machines that catalyze elementary biochemical reactions, and (ii) has evolved to orchestrate the molecular machines in a manner that whatever form the organism takes in response to external or internal stimuli the organism remains alive (Benecke 2006). This seemingly trivial concept that any transitions from one functional (active) form of the genome/organism to another can only happen at the condition that any intermediate represents a viable genome/organism needs to be exploited as it is the strongest constraint on the system. The true 'miracle' does not lie within the elementary machines but within

the fact that they self-organize across different time and space scales into a functional form whether it be at an embryonic or an adult state (Smet-Nocca et al. 2010). It is the rules of interaction (direct or indirect) that are at the essence of the genome. These rules of interaction are coded in the genome at its sequence level, but also on the level of its structural and spatial dynamics (for instance: activity-dependent subnuclear localization, or localization-dependent activity). Thereby, any elementary information in the genome (such as a single nucleotide) has a role (even seemingly negligible) of coding for any part of the functional forms of the genome at different time and space scales (Benecke 2006). The functional forms of a genome are thus expressed through nonzero contributions (weights) from individual elements which interact within a highly constrained, hence rigid structure. Note that from a computational viewpoint, an active genome is presumably a universal Turing machine (Benecke 2006). Recently, an initial proposition for a mathematical representation has been made where nucleotide frequencies as well as measurements on the activity of any part of the genome under defined biologic conditions are simultaneously expressed as probability distributions (Lesne and Benecke 2008a, b). This mathematical structure allows, which yet also has some questionable properties, see below, allows to introduce concepts from algebraic geometry for data analysis and modeling. We thereby use three independent paradigm shifts which lead to a modified approach to the inference problem in functional genomics (Benecke 2008).

4.3 Probability Landscapes

A genome is currently represented as a string composed of a four to six (DNA methylation, gaps) letter alphabet. Most approaches consist of identifying meaningful 'words' within this text, often by trying to identify over-represented subsequences that coincide with measurable quantities or changing quantities such as a gene, the amount of RNA transcribed from a gene, or the presence of a gene regulatory factor or particular chromatin modifications associated with the studied process in a given biologic condition. The genomic sequences obtained over the past decade reveal a low complexity of the genomic sequence, especially in non-coding regions, and consequently high-fidelity statistical inference of functional elements is essentially limited to protein coding sequences which account for only ≈ 2 % of the total human genomic sequence. Paradoxically, even what was considered to be a well-defined concept, the notion of a gene, is being challenged by the recent discovery of short and long, untranslated RNA sequences (microRNAs, ncRNAs), and the discovery of increasingly complex patterns of alternate promoter and splice-site usage. The concept of probability landscapes replaces the one-dimensional view of a genome by a stacked structure over genome positions, where the stack contains the representation of all biologic objects and events relative to the position n along the genome (Fig. 4). This mathematical structure gives at the same time the framework to

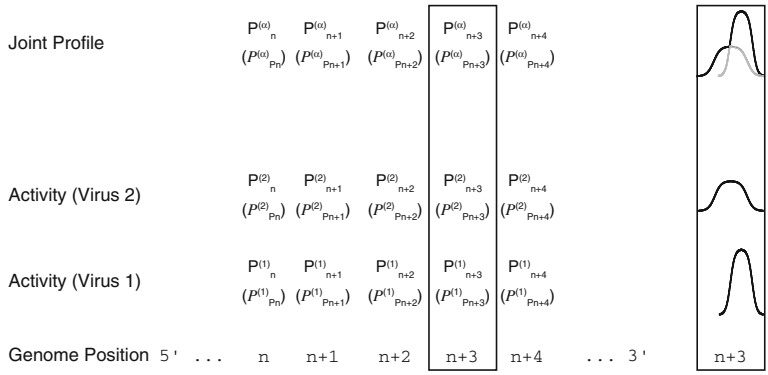


Fig. 4 Probability landscapes, which include as reference set the probabilistic representation of the genomic sequence obtained from several to many individuals, can be used to discover and analyze longitudinal correlations efficiently among the initially heterogeneous and unrelatable descriptions and genome-wide measurements. The structure consists of probability density distributions stacked on any genome position n defining the vertical extension. Horizontally, along the one-dimensional genome, a layer is generated for every biologic condition and every experimental measure. In this schematic representation, the probability distributions for two measures of activity of two different viruses over a five base genome is illustrated (Lesne and Benecke 2008). These profiles than can be integrated vertically (schematized on the *right*) using appropriate formalisms. A large collection of such geometric and algebraic ways to generate what is here referred to as joint profile exist (Lesne and Benecke 2008)

analyze data, to reconstruct missing information using rigidity-like and coherence arguments, and to express inherently multiscale causal relationships that can be used to explain genome function. Mathematical does not mean abstract, since on the contrary any set of experimental data or concrete interactions are transformable into the probability distributions (Lesne and Benecke 2008a). In turn, the probability distributions used allow the inference of a more integrated knowledge without having to prescribe all local properties and connected relationships. Rather than considering individual states of an active genome, probabilities describe the relevance of any object mappable to the genome (for instance: physical properties of chromatin, or transcription factor binding) to these states (Lesne and Benecke 2008a). As any relevant information on all levels, features (objects such as genes, regulatory sequences), and experimental data can be expressed as probabilities, a unified representation is obtained. The *ensemble* of probability distributions at site n constitute the stack and horizontally, thus over all positions n_i, a profile. Finally, rather than focusing on objects and states (or their probabilities) the aim of this form of representation is to be able to access the transformations between the probability distributions that govern their mechanistic, biologic relations. The set of transformations thereby constitutes the mapping functions f and Φ from Fig. 1 for the phenotypes associated with genome activity provided sufficient data have been integrated.

Probability landscapes provide, thus, a unified structure consisting of probabilities $(P_n)_n$ and associated quality estimates $(P_{P_n})_n$—in the form of functional probability densities (probabilities of probabilities)—to integrate any type of relevant genomic information into a coherent annotation. Most importantly, genomic sequence itself, its annotation with empirically derived features such as genes and regulatory sites, and any type of functional genomics data can be described in this manner. The rationale of this probabilistic description is not necessarily to account for an underlying stochasticity, though for some biologic processes this is indeed relevant, but rather to provide an efficient way to formulate partial knowledge and turn relative data of very heterogeneous nature and origin into absolute values and a homogeneous representation of the initial observations. Genome probability landscapes are systematic as any type of relevant information can be correctly and sensibly projected upon the genome positions. This projection has a single nucleotide resolution, producing a (at least locally) continuous profile. The proposed framework is coherent, as any information is converted without exception into the very same structure: probabilities with associated probability densities for local quality estimation. While the proposed representation of information is far from optimal in terms of compression, it provides a direct, systematic, and coherent interface for analysis, thus rendering numerical calculation efficient. The systematic nature of genome probability landscapes and their coherent structure allows easy exchange of information between different research teams. The simple structure of the resulting data also makes the framework easily portable between different computing environments as there is no real need for a solid database structure to generate, store, and handle the information provided that the same metrics are used to generate the profiles. Note that this aspect is a little oversimplification, as using the same metrics is not trivial when all aspects of quality control of the raw data, missing value imputations, and normalization have to be considered. It also appears that the concept is future compatible, as any type of relevant information can also be included in the very same manner into the existing landscapes (we disregard here whether or not this information makes previous data obsolete). This latter point is certainly of heightened interest giving the speed at which technology is developing for instance with respect to 'deep'/next generation-sequencing (NGS) and digital PCR. A structure that thus can meaningfully combine 'old' e.g., microarray type of data with 'new' NGS data will reduce the requirement for rerunning the same biologic conditions with the latest technology. Finally, the proposition to use probability landscapes for the integration of such data is—as it is inspired by and organized along the DNA sequence—a natural solution. Importantly, probability profiles can also accommodate the description of physical properties of DNA (for instance bending and intrinsic curvature) and chromatin fiber (local elastic constants, compactness), as well as the conformation of its nucleosomes and topologic constraints (conserved linking number within a loop); all these features are expected to play a key role in for instance transcriptional regulation (Widom 1998; Lesne and Victor 2006; Lesne and Benecke 2008a). Even nuclear dynamics could possibly be expressed

through the location, either central or peripheral of chromatin loci within the nucleus (Spector 2003; Cabal et al. 2006).

4.3.1 P-Landscape Based Analysis

Genome probability landscapes essentially provide the first step into processing any raw experimental data into a unified expression suitable for systematic genome-wide integration and analysis. To reduce unnecessary formal, mathematical, and computational complexity, we have developed methods for collapsing subsets of the landscapes whose basic step is an analysis of the stacks at a given genome location n (Lesne and Benecke 2008b). In the toy example given in Fig. 4, one might for instance want to ask whether it is necessary to consider the activity profiles of Virus$\sim$1 and Virus$\sim$2 as distinct or whether it is more meaningful to pool them. In other words, does the profile of Virus$\sim$1 when jointly considered with the one of Virus$\sim$2 provide independent information which needs to be considered or can the one be used to rather back the other? To answer, a measure called Kullback–Leibler divergence (Kullback and Leibler 1951) can be employed to measure the relative contribution of either activity profile to the joint profile. Each individual profile's weight to the combined measure is obtained using the average presumed frequency of these subsets (rather subpopulations). This amounts to one example of a vertical comparison which can be performed along the genome. Then, a longitudinal integration of the local divergencies is performed along genome regions of relevance (e.g. over the location associated to a given gene) allowing to analyze the feature divergence profile of a biologic condition over the entire genome or defined intervals. This genome-wide distance measure is meaningful, unlike the individual feature profiles. If the conditioning by any combination of individual or averaged profiles leads to a statistically significant divergence (suggesting that the associated subpopulation is well delineated and has a specific signature) the profile is kept as a separate entity. In contrast, if statistical significance is not reached, the condition is considered non-pertinent to the biologic question posed as it does not provide a measurable constraint on the value of the joint profile and can be combined with any other statistically insignificant conditions. This process, thus, integrates and thereby collapses part of the landscape to restrict to statistically divergent information (whether this is also biologic meaningful information can not be determined at this stage). Two advantages arise in this case: (i) the complexity of the structure is reduced in a controlled manner in so far as it is irrelevant to the biologic question investigated, and (ii) the statistical power of the joint probability profile is increased. As shown in Lesne and Benecke (2008b), this procedure can be performed at any interesting scale or functional level and thus the probability landscape over the genomic sequence can be reduced in complexity until all remaining context-dependencies reach statistical significance at which an optimum for computational complexity and statistical power is reached. Different biologic conditions can thereby be defined with maximum

flexibility using separate or overlapping subsets of subconditions in a hierarchical manner. The Kullback–Leibler divergence-based method discussed in Lesne and Benecke (2008b) represents, thus, a systematic and simple way of testing the statistical limits of complexity reduction and hence explanatory power of the integrative genomics data in their respective contexts. Note that since we are comparing the distributions of the same random variable under different conditions, it is only the distance (or divergence) between the two distributions that is meaningful. A joint probability, such as mutual information, could not be envisioned. This also holds for the case of two different variables because the joint probability distribution is inaccessible. From a general perspective, our method represents an application of concepts related to context trees to the probability landscape idea. Context analysis and landscape collapse thereby operate in similar manners to Markov chains with variable length for the analysis of time-series and historic context (Bühlmann and Wyner 1999; Maubourguet et al. 2008). We also note that the Kullback–Leibler divergence calculation provides measures that can be used directly for clustering of probability profiles. Clustering of probability profiles might help to establish and analyze relatedness among data otherwise not compared directly.

4.3.2 Expressing Time in P-Landscapes

As discussed earlier (Sect. 1), the successful integration of time over scales is one of the current bottlenecks of a systems biology description aiming at a discovery mechanism for mapping functions between objects and phenotypes. The two cited examples from virology (Sects. 2, 3) underline the potentially crucial importance of molecular dynamics and their coupling to macroscopic behavior. There are two different possibilities to incorporate time into probability landscapes. First, explicit integration using which will be based on directly using the different time points from the kinetic, to stay within the perimeter of the examples from above, transcriptome profiles to generate individual probability profiles now depend on time: $P_n^{(\text{Virus } 1)}(t)$ (probability to observe activity of Virus 1 in the experimental condition at site n and time t). It is then possible, generalizing the methodology developed for single time P-landscapes to compare those using for instance the Kullback–Leibler formalism, to align profiles from different biologic conditions (Virus ~ 1 vs. Virus ~ 2) using mutual information optimization to determine a local or global shift (compare Fig. 2), and finally fit a model of the evolution over time using a stochastic operator.

Alternatively, time might be captured only abstractly, and thus indirectly. Consider once more, the schematized behavior of the respiratory virus induced host response signature from Fig. 3. Whatever the interpretation of the experimentally measured result (center), thus whatever the underlying mechanism (rapid or slow turnover of key regulator) in both scenarios a density (here: pathogenicity) function over time is at the origin of the measured result. As discussed above,

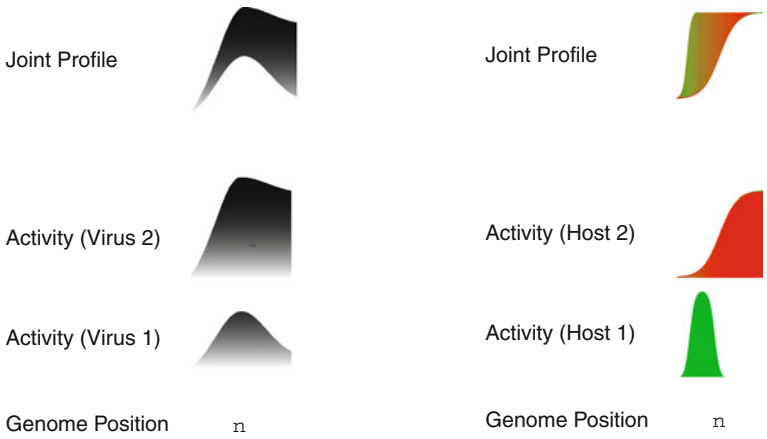


Fig. 5 Capturing time abstractly within the framework of probability landscapes (Lesne and Benecke 2008a, b). Both proposed mechanism (rapid or slow turnover of key regulator) which would lead to the remarkable correlation (and anti-correlation) between the expression levels of key signature genes for respiratory virus infection as a function of the pathogenicity of the analyzed virus lead to density distributions of gene activity with respect to time. These density distributions are characteristic for the virus and can be expressed as probability profiles along the host genome (here illustrated for a single genome position, which might be as discussed in Sect. 3, either indeed a single nucleotide or a consecutive stretch of the genome associated to a measured activity—simplest example would be the difference of resolution of NGS vs. microarray based transcriptomics). The virus-dependent, time-abstracted profiles then can be integrated into joint profiles using the same or similar formalisms as discussed in Sect. 4 and Lesne and Benecke (2008b)

probability density distributions are at the basis of the P-profiles generated from the to-be-annotated data. While so far only symmetric distributions have been described and studied (Lesne and Benecke 2008a, b), the formalism does not exclude the use of skewed, nontrivial distributions (Fig. 5). Furthermore, distance or divergence measures for skewed distributions, or parts thereof, can be defined. Thus instead of describing variability across individual measurements or different genetic backgrounds, the P_Pn part of the probability annotation would capture a generalized evolution over time. In this manner, only a single profile would be created for the entire time-series where the actual number of measured discrete time points is replaced by a continuously modeled distribution. Those distributions then can be studied in a fashion similarly as to what has been briefly discussed in Sect. 4 and in more detail in Lesne and Benecke (2008b). Again, a number of different ways to achieve such integration have been proposed (Selinger 2012). Indeed, in the example of the respiratory virus infection (Sect. 3), the proposed integration mechanism provides a means of discerning which one is the more likely of the two possible mechanisms, and thus prioritize the experimentally testable hypotheses.

5 Concluding Remarks

Systems Biology is a rapidly evolving field with is receiving a great deal of attention in the field of infectious disease research owing to the potential to provide a greater understanding of the pathogen–host interactions that control infection phenotype and disease outcome. A key aspect of the systems approach is the use of computational methods to collectively integrate high-throughput omics and traditional virologic or histopathologic data into a systems-level view that allows the identification of functional processes involved in pathogen-associated disease and the further illumination of host targets representing key points of control by pathogens.

Albeit having already made strong arguments in favor of a systemic analysis of the pathogen, the host, and most importantly their joint, interdependent activity, taking these analyses to the next level will require to overcome many current conceptual, technical, statistical, and computational bottlenecks. A key aspect of a higher level understanding, linking objects and mechanisms to organs and phenotypes, will be the integration of data on the one hand, and inference of network structure and dynamics on the other, over multiple scales. This problem is far from trivial, and ideas of how it can be overcome are still rare and in the early stage of development.

The potentially defining role of the network dynamics of host–pathogen interactions, as discussed on two recent examples, exemplifies the urgent need of identifying solutions of how to handle time across scales. Based on a recent proposition of a probability-theory derived approach for functional genome representations a first glimpse of methodology that might turn out to handle at least some of the problems arising through time disparity over scales was developed. Obviously, this approach, and even more so generalizable ideas of overcoming scales, will need many iterations of scientific thought and experimentation before we will see major breakthroughs.

References

Aderem A, Adkins JN, Ansong C, Galagan J, Kaiser S, Korth MJ, Law GL, McDermott JG, Proll SC, Rosenberger C, Schoolnik G, Katze MG (2011) A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. MBio 2(1):e00325-10

Bagnold RA (1936) The movement of desert sand. Proc Royal Soc Lond A 157(892):594–620

Bécavin C, Barbi M, Victor JM, Lesne A (2010) Transcription within condensed chromatin: steric hindrance facilitates elongation. Biophys J 98:824–833

Beigel JH, Farrar J, Han AM, Hayden FG, Hyer R, de Jong MD, Lochindarat S, Nguyen TK, Nguyen TH, Tran TH, Nicoll A, Touch S, Yuen KY (2005) Writing committee of the World Health Organization (WHO) consultation on human influenza A/H5. Avian influenza A (H5N1) infection in humans. N Engl J Med 353(13):1374–1385

Belisle SE, Yin J, Shedlock DJ, Dai A, Yan J, Hirao L, Kutzler MA, Lewis MG, Andersen H, Lank SM, Karl JA, O'Connor DH, Khan A, Sardesai N, Chang J, Aicher L, Palermo RE, Weiner DB, Katze MG, Boyer J (2011) Long-term programming of antigen-specific immunity from gene expression signatures in the PBMC of rhesus macaques immunized with an SIV DNA vaccine. PLoS One 6(6):e19681

Benecke A (2003) Genomic plasticity and information processing by transcriptional coregulators. Com Plex Us 1:65–76

Benecke A (2006) Chromatin code, local non-equilibrium dynamics, and the emergence of transcription regulatory programs. Eur Phys J E 19:353–366

Benecke A (2008) Gene regulatory network inference using out of equilibrium statistical mechanics. HFSP J 2:183–188

Benecke A, Gale M Jr., Katze MG (2012) Dynamics of innate immunity are key to chronic immune activation in AIDS. Curr Opin HIV & AIDS 7(1):71–78

Bosinger SE, Li Q, Gordon SN, Klatt NR, Duan L, Xu L, Francella N, Sidahmed A, Smith AJ, Cramer EM, Zeng M, Masopust D, Carlis JV, Ran L, Vanderford TH, Paiardini M, Isett RB, Baldwin DA, Else JG, Staprans SI, Silvestri G, Haase AT, Kelvin DJ (2009) Global genomic analysis reveals rapid control of a robust innate response in SIV-infected sooty mangabeys. J Clin Invest 119(12):3556–3572

Bosinger SE, Sodora DL, Silvestri G (2011) Generalized immune activation and innate immune responses in simian immunodeficiency virus infection. Curr Opin HIV AIDS 6(5):411–418

Bosinger SE, Jacquelin B, Benecke A, Silvestri G, Müller-Trutwin M (2012) Systems biology towards the understanding of nonpathogenic SIV infection in natural host primate species. Curr Opin HIV AIDS 7(1):71–78

Bradel-Tretheway BG, Mattiacio JL, Krasnoselsky A, Stevenson C, Purdy D, Dewhurst S, Katze MG (2011) Comprehensive proteomic analysis of influenza virus polymerase complex reveals a novel association with mitochondrial proteins and RNA polymerase accessory factors. J Virol 85(17):8569–8581

Brenchley JM, Silvestri G, Douek DC (2010) Nonprogressive and progressive primate immunodeficiency lentivirus infections. Immunity 32(6):737–742

Bühlmann P, Wyner A (1999) Variable length markov chain. Ann Stat 27:480–513

Cabal GG, Rodriguez-Navarro S, Genevesio A, Olivo-Marin JC, Zimmer C, Gadal O, Feuerbach-Fournier F, Lesne A, Buc H, Hurt EC, Nehrbass U (2006) Molecular analysis of SAGA mediated nuclear pore gene gating activation in yeast. Nature 441:770–773

Chang ST, Tchitchek N, Ghosh D, Benecke A, Katze MG A (2012) chemokine gene expression signature derived from meta-analysis predicts the pathogenicity of viral respiratory infections. BMC Syst Biol 5:202

Cilloniz C, Pantin-Jackwood MJ, Ni C, Goodman AG, Peng X, Proll SC, Carter VS, Rosenzweig ER, Szretter KJ, Katz JM, Korth MJ, Swayne DE, Tumpey TM, Katze MG (2010) Lethal dissemination of H5N1 influenza virus is associated with dysregulation of inflammation and lipoxin signaling in a mouse model of infection. J Virol 84(15):7613–7624

Donnelly CA, Ghani AC, Leung GM, Hedley AJ, Fraser C, Riley S, Laith A, Abu-Raddad J, Ho LM, Thach TQ, Chau P, Chan KP, Lam TH, Tse LY, Tsang T, Liu SH, Kong JHB, Lau EMC, Ferguson NM, Anderson RM (2003) Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet 361:1761–1766

Ein-Dor L, Kela I, Getz G, Givol D, Domany E (2005) Outcome signature genes in breast cancer: is there a unique set?. Bioinformatics 21(2):171–178

Estes JD, Gordon SN, Zeng M, Chahroudi AM, Dunham RM, Staprans SI, Reilly CS, Silvestri G, Haase AT (2008) Early resolution of acute immune activation and induction of PD-1 in SIV-infected sooty mangabeys distinguishes nonpathogenic from pathogenic infection in rhesus macaques. J Immunol 180(10):6798–6807

Favre D, Lederer S, Kanwar B, Ma ZM, Proll S, Kasakow Z, Mold J, Swainson L, Barbour JD, Baskin CR, Palermo R, Pandrea I, Miller CJ, Katze MG, McCune JM (2009) Critical loss of the balance between Th17 and T regulatory cell populations in pathogenic SIV infection. PLoS Pathog 5(2):e1000295

Goodman AG, Tanner BC, Chang ST, Esteban M, Katze MG (2011) Virus infection rapidly activates the P58(IPK) pathway, delaying peak kinase activation to enhance viral replication. Virology 417(1):27–36

Gregorius HR (2006) The isolation principle of clustering: structural characteristics and implementation. Acta Biotheor 54(3):219–233

Harris LD, Tabb B, Sodora DL, Paiardini M, Klatt NR, Douek DC, Silvestri G, Müller-Trutwin M, Vasile-Pandrea I, Apetrei C, Hirsch V, Lifson J, Brenchley JM, Estes JD (2010) Downregulation of robust acute type I interferon responses distinguishes nonpathogenic simian immunodeficiency virus (SIV) infection of natural hosts from pathogenic SIV infection of rhesus macaques. J Virol 84(15):7886–7891

Hopfield JJ (1974) Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proc Natl Acad Sci U S A 71(10):4135–4139

Hopfield JJ, Yamane T, Yue V, Coutts SM (1976) Direct experimental evidence for kinetic proofreading in amino acylation of tRNAIle. Proc Natl Acad Sci U S A 73(4):1164–1168

Israeli N, Goldenfeld N (2006) Coarse-graining of cellular automata, emergence, and the predictability of complex systems. Phys Rev E 73:026203

Jacquelin B, Mayau V, Targat B, Liovat AS, Kunkel D, Petitjean G, Dillies MA, Roques P, Butor C, Silvestri G, Giavedoni LD, Lebon P, Barr P, Sinoussi F, Benecke A, MŸller-Trutwin MC (2009) Nonpathogenic SIV infection of African green monkeys induces a strong but rapidly controlled type I IFN response. J Clin Invest 119(12):3544–3555

Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet 6(6):451–464

Kash JC, Basler CF, Garc'a-Sastre A, Carter V, Billharz R, Swayne DE, Przygodzki RM, Taubenberger JK, Katze MG, Tumpey TM (2004) Global host immune response: pathogenesis and transcriptional profiling of type A influenza viruses expressing the hemagglutinin and neuraminidase genes from the 1918 pandemic virus. J Virol 78(17):9499–9511

Kullback S, Leibler R (1951) On information and sufficiency. Ann Math Stat 22:79–86

Lavelle C, Benecke A (2006) Chromatin physics: replacing multiple, representation-centered descriptions at discrete scales by a continuous function-dependent selfscaled model. Eur Phys J E (Soft Matter) 19:379–384

Lederer S, Favre D, Walters KA, Proll S, Kanwar B, Kasakow Z, Baskin CR, Palermo R, McCune JM, Katze MG (2009) Transcriptional profiling in pathogenic and non-pathogenic SIV infections reveals significant distinctions in kinetics and tissue compartmentalization. PLoS Pathog 5(2):e1000296

Lepelley A, Louis S, Sourisseau M, Law HK, Pothlichet J, Schilte C, Chaperot L, Plumas J, Randall RE, Si-Tahar M, Mammano F, Albert ML, Schwartz O (2011) Innate sensing of HIV-infected cells. PLoS Pathog 7(2):e1001284

Lesne A (1998) Renormalization methods. Wiley, Chichester, ISBN 0-471-96689-4

Lesne A (2011) Morphogenesis. Springer, ISBN 978-3-642-13173-8

Lesne A, Benecke A (2008a) Probability landscapes for integrative genomics. Theor Biol Med Mod 5:9

Lesne A, Benecke A (2008b) Feature context-dependency and complexity reduction in probability landscapes for integrative genomics. Theor Biol Med Mod 5:21

Lesne A, Lagües M (2012) Scale invariance. Springer, ISBN 978-3-642-15122-4

Lesne A, Victor JM (2006) Chromatin fiber functional organization: some plausible models. Eur Phys J E 19:279–290

Liu HM, Gale M (2010) Hepatitis C Virus Evasion from RIG-I-Dependent Hepatic Innate Immunity. Gastroenterol Res Pract. 2010:548390

Loo YM, Gale M Jr. (2011) Immune signaling by RIG-I-like receptors. Immunity 34(5):680–692

Manches O, Bhardwaj N (2009) Resolution of immune activation defines nonpathogenic SIV infection. J Clin Invest 119(12):3512–3515

Maubourguet N, Lesne A, Changeux JP, Maskos U, Faure P (2008) Behavioral sequence analysis reveals a novel role for beta2* nicotinic receptors in exploration. PLoS Comput Biol 4(11):e1000229

McKeithan TW (1995) Kinetic proofreading in T-cell receptor signal transduction. Proc Natl Acad Sci U S A 92(11):5042–5046

Mir KD, Gasper MA, Sundaravaradan V, Sodora DL (2011) SIV infection in natural hosts: resolution of immune activation during the acute-to-chronic transition phase. Microbes Infect 13(1):14–24

Moore C (1990) Unpredictability and undecidability in dynamical systems. Phys Rev Lett 64:2354–2357

Pandrea I, Gaufin T, Gautam R, Kristoff J, Mandell D, Montefiori D, Keele BF, Ribeiro RM, Veazey RS, Apetrei C (2011) Functional cure of SIVagm infection in rhesus macaques results in complete recovery of CD4+ T cells and is reverted by CD8+ cell depletion. PLoS Pathog 7(8):e1002170

Ross AL, Brave A, Scarlatti G, Manrique A, Buonaguro L (2010)) Progress towards development of an HIV vaccine: report of the AIDS vaccine 2009 conference. Lancet Infect Dis 10(5):305–316

Rotger M, Dalmau J, Rauch A, McLaren P, Bosinger SE, Martinez R, Sandler NG, Roque A, Liebner J, Battegay M, Bernasconi E, Descombes P, Erkizia I, Fellay J, Hirschel B, Mir– JM, Palou E, Hoffmann M, Massanella M, Blanco J, Woods M, GŸnthard HF, de Bakker P, Douek DC, Silvestri G, Martinez-Picado J, Telenti A (2011) Comparative transcriptomics of extreme phenotypes of human HIV-1 infection and SIV infection in sooty mangabey and rhesus macaque. J Clin Invest 121(6):2391–2400

Schreiber F, Lynn DJ, Houston A, Peters J, Mwafulirwa G, Finlay BB, Brinkman FS, Hancock RE, Heyderman RS, Dougan G, Gordon MA (2011) The human transcriptome during nontyphoid Salmonella and HIV coinfection reveals attenuated NFκB-mediated inflammation and persistent cell cycle disruption. J Infect Dis 204(8):1237–1245

Selinger C (2012) On diffusion processes arising from optimal transport with applications to negative selection. Ann Math (to appear)

Smet-Nocca C, Paldi A, Benecke A (2010) From epigenomic to morphogenetic emergence. In: Bourgine P, Lesne A (eds) Morphogenesis. Springer, ISBN-13: 978-3-642-13173-8

Sonnenschein N, Geertz M, Muskhelishvili G, HŸtt MT (2011) Analog regulation of metabolic demand. BMC Syst Biol 5:40

Spector DL (2003) The dynamics of chromosome organization and gene regulation. Ann Rev Biochem 72:573–608

Suthar MS, Ma DY, Thomas S, Lund JM, Zhang N, Daffis S, Rudensky AY, Bevan MJ, Clark EA, Kaja MK, Diamond MS, Gale M Jr (2010) IPS-1 is essential for the control of West Nile virus infection and immunity. PLoS Pathog 6(2):e1000757

Tisoncik JR, Katze MG (2010) What is systems biology?. Future Microbiol 5(2):139–141

Tisoncik JR, Belisle SE, Diamond DL, Korth MJ, Katze MG (2009) Is systems biology the key to preventing the next pandemic?. Future Virol 4(6):553–561

Widom J (1998) Structure, dynamics and function of chromatin in vitro. Annu Rev Biophys Biomol Struct 27:285–327

Ye J, Maniatis T (2011) Negative regulation of interferon-β gene expression during acute and persistent virus infections. PLoS One 6(6):e20681

Index

Current Topics in Microbiology and Immunology (2013) 363: 261–263
DOI: 10.1007/978-3-642-33099-5