Integrative Statistical Learning with Applications in Predicting Features of Diseases and Health

by

Yongsheng Huang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2011

Doctoral Committee:

Professor Alfred O. Hero III, Co-Chair
Professor Jay L. Hess, Co-Chair
Professor Daniel Burns Jr
Professor Gilbert S. Omenn
Associate Professor Kerby Shedden

I dedicate this dissertation to my parents and my sisters. It is their unconditional love that gave me the courage and perseverance to continue on this long and winding road towards personal and professional improvement. For so many years, they have quietly and patiently waited for me to grow up. I dedicate this work to my true friend Jiehua Guo who is like my brother and helped me tremendously at many critical moments along this journey. I also dedicate my dissertation to the University of Michigan for granting me such a privilege to its invaluable educational resources. The time I spent studying here will always be one of the most significant parts of my life.

ACKNOWLEDGEMENTS

This dissertation is not even remotely possible without the guidance from Professor Alfred Hero, period. Professor Hero brought me into the world of mathematical statistics and taught me the true meaning of statistical thinking. He often worked with me late into the night and early in the morning, going through each analysis that I have performed and every sentence of manuscripts that I have written. He demonstrated the dedication and rigorous attitude towards science. Most important, he always challenged my intellectual capacity by showing me multiple elegant statistical approaches to a problem. That is the most powerful formula to motivate my desire to study more and think harder. And that is when I overcome the laziness inside of me and abandon the temptation to settle for easy solution. Dr. Jay Hess, kindly welcomed me into his laboratory to work with him and his team on the important problem of Hoxa9 protein. His scientific vision provided the biggest support to my research at the most challenging time when things do not piece together and make sense. He encouraged me to take the responsibility and ownership of my research and accepted my mistakes with forgiveness. As a mentor, both Prof. Hero and Dr. Hess genuinely care about my career development. They provided me complete academic freedom to pursue my research interests and to develop my professional skills. They showed, by example, the passion, scholarship, and mastery to scientific research. They emphasize and foster independent thinking and problem-solving ability. There has never been a doubt in my mind that I was granted a once-in-a-lifetime privilege to work with these two great mentors.

During my study in Bioinformatics, Dr. Omenn has always been on my side whenever I make important decisions. From finding academic mentors to choosing post-graduation career path, I have always been prepared and blessed with his wisdom, encouragement, and positive energy. Professor Daniel Burns went his way to help me even before I arrived at Michigan. Over the years, he helped me so many times that I lost my count. But I do know, my experience would have been much much harder without his help. Professor Kerby Shedden is the first committee member I met before everyone else. He interviewed me on the recruitment day and we got to know each other since. He is kind and supportive to me and my research. But, what I would say if i were asked what I will always remember from all my interactions with him? Without a second of doubt, I would say it has to be the four words he gave me on the study of statistics — "know the stuff cold". Plain simple! I never stopped working on it.

Over the years, I have also been very fortunate to have the opportunities to collaborate with four groups of exceptional researchers from all around the world. I thank Dr. Aimee Zaas, Dr. Geoffrey Ginsberg, Dr. Christopher Woods, Dr. Timothy Veldman, Ms. Christine Øien in the Duke University for the exceptional challenge study they have managed and the invaluable discussion they provided to my research. I also thank members of the Hess laboratory, especially Dr. Kajal Sitwala, Joel Bronstein, Daniel Sanders, and Monisha Dandekar. They have provided me superb data and biological insights. I thank Dr. Gordon Robertson and Mr. Timothee Cezard at the Genome Science Center in Vancouver, Canada. They generously shared with me their valuable knowledge on ChIP-sequencing. Particularly, I thank Dr. Robertson who literally taught me everything that I needed to know about next-generation sequencing analysis and showed me how quality research is done in high speed. I am also thankful to the members of Hero group, particularly Dr. Mark Kliger, Dr. Arvind Rao, Dr. Patrick Harrington, Mr. Yilun Chen, Mr. Kevin Xu, and

Mr. Arnau Tibau Puig, who helped my research in many ways. They always reminded me how similar statistics problems are approached from engineering perspectives — different disciplines, distinct applications, but same magic.

I am grateful to Ms. Julia Eussen, Ms. Denise Taylor-Moon, and Ms. Michelle Curry in the Bioinformatics Program; Ms. Lynn McCain in the Department of Pathology; Ms. Michele Feldkamp in the Department of Electrical Engineering and Computer Science. Their wonderful assistance have protected me from things that would have taken lot of time away from my research.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Integrative Statistical Learning with Applications in Predicting Features of Diseases and Health

by
Yongsheng Huang

Co-Chairs: Alfred O. Hero III and Jay L. Hess

This dissertation develops methods of integrative statistical learning to studies of two human diseases - respiratory infectious diseases and leukemia. It concerns integrating statistically principled approaches to connect data with new knowledge for improved understanding of diseases. A wide spectrum of temporal and high-dimensional datasets were considered, including various types of high-throughput measurements and clinical observations.

The first question studied in this thesis examined the host response to viral insult. In a human challenge study project, eight transcriptional response patterns were identified in hosts who were experimentally challenged with influenza H3N2/Wisconsin viruses. These patterns are highly correlated with and predictive of symptoms. A non-passive asymptomatic state was revealed and associated with subclinical infections. The findings were validated and extended to three additional viral pathogens (influenza H1N1, Rhinovirus, and RSV). Their differences and similarities were compared and contrasted. Unsupervised statistical models were constructed for exposure detection and risk stratification. Experimental validations have been performed by our collaborators at the Duke University.

The second question studied in this thesis investigated the regulatory roles of Hoxa9 and Meis1 in hematopoiesis and leukemia. Methods were developed to characterize their global *in vivo* binding patterns and to identify their functional *cofactors* and *collaborators*. The combinatorial effects of these factors were modeled and related to specific epigenetic signatures. A new biological model was proposed to explain their synergistic functions in leukemic transformation. Experimental validations have been performed by members of the Hess laboratory. Cross-platform integration of transcriptional, sequencing, and epigenetic profiles was provided to facilitate future study of acute leukemias.

Motivated by problems encountered in these studies, two algorithms were developed to identify spatial and temporal patterns from high-throughput data. The first method de-

termines temporal relationships between gene pathways during disease progression. It performs spectral analysis on graph Laplacian embedded significance measures of pathway activity. The second algorithm proposes probabilistic modeling of protein binding events. Based on information geometry theory, it applies hypothesis testing coupled with jackknife-bias correction scheme to characterize relationships between proteins. Experimental validations were shown for both algorithms.

In conclusion, this dissertation addressed practical issues in the design of statistical methods to identify characteristic and predictive features of human diseases. This thesis demonstrated the effectiveness of integrating simple techniques in bioinformatics analysis. Several bioinformatics tools were developed to facilitate the analysis of high-dimensional time-series biomedical datasets.

CHAPTER I

Introduction

1.1 Research Overview

Statistical learning is at the heart of bioinformatics analysis that connects data with knowledge for improved understanding of human health and diseases. It concerns the development and application of statistically principled approaches to model observed data, to identify unknown patterns, and to build predictive rules. Above all, it emphasizes on quantifying the intrinsic uncertainty associated with the derived knowledge in the presence of noise and strives to protect models from spurious artifacts. Biomedical data has become increasingly time-dependent, heterogeneous, large in quantity, and high-order in dimension. Consequently, it is a necessity rather than an option for any data-driven learning process to be highly integrative. On one hand, the inferences and conclusions about the underlying biomedical phenomena need to be drawn from combined evidence and *a priori* cross-disciplinary knowledge. On the other hand, various statistical techniques need to be appropriately chosen and carefully assembled for optimal learning performance. These two aspects constitute inseparable elements in the design of a learning strategy and require critical treatment for a given bioinformatics problem.

Most biomedical data consist of a mixture of discrete and continuous measurements that were made mover time or from repeated experiments. They are often sparse, skewed,

and small in sample size but huge in feature space. This makes the process of learning from data an extremely challenging task because a typical model fitting involves estimation of thousands of parameters and searches through a large collection of models. Although many statistical theories have been developed to address these issues, the overall knowledge acquisition process in a bioinformatics analysis remains to be problem-driven. This is so because each biomedical study presents a unique set of measurements that are pertaining to a specific scientific question. The statistical learning and inference from such dataset need to be considered in the context of the problem under study and different methodologies need to be integrated with data irregularities being properly handled.

The integrative characteristic of statistical learning is also reflected in its *iterative* applications to the scientific inquiry being pursued. The study subjects in biomedical domain are highly complex and dynamic in nature. It is unlikely that an one-size-fit-all approach will produce a definitive answer and solve the problem once for all. Instead, the analytic findings often invite new questions to be asked and/or new hypotheses to be formed. The direct result is additional or modified design of experiments followed by a new round of data collection and analysis. The learning strategy will then need to be adjusted to incorporate new data and modeling techniques. In addition, the learning strategy may also be significantly affected by the interactions from domain experts, namely biomedical research scientists alike. This further requires a highly flexible and robust modeling technique in its integration capacity.

The objective of this research is to tackle the challenge of integrative statistical learning on large-scale biomedical datasets obtained in two independent studies on acute respiratory infectious diseases and leukemia cancer. The two datasets are high-dimensional and temporal, composed of measurements from gene expression profiling, high-throughput ChIP-sequencing, biased and unbiased proteomic profiling, and clinical observations. They were

systematically and thoroughly analyzed using an array of statistical modeling techniques. More specifically, the first study investigates the temporal transcriptional pattern of host-virus interactions involved in respiratory viral infections. The aim is to identify differential characteristics of viral infection and build predictive models for current state estimation and forward state prediction of symptomatic diseases. The second project concerns with elucidating the mechanistic roles of two transcription factors Hoxa9 and Meis1 in leukemic transformation. We identified their global genomic binding characteristics. For both projects, the main focuses are the same - to conduct integrative statistical learning from tens and thousands of data points, to derive inference, to measure uncertainty, and to predict. To this end, the results presented here provide useful insights to the pathogenesis and progression of two important human diseases.

By presenting analysis and modeling rationale, this research demonstrates effective analysis strategies that allow multiple models to be constructed simultaneously and combined in a cohesive manner. The findings suggest that, in a large-scale data modeling situation, models may be built on a rather smaller scale from subsets of data using relatively simple techniques. These models are subsequently combined to yield new knowledge. Computationally, this is analogous to the parallel computing paradigm where a *divide-and-conquer* approach is used to handle computation-intensive tasks. It effectively reduces the size and complexity of a given problem. From a statistical point of view, simple models are favored because they impose less model assumptions and have the nice property of being close to the physical data. Contrastingly, a more complicated model can become more abstract and further away from the observed data being modeled. Although it provides better fitting in some cases, it also runs the risk of overfitting. The simplicity rule is thus less likely to suffer from technical issues such as data irregularities or model misspecification. This implies increased transparency between models and allows a more

straightforward and efficient model integration, resulting in more interpretable analytic results.

1.2 Outline of Dissertation

This dissertation is organized into three logical parts. The first part presents temporal analysis of human virus-induced acute respiratory illness. In Chapter II, we describe the temporal gene expression patterns that differentiates symptomatic hosts from asymptomatic hosts during Influenza H3N2/Wisconsin viral infection. A boosting regression model is introduced to detect exposure and stratify subjects into four groups that are associated with different level of risk for developing clinical overt symptoms. In Chapter III, we extend the analysis and validate the results from Chapter II by including three additional upper respiratory infectious viruses - Rhinoviruses (HRV), respiratory syncytial viruses (RSV), and influenza virus type H1N1.

In the second part, we focus on the genome-wide functional study of two transcription factors Hoxa9 and Meis1. These two oncogenic proteins are closely related to leukemic transformation. Their increased expression have been linked to poor prognosis phenotype in acute myeloid leukemia. Chapter IV details an integrative analysis on genomic, genetic, and epigenetic profiling data. A variety of statistical and computational modeling techniques were combined to identify the patterns of regulatory controls in transcription by Hoxa9 and Meis1 on both sequence-binding and epigenetic modifications. Based on these findings, a new biological model is proposed to explain regulatory functions by which Hoxa9 and Meis1 collaboratively promote acute myeloid leukemia transformation.

The third part of this dissertation proposes two algorithms that are motivated by two problems that arose from the two aforementioned studies. They currently do not have satisfactory solutions. Chapter VI describes an information geometry based method for

inferring protein-protein interactions from ChIP-seq data. The existing ChIP-sequencing related analysis methods focus on single transcription factor and are not adequate to address potential combinatory effects by multiple factors. We extend current sequence motif analysis with capacity for non-parametric inference and estimation of relationship between proteins. We show that this simple method for analyzing protein binding pattern in whole-genome sequencing studies. Existing methodologies on sequence motif analysis are reviewed in order to provide a context for discussion. Chapter V describes a spectral analysis method for studying temporal activities of biological pathways. Using graph Laplacian embedding of the significance measures of gene pathway activities, this algorithm can partition pathways into groups based on their temporal expression trajectory. Chapter VII concludes our work and outlines potential future research directions that we see can further validate and extend our findings.

1.3 Contributions of Dissertation

As this dissertation research applies integrative statistical learning theory in the solving of two real-world biomedical problems, it contributes to a broad spectrum of research areas. The original contributions of this research work are summarized as the following:

- To the field of bioinformatics and statistical analysis
  - Demonstration of the effectiveness of model simplicity. We show how simple statistical modeling techniques are effective in deriving and integrating knowledge from complex biomedical data without over-complicating the model.
  - An spectral method for studying temporal disease dynamics. In Chapter V, we develop an algorithm to perform spectral analysis of temporal gene pathway activities by embedding of their statistical significance measures of with graph

Laplacian.

  - An information geometry based method for inferring protein-protein interactions. In Chatper VI, we model the spatial distributions of protein binding sequence motifs with probability densities on a statistical manifold. This is followed by estimation of dissimilarities between probability distributions. The statistical significance of the estimates are assessed with simple hypothesis testing with a jack-knife bias correction scheme. This allows putative protein-protein interactions to be inferred.
  - Development of bioinformatics tools. Several bioinformatics tools are developed using R for (i) analyzing and visualizing high-throughput sequencing analysis; (ii) performing Bayesian factor analysis in microarray and potentially other high-throughput experiments; (iii) building classification models with bootstrap estimation of performance; (iv) temporal gene pathway analysis.

- To the study of infectious disease
  - Characterization of temporal host response towards influenza viral infection. In Chapter II, eight distinct host transcription patterns are identified. They differentiate symptomatic hosts from asymptomatic individuals who are exposed to influenza viruses.
  - A disease risk stratification model. In Chapter II, we built an unsupervised statistical model based on host gene expression profiles. This model is capable of detecting molecular signatures associated with influenza-mediated disease and stratifying observations into classes of different risks of developing symptomatic diseases.

  - Pan-viral study of multiple respiratory viral infections. In Chatper III, the temporal host response patterns are analyzed and compared for four different viral pathogens - influenza H3N2, influenza H1N1, Rhinoviruses, and RSV. Their similarities and differences are contrasted to provide new insights to virus-induced respiratory illnesses.

- To the study of acute myeloid leukemia

  - Characterization of Hoxa9 and Meis1 *in vivo* binding patterns. In Chapter IV, a catalog of Hoxa9 and Meis1 binding sites are identified and their genomic distribution patterns are analyzed.

  - Model of transcriptional control by Hoxa9 and Mei1. Chapter IV highlights a combinatorial regulation scheme by a group of transcription factors who collaborate with Hoxa9 and Meis1. This model accounts for a sophisticated multi-tier organization on their regulatory functions on target gene transcription. We provide functional and experimental validation of these results.

  - Integration and visualization of multi-modal Hoxa9 and Meis1 study data. Chapter IV provides a unified set of data including the transcriptional, genomic, epigenetic, and combinatorial protein interaction profiles with graphical visualization. This facilitates the study of human leukemia.

1.4 List of Relevant Publications and Softwares

This dissertation research results in several publications and bioinformatics tools that are published, submitted, or in preparation:

Selected Journal Publications

Y. Huang, A. Zaas, A. Rao, N. Dobigeon, P. Wolfe, T. Veldman, N. Oien, L. Carin, S. Kingsmore, C. Woods, GS. Ginsburg, AO. Hero. Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza A infection. Submitted 2010.

Y. Huang, K. Sitwala, J. Bronstein1, D. Sanders, M. Dandekar, G. Robertson, J. MacDonald, T. Cezard, M. Bilenky, N. Thiessen, Y. Zhao, T. Zeng, M. Hirst, A. Hero, S. Jones and J. Hess. Genome-wide functional characterization of Hoxa9 and Meis1 binding sites in hematopoietic cells. Submitted 2010.

Y. Huang, A. Hero. Towards Early Detection: Temporal Spectrum of Host Response in Symptomatic Respiratory Viral Infection. In preparation for submission in 2011.

Y. Huang, A. Rao, A. Hero III. Spectral Analysis of Temporal Gene Pathway Activities. In preparation for submission in 2011.

Y. Huang, G. Robertson, J. Bronstein1, D. Sanders, K. Sitwala, A. Hero, J. Hess. Information Geometry Based Inferrence of Motif Sptial Distribution. In preparation for submission in 2011.

Y. Huang, N. Dobigeon, A. Hero. rBLU: an R implementation for joint Bayesian feature extraction and linear unmixing. In preparation for submission in 2011.

Conference Publications

Y. Huang, A. Rao, A. Hero III. Assessing Temporal Correlation of Significant Gene Pathways Using String Edit Distance. *American Medical Informatics Association Summit*

on Translational Bioinformatics (Oral presentation). March 2010.

K. Sitwala, **Y. Huang**, M. Dandekar, G. Robertson, J. Hess. Genome-wide binding profile of Hoxa9 and Meis1 in leukemia cells. *American Society of Hematology Annual Meeting* (Abstract). December 2008.

Collaborative Works

A. Muntean, J. Tan, K. Sitwala, **Y. Huang**, J. Bronstein, J. Connelly, V. Basrur, K. Elenitoba-Johnson, J. Hess. The PAF complex synergizes with MLL fusion proteins at HOX loci to promote leukemogenesis. *Cancer Cell*. 2010 Jun 15;17(6):609-21.

A. Zaas, M. Chen, J. Varkey, T. Veldman, AO. Hero, J. Lucas, **Y. Huang**, R. Turner, A. Gilbert, R. Lambkin-Williams, N. Oien, B. Nicholson, S. Kingsmore, L. Carin, C. Woods, GS. Ginsburg. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host Microbe*. 2009 Sep 17;6(3):207-17.

A. Rao, **Y. Huang**, A. Hero. Identification and Query of Activated Gene Pathways in Disease Progression. In preparation for submission in 2011.

Software Implementations

cMotif: ChIP-sequencing motif analysis utilities.

SpecPath: Spectral analysis of temporal pathway activities.

rBLU: an R implementation for joint Bayesian feature extraction and linear unmixing.

PathEdit: Parallel implementation of pair-wise string edit distance computation for temporal gene pathway expression

CHAPTER II

Temporal Dynamics of Host Molecular Responses Differentiate Symptomatic and Asymptomatic Influenza A Infection

2.1 INTRODUCTION

Influenza viruses are highly infectious and can cause acute respiratory illness in human hosts. Infected hosts present a variety of clinical symptoms including fever, runny nose, sore throat, myalgias, and malaise with potentially more serious complications such as viral pneumonia (Cox and Subbarao, 1999). Many hosts also withstand comparable level of viral insult with little or no overt symptoms, exhibiting a higher degree of tolerance (Carrat et al., 2008, De Jong et al., 2006). Clearly, these asymptomatic infected hosts are able to control and eradicate viral threats more effectively than those who become symptomatic. Given the dynamic nature of viral infection, it is now recognized that interactions between hosts and viruses play a crucial role in determining the presence and absence of symptoms (Palese, 2004). This leads to an interesting question — what are the principal factors associated with such divergent disease outcome?

In recent years, seminal studies on the sensing of pathogens by pattern-recognition receptors (PRRs) and their related signaling cascades have advanced our understanding of innate immunity (Kawai and Akira, 2007, Stetson and Medzhitov, 2006, Kawai et al.,

2004, Honda et al., 2005, Yamamoto et al., 2003). Many elegant experimental analyses have further elucidated the mechanistic activation and modulation of host response to invading pathogens (Ichinohe et al., 2009, Yoneyama et al., 2004, Huang et al., 2001, Zhu et al., 2008, Fenner et al., 2006, Ryo et al., 2008, Proud et al., 2008). By design, however, host responses in these experimental conditions are often characterized for individual cells via cell culture; or they represent a snapshot of the immune response pertaining to a limited number of time points. The components of the host immune system are diverse and they interact in a complicated manner. Owing to both technical and ethical difficulties, it has not been practical to experimentally determine the full course of immune responses leading to severe symptoms in otherwise healthy human hosts. Thus the time sequence and orchestration of host response events remain to be fully understood.

The peripheral blood contains key elements of the immune system and the circulating immune cells recruited by the host in response to viral infection and virus-induced tissue damage provides a global view of the host immune response. Thus, we hypothesized that it can be used to monitor the temporal dynamics of host-virus interactions. Analyzing whole-genome gene expression profiles from healthy human subjects challenged with influenza H3N2/Wisconsin, we studied the full temporal spectrum of virus-mediated disease dynamics. This report offers an hour-by-hour detailed view of host immune response as a continuum, spanning the time from exposure to peak symptom manifestation and beyond. Utilizing biological and mathematical models, we highlight key immune response events representing potential factors that determine the pathogenecity of influenza viral infection. We further present a statistical risk-stratification model for estimating current disease state with potential forward risk assessment capability.

2.2 RESULTS

2.2.1 Temporal gene expression profiling of host transcriptional response

A cohort consisting of 17 healthy human volunteers (Table S2.1) received intranasal inoculation of influenza H3N2/Wisconsin and 9 of these subjects developed mild to severe symptoms based on standardized symptom scoring (Zaas et al., 2009). Gene expression profiles (GEP) were measured on whole peripheral blood drawn from all subjects at an interval of $\sim$8 hours post inoculation (hpi) through 108hpi. A total of 267 gene arrays were obtained for all 17 subjects at 16 different time points including baseline at $-$24hpi.

2.2.2 Screening for genes with different temporal profiles between asymptomatic and symptomatic hosts

We sought to determine genes whose temporal expression profiles differed significantly between asymptomatic (Asx) and symptomatic (Sx) subjects. A total of 5,076 significant genes were identified at false discovery rate significance level (q-value) $< 1\%$ (Storey et al., 2005). The changes of transcriptional program between the two phenotypes were dramatic both in terms of sheer number of transcripts affected and in terms of the magnitude of these changes. Furthermore, more than half of these significant genes showed marked time course activity in Asx subjects alone. The temporal changes observed in asymptomatic phenotype differ from those in symptomatic hosts in three important aspects - the particular time point at which a change occurred; the direction of the change; and the magnitude of the change. The temporal dynamics of these phenotypic differences elucidate the pathogenesis of symptomatic illness and uncover a dynamic state of the immune response in exposed individuals who ultimately do not develop symptomatic disease.

2.2.3 Co-clustering differentially expressed genes based on temporal expression dynamics

To reveal groups of genes sharing coherent temporal expression profiles, we applied Self-Organizing Maps (SOM) (Kohonen, 1995) to classify the 5,076 genes into clusters in a manner similar to a previous study (Huang et al., 2001), except that here both Asx and Sx gene expression profiles were simultaneously co-clustered. A total of eight clusters were identified and their associated prototypes (locally weighted average temporal gene expression) are shown as polar plots (Figure 2.1A) to display phenotypic contrast of expression dynamics. Each prototype captures the contrast between Asx and Sx expression pattern of genes in an individual SOM cluster. Heatmaps are shown for the top 5 genes from each SOM cluster (Figure 2.1B). The average magnitude of the expression level of each SOM cluster is shown, along with corresponding error bars as an indication of cluster purity (Figure 2.1C).

These eight clusters include genes that are differentially expressed at immediate early to early (0 – 12hpi), middle (12 – 45hpi), and late ($>$ 45hpi) stages of infection. The differential expression dynamics of these prototypes either sustain over all time points or extinguish after a short period of time. The contrasts in expression patterns between phenotypes are all statistically significant at q-value < 0.0001 (Figure 2.1C). For individual phenotypes, most clusters also show significant monotonic increase or decrease in expression over time (Table S2.3).

Specifically, cluster 1, denoted as (A^{nc}, S^{up}_{late}) where *nc* stands for no change, consists of genes whose expression rapidly increased in the Sx individuals, starting at approximately 45hpi (late stage). The average expression of Asx subjects in this cluster is virtually constant except for a small transient increase between 36 and 84hpi. Cluster 2, named $(A^{dw}_{early}, S^{up}_{mid})$, includes genes exhibiting sustained decrease unique to the Asx phenotype

from early time onward. In Sx, the expression of cluster 2 genes increased to their peak level at the middle of challenge (45 – 69hpi), followed by a rescinding trend. Cluster 3, (A^{nc}, S_mid^{up}), is characterized by strong activation, in Sx phenotype, of genes responsible for proinflammatory responses. Compared to cluster 2, genes in cluster 3 remain up-regulated several hours after peak symptom time (80hpi). Cluster 4, (A^{nc}, S^{dw}_{mid}), contains genes that were continuously down-regulated in the Sx phenotype in contrast to nearly no change in the Asx phenotype. Cluster 5, $(A^{dw}_{mid}, S^{dw}_{late})$, associates the Sx phenotype with a delayed decreasing expression at 45hpi in comparison to a much earlier decline ($\sim$ 12hpi) in the Asx phenotype. Cluster 6, $(A^{up}_{early}, S^{dw}_{mid})$, is populated by genes whose expression steadily increases in the Asx phenotype over all time. In contrast, for the Sx subjects these genes exhibit a transient but significant decrease beginning at 29hpi and return to baseline after 60hpi. Cluster 7, $(A^{dw}_{mid}, S^{dw}_{early})$, features an early transient (5 – 60hpi) decline in expression of its constituent genes in the Sx phenotype versus a sustained suppression of expression in the Asx phenotype starting from the middle stage ($\sim$ 12 – 21hpi). Cluster 8, (A^{up}_{mid}, S^{nc}), comprises genes that are strongly up-regulated in Asx subjects across the course of the study in contrast to a relatively weak response in Sx subjects. Evidently, clusters 2, 6 and 8 provide the best characterization of the asymptomatic host response and highlight the active state of the immune system as viral control is achieved. Functional pathway enrichment analysis shows that genes from these clusters are involved in a variety of biological functions, many of which directly relate to the activation and modulation of host immune and inflammatory responses (Table 2.1).

Clusters 2, 3, 4, and 6 (Figure 2.1) contain more than 78% of all significant genes that were differentially expressed between the Asx and Sx hosts. They highlight the sharp contrasts in expression dynamics between Asx and Sx phenotypes. Clusters 3 and 4 contain genes associated with equally strong Sx response, but responding in opposite directions.

On the other hand, genes in clusters 2 and 6 are associated with strong responses in both Asx and Sx individuals. Thus, the post-infection host transcription program differs between Asx and Sx subjects and, in particular, shows the existence of a non-passive physiological state in Asx hosts. In the following we provide details on the genes composing each of these clusters.

2.2.4 Intense activation of TLR and non-TLR mediated signaling in symptomatic subjects

Each of the eight clusters (Figure 2.1) represents a molecular signature with unique and contrasting temporal dynamics. We evaluated the relationship between these signatures and the dynamics of symptom development. By fitting a random-effects model, we determined the temporal correlation between gene expression profile in each cluster and standardized clinical symptom scores (18). Both positive and negative correlations were observed (Figure 2.2B). In particular, cluster 3 (A^{nc}, S^{up}_{mid}) shows the strongest positive correlation with symptom scores ($\rho = 0.77$) followed by cluster 2 ($\rho = 0.58$). Indeed, the temporal expression pattern of cluster 3 genes closely resembles the disease progression trajectory of each individual subject who developed symptoms. In comparison, the lack of symptoms in asymptomatic subjects was consistent with their nearly-constant low-level expression of this same cluster of genes (Figure 2.2A). Moreover, the molecular signature of cluster 3 attains its highest expression level at 45hpi, preceding the clinical peak symptom time (80hpi) by nearly 36 hours. This suggests the existence of a time window in which the host transcriptional changes preceded overt clinical disease development. Interestingly, the two largest clusters, cluster 4 (A^{nc}, S^{dw}_{mid}) and cluster 6 ($A^{up}_{early}, S^{dw}_{mid}$), were the most negatively correlated with the development of symptoms, $\rho = -0.54$ and $\rho = -0.41$ respectively.

A close examination of the highest ranked genes in cluster 3 (A^{nc}, S^{up}_{mid}) reveals strong

activation of a group of PRR genes that are key to innate immune responses, including Toll-like receptor 7 (TLR7) and two non-TLRs, the RNA helicases (RIG-I or DDX58) and interferon induced with helicase C domain 1 (IFIH1 or MDA-5). As a membrane-bound receptor, TLR7 is known to recognize single-stranded viral RNA (Hemmi et al., 2002). It is the most statistically significant ($p < 0.0001$) among all differentially expressed TLR genes and the only TLR gene present in cluster 3. Likewise, RIG-I and IFIH1 have been identified as cytoplasmic double-strand viral RNA sensors (Yoneyama et al., 2004, Andrejeva et al., 2004, Kang et al., 2002). Consistent with cluster 3 expression dynamics, these genes exhibit a dramatic increase, starting at 45hpi in Sx hosts (Figure 2.3A)(Fig. S8). Studies have demonstrated that the downstream signaling triggered by these PRRs converge at TBK1, resulting in direct phosphorylation of interferon regulatory factor 7 (IRF7) (Akira et al., 2006). In support of this, both TBK1 and IRF7 (Fig. S1) are found in the same cluster and have similar expression dynamics. Furthermore, cluster 3 contains a total of 11 genes that are directly involved in the TLR signaling pathway, including MyD88, TRAF6, and STAT1. When analyzed as a group, they showed an aggregated effect that is significantly associated ($p < 0.05$) with the Sx phenotype. Although this association does not reach statistical significance until 53hpi, a putative increase can be traced back to times as early as 36hpi. At 93-101hpi, this pathway attains its maximum level of significance with all 11 member genes exhibiting nearly identical expression dynamics (Figure 2.2C)(Fig. S1).

The engagement and activation of PRRs by viral ligands directly triggers many downstream signaling cascades that function in both antiviral and inflammatory responses. In line with this, cluster 3 contains many such downstream effector genes that were fully activated showing similar dynamics. In particular, a group of interferon-stimulated antiviral genes, such as MX1, OAS1, RSAD2, exhibit Sx-specific strong activation beginning at

36-45hpi (Figure 2.3B)(Fig. S9). Their increased expression persists many hours beyond symptom peak time, suggesting non-rescinding efforts in viral resolution by the host. It is noteworthy that none of the type-I interferon genes themselves is differentially expressed. Cluster 3 also contains many elements of the inflammatory branch of TLR signaling, e.g., the interferon regulatory factor 5 (IRF5). As a master regulator of the inflammatory arm of TLR7 signaling (Takaoka et al., 2005), IRF5 directly activates proinflammatory cytokine tumor necrosis factor alpha (TNF), which has been directly implicated in flu-like symptoms in many types of diseases with excessive inflammation. Together with other mediators of inflammatory response including IL15 and IL10, these genes share similar Asx-specific increasing pattern (Figure 2.3C)(Fig. S7). Of interest, we also observed the activation of sialic acid binding Ig-like lectin 1 (SIGLEC1 or Sialoadhesin) in symptomatic hosts at mid-to-late stage of infection (Figure 2.3C). As a macrophage-specific adhesion molecule, SIGLEC1 has recently been related to pro-inflammatory function of macrophages in HIV infections (Pulliam et al., 2004). Combined, the expression kinetics of cluster 3 genes constitutes a unique transcriptional signature that is closely related to the activation of multiple PRRs and the development of disease symptoms. Also notable is the correspondence between cluster 3 genes and the "acute respiratory viral" gene signature previously reported by our group (Table S2.5) (Zaas et al., 2009). We further note that many of these genes are also IFN-inducible. It is therefore necessary to confirm with biochemical experiments (e.g., p-IRF3) that their up-regulation is due to viral RNA sensing rather than a direct consequence of IFN activation.

2.2.5 A non-passive asymptomatic state is characterized by down-regulated expression of the NLRP3 inflammasome, CASP5 and the IL1B pathway

Members of cytoplasmic Nod/NACHT-LRR (NLRs) have recently been linked to pathogen pattern recognition. Originally identified in bacterial infections, this family of molecules

is important to the function of innate immunity (Chen et al., 2009, Kanneganti et al., 2006, Kobayashi et al., 2005). A recent study showed that nucleotide-binding oligomerization domain 2 (NOD2) recognizes ssRNA of both Influenza and respiratory syncytial viruses (Sabbah et al., 2009). Furthermore, activated NODs were linked to the activation of receptor-interacting serine-threonine kinase 2 (RIPK2) and subsequently nuclear factor kappa-B (NFkB) activation whereas activated NLPRs result in forming so-called inflammasome complexes, a process involving caspase-1 (CASP1) and caspase-5 (CASP5) and ultimately the release of pro-inflammatory and pro-oxidant cytokine interleukin 1-beta (IL1B) (Martinon et al., 2002, Martinon and Tschopp, 2005). Notably, influenza-induced inflammasome signaling depends on the production of reactive oxygen species (ROS) due to influenza infection (Allen et al., 2009).

The NLR-related genes are among the most highly differentially expressed genes discovered by our analysis. These genes are spread into two clusters, cluster 2 ($A^{dw}_{early}, S^{up}_{mid}$) and cluster 3 ($A^{nc}, S^{up}_{mid}$), exhibiting two distinctive temporal patterns (Figure 2.1). Residing in cluster 3, NOD1, RIPK2 and CASP1 show no significant change in asymptomatic subjects, in contrast to the dramatic increase among symptomatic individuals (Figure 2.4A)(Fig. S2). On the other hand, NOD2, NLPR3, and CASP5 are found in cluster 2. While they decrease in Asx, they increase evidently in Sx (Figure 2.4B)(Fig. S2). Another essential inflammasome gene, PYD and CARD domain containing gene (PYCARD), is located in cluster 7 ($A^{dw}_{mid}, S^{dw}_{early}$). It appears to be transiently suppressed in Sx at early phase of infection only to increase later, albeit to a lesser extent than cluster 2 genes (Fig. S2).

Finally, we observe that the expression level of IL1B, in cluster 2 ($A^{dw}_{early}, S^{up}_{mid}$), is evidently suppressed in the Asx phenotype whereas activated in the Sx phenotype (Figure 2.4C). Given the critical roles of NOD2, NLPR3, and PYCARD in the processing of

IL1B, it is tempting to speculate that their divergent expression patterns may cause lower expression of IL1B that is unique to Asx hosts. This hypothesis is supported by a new study in which Nod2-deficient mice showed reduced inflammatory response as reflected by decreased levels of TNF and IL1B in PBMC (Sabbah et al., 2009).

Of relevance to the phenotypically different expression dynamics of NLR-mediated inflammasome activation, an opposite trend is observed in two cluster 6 ($A^{up}_{early}, S^{dw}_{mid}$) genes that are related to cellular response to oxidative stress. The superoxide dismutase (SOD1) and serine/threonine kinase 25 (STK25 or SOK1) are markedly activated in Asx subjects, contrasting to the transient suppression pattern (45 – 60hpi) in Sx hosts (Figure 2.4D)(Fig. S10). As SOD1 and STK25 both have been linked to anti-oxidant/stress response and reduced concentration of ROS (Durand et al., 2005, Oda et al., 1989, Pombo et al., 1996), their sustained up-regulation in asymptomatic hosts highlights a host response signature unique to the Asx phenotype. This signature may relate to the concomitant suppression of NLRP3 and IL1B in Asx individuals. Collectively, these data reveal a phenotypically divergent expression of NLR family genes and related inflammasome signaling, which may be associated with the host anti-oxidant defense system.

2.2.6 Distinct temporal kinetics of JAK-STAT pathway and SOCS family genes reveals a potential method of viral control in asymptomatic hosts

A hallmark of host recognition of viral RNA is the activation of Janus kinase-signal transducer and activator of transcription (JAK-STAT) pathway, which is crucial for the antiviral function of interferons. However, such activation is tightly controlled to limit the possibility of over-stimulating inflammatory cytokine-receptor signals. As an integral component of the JAK-STAT pathway, the family of suppressor of cytokine signaling (SOCS) proteins have recently been reported to negatively regulate the response of immune cells to cytokine signals (Yasukawa et al., 2000). Using pathway analysis, we

detected significantly distinct JAK-STAT signaling dynamics ($p < 0.05$), involving two different sets of SOCS genes. The first set included SOCS1 and SOCS3 from cluster 2 ($A^{dw}_{early}, S^{up}_{mid}$) while the second group consists of SOCS2 and SOCS5 from cluster 6 ($A^{up}_{early}, S^{dw}_{mid}$). The expression of SOCS1 and SOCS3 declines at early time points among Asx but strongly increases among Sx (Figure 2.5A)(Fig. S11). Growing evidence suggests that SOCS1 and SOCS3 are important inhibitory modulators in limiting the inflammatory effect of IFN signaling during viral infection (Rothlin et al., 2007, Pothlichet et al., 2008). Our data supports such a protective role of SOCS1 and SOCS3 given their much higher levels of expression during late infection phase (45hpi onward).

Consistent with cluster 6 but contrasting with the cluster 2 expression pattern (Figure 2.1), SOCS2 and SOCS5 exhibits expression dynamics that clearly differed from that of SOCS1 and SOCS3. Starting from the early infection stage ($\sim$12hpi), SOCS2 and SOCS5 show marked increasing trend in Asx and this trend persists throughout the entire infection period (Figure 2.5B)(Fig. S11). In contrast, their expression diminishes in Sx, especially between 45hpi and 69hpi. A recent study showed that the anti-inflammatory actions of aspirin-induced lipoxins depend upon the function of SOCS2 (Machado et al., 2006). Highly expressed in lymphoid organs, SOCS5 was hypothesized to be important for the generation of Th1 responses by repressing IL-4-induced signals that promote Th2 differentiation (Seki et al., 2002). In addition, we observed a significant positive association of interleukin 7 (IL7) and STAT4 (Figure 2.5B). Of these, STAT4 transduces IL12 and IFN-A cytokine signals in T cells and monocytes (Korman et al., 2008) whereas IL7 is critical for proper T cell response and expansion during viral infection (Ma et al., 2006, Schluns and Lefrancois, 2003, Sun et al., 2006). Taken together, the distinct expression patterns of SOCS family genes and related JAK-STAT signaling suggest possible early involvement of Th1-type adaptive immune response in asymptomatic hosts with no sign

of excessive inflammation.

2.2.7 Ribosomal protein synthesis is enhanced in asymptomatic subjects as compared to symptomatic subjects

In addition to expression changes in magnitude, a large number of significant genes in clusters 2 ($A^{dw}_{early}, S^{up}_{mid}$) and 6 ($A^{up}_{early}, S^{dw}_{mid}$) also exhibit directional contrast between two phenotypes. As the largest cluster with a total of 1,326 member genes, cluster 6 contains genes with expression profiles similar to those of SOCS2 and SOCS5. Functional pathway analysis reveals that many of these genes are implicated in innate immune response (Table S2.1). In particular, we found an unusual saturation of genes related to ribosomal protein synthesis. Out of 47 significant genes in this pathway, 35 (76%) of them are located in cluster 6 ($p < 0.0001$). Together, these 35 genes correlate positively with Asx phenotype ($p < 0.05$) and their expression increases over the course of the study (Figure 2.5C). Such association emerges at 45-53hpi and peaks at 60hpi, at which point every one of the 35 genes becomes highly expressed. Individually, all genes showed increased expression trend (Fig. S4). Similar to other genes in this cluster, the trend can be seen at as early as 5hpi and as late as 108hpi. In contrast, symptomatic subjects show sustained down-regulation of the same set of genes, with lowest expression level at 60hpi. This decreasing trend continues until $\sim$ 84hpi, which coincides with the peak symptom time observed in symptomatic subjects (Figure 2.2). It is notable that down-regulation of ribosomal proteins has been reported in measles-infected dendritic cells (Zilliox et al., 2006). Given the markedly contrasting trends observed between Asx and Sx phenotypes, we conclude that asymptomatic hosts responded differently to the viral insult, exhibiting enhanced cellular protein biosynthesis relative to symptomatic subjects.

2.2.8 Unsupervised detection of disease signature with Bayesian Liner Unmixing (BLU)

To supplement the supervised methods implemented above that drew contrasts between Asx and Sx responses using the knowledge of ultimate disease outcome, we also applied an unsupervised factor analysis to explore associations between gene expression responses over time and over subjects. BLU is a signal processing algorithm originally developed for unmixing composite spectra in hyperspectral imaging (Dobigeon et al., 2009). Blind-folded to the clinically determined phenotype labels, BLU operates on the expression data matrix alone and detects molecular signatures of symptomatic disease by discovering groups of genes, called factors, which best explain common gene profiles contributing to the overall gene expression response. Examination of the factor scores, which take values between zero and one, reveal that one of the factors, called the leading enriched factor, is highly correlated with symptom severity (Figure 2.6A and B). Of the genes in this factor, 70% can be mapped to SOM cluster 3 ($p < 0.05$; Fisher's exact test). This is in strong concordance with the high correlation between cluster 3 (A^{nc}, S^{up}_{mid}) expression profile and temporal disease progression (Figure 2.2A and B). Thus BLU independently validates the results of supervised SOM co-clustering analysis. When post hoc re-arranged according to phenotypes and time, the expression signature detected by BLU clearly distinguished Sx from Asx subjects and delineated pre-onset from post-onset phases of the symptomatic infection. The image of the BLU factor score bears striking resemblance to the standardized clinical symptom observation matrix (Figure 2.6A and B). This confirms the power of BLU as an unsupervised analysis method in the de novo discovery of molecular signatures underlying symptomatic disease.

2.2.9 Early and late phase disease stratification using a logistic boosting model

We next used the factors discovered by BLU to identify gene discriminants that best differentiate between early and late phases of host response. Based on the BLU results, we assigned individual expression profiles into four classes, namely regions 1 - 4 (Figure 2.6A). Of these, class 1 includes all samples collected prior to inoculation. Class 2 samples are from Asx subjects post-inoculation whereas class 3 samples are from Sx subjects prior to symptom onset time. Class 4 are post-onset samples of Sx subjects that include samples at their peak symptom times. Essentially, such discretized four-class designation encapsulates distinct risk levels of four intrinsic disease states — uninfected (class 1), infected with low-risk for symptom development (class 2), infected with high-risk for symptom development (class 3), and infected with overt symptoms (class 4). By fitting a logistic boosting model to the data in each pair of classes (Bhlmann and Hothorn, 2007), we identified different gene sets that are capable of discriminating samples between any pair of the four disease states (Table S2.4). Overall, classifiers between each class pair achieve nearly-perfect true positive rate at 5% average false positive rate (Figure 2.6C). Not unexpectedly, the model has the most difficulty discriminating between the uninfected versus high-risk class pair (1vs3). At average false positive rate of 10%, the model can classify with 90% accuracy rate but with a rather wide 95% confidence interval (78%, 100%).

The numbers of discriminating genes selected by the models range from as few as 6 for the high-risk versus overt symptom class pair 3vs4 to as many as 19 for the low-risk versus high-risk class pair 2vs3 (Table S2.4). Many of these genes are relevant to modulation and signaling in innate immune responses. For example, IFI44L, IFI27, GBP1, and OAS1 are directly involved in type-I IFN's antiviral response and appear as highly discriminating between (2vs4), (1vs4), and (3vs4). Similarly, gene DDX17 that discriminates class pair

(3vs4) has been shown to be important for optimal function of zinc-finger antiviral protein in limiting the accumulation of viral mRNA in the cytoplasma (Chen et al., 2008). The gene for complement component 3a receptor 1 (C3AR1) is the second highest ranked gene for discriminating between (2vs3), suggesting an important role by the complement system activation against viral infection. For the same classifier, resistin (RETN) appears to be even more discriminatory than C3AR1. It has been reported that RETN induces insulin resistance partly via induction of SOCS3 expression in HepG2 cells (Luo et al., 2009). There are reports that SOCS1 and SOCS3 block insulin signaling by facilitating insulin substrate receptors (IRS 1 and IRS2) for degradation (Rui et al., 2002). This is consistent with our finding that IRS2 is a highly discriminatory gene for class pairs (1vs3), (1vs4), and (3vs4). Studies have shown that IFNA activates both IRS1 and IRS2 independent of the JAK-STAT pathway and transduces antiviral signals (Platanias et al., 1996, Uddin et al., 2000, 1995). FOXO3 was also implicated in insulin receptor signaling (Kino and Chrousos, 2004, Kino et al., 2005). Other than IRS2, genes MS4A1, C3AR1, HBG2 have also been associated with infectious diseases. Taken together, these suggest an underappreciated role by IRS pathway in innate immunity. Other genes in Table S2.4 are less well-known, especially those selected for uninfected versus high-risk (1vs3), uninfected versus low-risk (1vs2), and low-risk versus high-risk (2vs3). On the basis of this analysis we believe that their function in immune response merits further investigation.

2.3 DISCUSSION

Pathogenic influenza A viral infection is a complex and dynamic process that involves various components of the host immune system at different stages of infection in response to virus-induced physiological changes. Dissecting the temporal host response to invading viruses and subsequent symptomatic disease process therefore provide insight into the

pathogenesis of influenza A infection and related host factors. Equally important is to understand the complexity of the host response in individuals who are exposed but effectively contain the virus and avoid symptomatic disease.

In this study, we presented key transcriptional differences between asymptomatic and symptomatic host responses, and highlighted the active state (on a gene transcription level) of viral control. We showed that the transcriptional patterns in symptomatic hosts directly correlate to the development of clinical symptoms over time. As mounting evidence has established the role of various PRRs in sensing viral components of influenza viruses, our results confirm the concurrent activation of all known classes of PRRs and their signaling cascades by influenza viruses in human challenge models. In contrast, asymptomatic hosts showed not only an absence of such activation, but also negative regulation of related inflammatory signals, especially in the case of NLRP3 and NOD2. This directly corresponds to their lack of clinical apparent symptoms. Since these PRRs serve as the link between the innate immune function to the adaptive immunity, it is likely that adaptive immune response was never fully activated in asymptomatic individuals.

It has long been postulated that multiple PRRs represent a functional redundancy of host defense and that there exists signaling crosstalk among them, stimulating similar cytokine profiles that are both pro-inflammatory and pro-oxidant (Martinon and Tschopp, 2005). The simultaneous and continued activation of these PRRs in symptomatic hosts may indicate undesired over-reaction by the host immune system. Without proper control, relentless PRR-stimulated signals may do more harm than good to the host. For example, abnormally expressed NOD2 has been implicated in inflammatory bowel disease (Abraham and Cho, 2009, Hugot et al., 2001). Conversely, a study on chronic arthritis has shown that Nod2 gene-deficiency resulted in reduced joint inflammation and increased protection against early cartilage damage in mice (Joosten et al., 2008). Our results pro-

vide additional evidence but further investigation is needed to elucidate the aggregated consequence of co-activated PRRs and to study the potential benefit in limiting inflammation during infection. Of note, a recent study on 3,921 lab-confirmed influenza associated hospitalizations linked the anti-inflammatory function of HMG-CoA reductase inhibitors (or statins) to favorable clinical outcome in infectious disease caused by influenza viruses (Vandermeer et al., 2009).

Importantly, activation of the inflammasome and production of pro-inflammatory cytokines have been associated with increased level of oxidative stress during viral infection (Kofler et al., 2005, Floyd et al., 1999, Schwarz, 1996). A recent report showed in mouse model that Nlrp3 inflammasome activation depends on reactive oxygen species (ROS). Furthermore, the inhibition of ROS induction abolished IL1B production during influenza infection (Allen et al., 2009). It is intriguing that our data shows a temporal Asx-specific upregulation and Sx-specific suppression of SOD1 and SOK1. This coincides with the observed negative correlation between these genes and NLRP3. Since SOD1 and SOK1 are capable of reducing ROS and of suppressing oxidative stress (Durand et al., 2005), their increased expression in asymptomatic hosts may play a role in negatively regulating the NLRP3 expression and inflammasome signaling. In support of this hypothesis is a study on the efficacy of antioxidant therapy found that pyran polymer-conjugated SOD1 protected mice against potentially lethal influenza virus infections (Oda et al., 1989). Together, these results provide further evidence for a protective role of antioxidants such as SOD1 and SOK1. Their increased mRNA expression may constitute an effective antiviral mechanism by which aberrant immune responses are avoided in asymptomatic hosts.

Shutting down protein synthesis helps control infection by inducing apoptosis of infected cells (Castelli et al., 1997, Samuel et al., 1997, Clemens and Elia, 1997). Our findings are consistent with this as we observe marked expression decrease of protein

biosynthesis-related genes in symptomatic hosts (Fig. S4) at mid-to-late stages, a result likely due to increased viral replication as evidenced by elevated expression of protein kinase R (PKR) (Fig. S3). However, a surprising finding in our study is the sustained increase in the expression of as many as 35 ribosomal proteins in only asymptomatic subjects (Fig. S4). This suggests that enhanced protein synthesis help hosts improve viral clearance and prevent over-stimulation of inflammatory responses. Confirmation of this hypothesis may provide insight into protein targets for therapy to elicit effective host antiviral capacity.

It is estimated that asymptomatic infections account for 30 to 50 percent of seasonal flu cases (Carrat et al., 2008), which is consistent with the attack rate in our study. Since both asymptomatic and symptomatic subjects were challenged under the same protocol, this raises a critical question concerning the nature of the factors that lead to subclinical infections. Our results indicate that genes in cluster 2, 4, and 6 may prove most relevant in this regard. From the standpoint of disease control and pandemic prevention, the asymptomatic hosts represent a population that is at least as important as their symptomatic counterpart since both were exposed to viral pathogens. Likewise, pre-symptomatic subjects can benefit from appropriate early intervention. Having the capacity to stratify subjects into different risk groups according to their disease states may be a key to implementing effective therapeutic measures. Using no prior clinical label information, our risk-stratification model suggests that host molecular signatures may be used for this purpose with potential prognostic value at the point-of-care.

In conclusion, to our knowledge this study presents for the first time a systematic analysis of full temporal spectrum of pathogen-elicited host responses during influenza viral infection. In addition to differentiating host response mechanisms of asymptomatic and symptomatic subjects, we have introduced new concepts and novel applications of modern

statistical modeling. Well-designed and meticulously organized, this multi-institutional collaborative work represents by far the most extensive in vivo human challenge study on influenza viruses. Such challenge study allows the sophisticated human immune system to be examined in a unified manner where concerted immune responses are studied as a whole. Combined with key clinical parameters, our results offer an opportunity to look beyond individual signaling events and into their collective modular effects on symptomatic disease pathogenecity. Understanding the timing of various immune response events in vivo will enable us to assess their biological and clinical relevance to disease progression. Our previous work (Zaas et al., 2009) identified a molecular signature in blood that is characteristic of upper respiratory viral infection involved in several pathogens. We speculate that the present findings will extend to other type of respiratory viruses as well and that infection result in the selection of specific molecular pathways at various infection stages. These pathways serve to eradicate viremia while minimizing host debilitation and post-infection sequelae. Highly selected and stereotyped responses are generalizable to many viral upper respiratory infections.

2.4 MATERIALS AND METHODS

Human Influenza viral challenges

A healthy volunteer intranasal challenge with influenza A/Wisconsin/67/2005 (H3N2) was performed at Retroscreen Virology, LTD (Brentwood, UK) in 17 pre-screened volunteers who provided informed consent. All volunteers were influenza A antibody negative at pre-inoculation testing. On day of inoculation, a dose of 106 TCID50 Influenza A manufactured and processed under current good manufacturing practices (cGMP) by Bayer Life Sciences, Vienna, Austria) was inoculated intranasally per standard protocol at a varying dose (1:10, 1:100, 1:1000, 1:10000) with four to five subjects receiving each

dose. Subjects were not released from quarantine until after the 216th hour. Blood and nasal lavage collection continued throughout the duration of the quarantine. All subjects received oral oseltamivir (Roche Pharmaceuticals) 75 mg by mouth twice daily prophylaxis at day 6 following inoculation. All patients were tested negative by rapid antigen detection (BinaxNow Rapid Influenza Antigen; Inverness Medical Innovations, Inc) at time of discharge. All exposures were approved by the relevant institutional review boards and conducted according to the Declaration of Helsinki.

Case definitions

Symptoms were recorded twice daily using standardized symptom scoring (2). The modified Jackson Score requires subjects to rank symptoms of upper respiratory infection (stuffy nose, scratchy throat, headache, cough, etc) on a scale of 0 – 3 of "no symptoms", "just noticeable", "bothersome but can still do activities" and "bothersome and cannot do daily activities". For all cohorts, modified Jackson scores were tabulated to determine if subjects became symptomatic from the respiratory viral challenge. A modified Jackson score of $>= 6$ over the quarantine period was the primary indicator of successful viral infection (Turner, 2001) and subjects with this score were denoted as "Symptomatic" (Sx). Viral titers from daily nasopharyngeal washes were used as corroborative evidence of successful infection using quantitative PCR (Table S2.2) (Turner, 2001, Barrett et al., 2006, Jackson et al., 1958). Subjects were classified as "Asymptomatic" (Asx) if the Jackson score was less than 6 over the five days of observation and viral shedding was not documented after the first 24 hours subsequent to inoculation. Successful inoculation in Asx hosts was further validated by analysis of multimodal data including serum neutralizing antibody and haemagglutination inhibition titers. For additional evidence see discussion in Supplementary Notes. Standardized symptom scores were tabulated at the end of each study to determine attack rate and time of maximal symptoms (time "T"). We note that the

clinical disease is mild and represents early stage infection (only a single fever was observed). Immune activation assays (such as antibody response) over the full time course of the challenge study were not available for our analysis. However, the reported high correlation between self-reported symptom severity scores (2.2) and the interferon/inflammatory response pathway (Cluster 2) is suggestive that genes in this pathway are themselves good severity markers.

Biological sample collections

During the challenge study, subjects had samples taken 24 hours prior to inoculation with virus (baseline), immediately prior to inoculation (pre-challenge) and at set intervals following challenge: peripheral blood for serum, peripheral blood for RNA PAXgene™, nasal wash for viral culture/PCR, urine, and exhaled breath condensate. Peripheral blood was taken at baseline, then at 8 hour intervals for the initial 120 hours and then 24 hours for the remaining 2 days of the study. For all challenge cohorts, nasopharyngeal washes, urine and exhaled breath condensates were taken at baseline and every 24 hours. Samples were aliquoted and frozen at $-80°$C immediately.

RNA purification and microarray analysis

RNA was extracted at Expression Analysis (Durham, NC) from whole blood using the PAXgene™96 Blood RNA Kit (PreAnalytiX, Valencia, CA) employing the manufacturer's recommended protocol. Hybridization and microarray data collection was performed using the Human Genome U133A 2.0 Array (Affymetrix, Santa Clara, CA) and expression profiles were analyzed using standard methods.

Statistical analysis

Temporal gene expression was analyzed using EDGE (Storey et al., 2005). Co-clustering of the significant genes found by EDGE was performed using Self-Organizing Map (Kohonen, 1995). Biological pathway enrichment analysis was performed using Ingenuity?

Pathway Analysis (IPA). We implemented the non-parametric Jonckheere-Terpstra (JT) method (Hero and Fleury, 2004) to test monotonicity of the expression patterns of individual gene clusters. Briefly, the JT test was applied independently to each cluster and configured to test the null hypothesis that there exists no monotonic trend in the temporal change of gene expression. This test was performed separately for each one of two phenotypes separately. The resulted p-values were adjusted for multiple comparisons with Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

To identify canonical gene pathways in each SOM cluster that were highly associated with disease phenotypes, we applied Globaltest (Goeman et al., 2004) using the pathway definition in MsigDB database (v2.5) (Subramanian et al., 2005) that include both pathway components and targets. We assessed the correlation between clinically determined symptom scores and the temporal gene expression of SOM clusters using standard linear mixed model regression. The correlation (R value) was estimated using a signed coefficient of determination (Faraway, 2004, Hssjer, 2008).

The Bayesian Linear Unmixing (BLU) (Dobigeon et al., 2009) factor analysis was used to detect disease signature. Unlike our implementation of EDGE, SOM and Globaltest, BLU is an unsupervised method requiring no prior class information. Like other unsupervised Bayesian factor analysis methods, BLU finds a decomposition of the data matrix Y, here a p by n matrix of abundances of the p mRNA transcripts for each of n gene expression profiles, into a matrix product MA where each column of M is a factor and each column of A is a set of factor loadings corresponding to individual factors in M for a given chip:

$$Y = MA + N \tag{2.1}$$

In essence, BLU estimates two matrix valued latent variables M and A, whose product

best approximates the most important information contained in the observation Y while minimizing the residual model fitting error (denoted as N in the formula above) with latent variable order selection according to an hierarchical Bayesian model. However, unlike other factor analysis, BLU decomposes the data into relative proportions such that the columns of M and the columns of A are non-negative and the columns of A sum to one. Intuitively, a BLU-discovered factor can be viewed as a gene expression profile, whose amplitudes represent the relative contribution of each gene present in that factor, and the factor loadings are the proportions of these factors that are present in each chip. Such positivity constraints aid in interpretation and are natural in gene microarray analysis as the expression intensity measurements of genes are always non-negative. Before applying BLU, we first performed standard ANOVA method to screen genes with most significant time-varying expression. At FDR q-value < 0.01 significance level, we obtained a list of 935 genes. This pre-selection of genes is completely independent of the Sx and Asx labeling and eliminates all genes except those having the strongest temporal dynamics. Subsequently, BLU was run on this smaller set of gene expression profiles of 935 genes and extracted a total of three major BLU factors. Based on one of the factors, called the enrichment factor, we clustered the factor scores into 4 classes as explained above and shown in Figure 2.6A.

The early and late phase disease stratification model was constructed using standard LogitBoost (Bhlmann and Yu, 2003) classification model with the aforementioned 4-class designation derived from BLU. The Akaike Information Criterion was used for model selection. Furthermore, we employed a standard bootstrap-resampling technique Efron, 1979 to post hoc cross-validate the classifiers, assessing the performance of the model and predictor genes. The training set (2/3 of data) was used to construct the boosting ensemble while the test set (1/3 of data) was used for cross-validation and testing. We report the

classifier performance using ROC curves constructed from applying the trained classifier to the test set and chose the area under the curve (AUC) as a measure of classification performance. We computed 95% confidence intervals based on the bootstrap re-sampled data, for both true positive prediction and false negative prediction at each threshold point on the ROC.

More detailed information about the materials and methods are available in Supplementary Information.

2.5 ACKNOWLEDGEMENTS

This research was partially supported by a grant from the DARPA PHD Program. We gratefully acknowledge Anthony Moody, Elizabeth Ramsberg and Micah T. McLaine of Duke University for their helpful comments on earlier versions of this paper.

2.6 SUPPLEMENTARY MATERIALS

Human Influenza viral challenges A healthy volunteer intranasal challenge with influenza A A/Wisconsin/67/2005 (H3N2) was performed at Retroscreen Virology, LTD (Brentwood, UK) in 17 pre-screened volunteers who provided informed consent. All volunteers were influenza A antibody negative at pre-inoculation testing. On day of inoculation, a dose of 10^6 $TCID_{50}$ Influenza A manufactured and processed under current good manufacturing practices (cGMP) by Bayer Life Sciences, Vienna, Austria) was inoculated intranasally per standard methods at a varying dose (1:10, 1:100, 1:1000, 1:10000) with four to five subjects receiving each dose. Subjects were not released from quarantine until after the 216th hour. Blood and nasal lavage collection continued throughout the duration of the quarantine. All subjects received oral oseltamivir (Roche Pharmaceuticals) 75 mg by mouth twice daily prophylaxis at day 6 following inoculation. All patients were neg-

ative by rapid antigen detection (BinaxNow Rapid Influenza Antigen; Inverness Medical Innovations, Inc) at time of discharge.

Case definitions Symptoms were recorded twice daily using standardized symptom scoring (Jackson et al., 1958). The modified Jackson Score requires subjects to rank symptoms of upper respiratory infection (stuffy nose, scratchy throat, headache, cough, etc) on a scale of 0–3 of "no symptoms", "just noticeable", "bothersome but can still do activities" and "bothersome and cannot do daily activities". For all cohorts, modified Jackson scores were tabulated to determine if subjects became symptomatic from the respiratory viral challenge. A modified Jackson score of ≥ 6 over the quarantine period was the primary indicator of successful viral infection (Turner, 2001) and subjects with this score were denoted as "SYMPTOMATIC". Viral titers from daily nasopharyngeal washes were used as corroborative evidence of successful infection using quantitative and quantitative PCR (Jackson et al., 1958, Turner, 2001, Barrett et al., 2006) (Table S2.2). Antibody neutralization assays were also performed to corroborate this hypothesis (see discussion in the section entitled "Supplementary Notes" below). Subjects were classified as "ASYMPTOMATIC" if the Jackson score was less than 6 over the five days of observation and viral shedding was not documented after the first 24 hours subsequent to inoculation. Standardized symptom scores tabulated at the end of each study to determine attack rate and time of maximal symptoms (time "T").

Biological sample collections During challenge study, subjects had the following samples taken 24 hours prior to inoculation with virus (baseline), immediately prior to inoculation (pre-challenge) and at set intervals following challenge: peripheral blood for serum, peripheral blood for RNA PAXgene^{TM}, nasal wash for viral culture/PCR, urine, and exhaled breath condensate. Peripheral blood was taken at baseline, then at 8 hour

intervals for the initial 120 hours and then 24 hours for the remaining 2 days of the study. For all challenge cohorts, nasopharyngeal washes, urine and exhaled breath condensates were taken at baseline and every 24 hours. Samples were aliquoted and frozen at -80^{o}C immediately.

RNA purification and microarray analysis RNA was extracted at Expression Analysis (Durham, NC) from whole blood using the PAXgene^{TM} 96 Blood RNA Kit (PreAnalytiX, Valencia, CA) employing the manufacturer's recommended protocol. Hybridization and microarray data collection was performed at Expression Analysis (Durham, NC) using the GeneChip Human Genome U133A 2.0 Array (Affymetrix, Santa Clara, CA). Raw gene expression profiles were further preprocessed using robust multi-array analysis (Bolstad et al., 2003) with quantile normalization and probe-level signals were summarized in log base 2 scale. We selected a custom Chip Definition File (CDF) version 10 for more accurate probe mapping to genome (Dai et al., 2005).

Differential gene expression analysis between symptomatic versus asymptomatic Temporal gene expression was analyzed using EDGE (Storey et al., 2005). Briefly, a gene-wise natural cubic smoother was fit to the temporal expression profiles for each individual subject. To prevent overfitting, we fixed the number of spline knots to four such that there were at least three time points available for each knot. Subsequently, a group-wise cubic spline was summarized for asymptomatic and symptomatic subjects, respectively. The spline-fitted gene expression was compared to test the null hypothesis that there is no significant difference between asymptomatic and symptomatic phenotypes. Statistical significance is assessed using F-test with simultaneous multiple-testing FDR control. The final set of candidate genes was selected as significantly differentially expressed between Sx and Asx with FDR adjusted p-value $< 1\%$. Of note, the samples collected at baseline

time ($-$24hpi) were not directly used in aforementioned differential expression analysis. Instead, they were used as quality assurance to ensure that none of the genes deemed as significant was differentially expressed relative to pre-challenge (0hpi) samples, using a standard paired t-test.

Co-clustering significant genes using Self-Organizing Map (Kohonen, 1995) The self-organizing map was used to cluster the complex high-dimensional temporal gene profiles of each phenotype (Figure 2.1). Like other metric clustering algorithms, SOM performs dimensionality reduction for visualization of complex relationships and trends by preserving the topological and metric relationships between profiles (Kohonen, 1995). In our analysis, we aim to place, in the same region of a 2D grid layout, those genes that are similar in temporal expression profiles, measured by their Euclidean distances. In consistency with differential expression analysis, a natural cubic spline was fitted on the temporal expression values of each gene using smoothing spline method (Hastie and Tibshirani, 1990) prior to clustering. Again, we fixed the degrees of freedom at four, yielding a more conservative model fit in terms of the amount of smoothing. This is in concordance to the parameter setting we used in determining significance level of genes in EDGE (Storey et al., 2005). The fitted values were subsequently z-score normalized prior to clustering. As there exists no gold standard in choosing the "best" map configuration among all possible maps, we proposed an analytical selection procedure in which we balance the complexity of the map (number of prototypes), the distances between genes and their prototypes, and the silhouette values (Rousseeuw, 1987) of genes (measure of the closeness of a gene to its within-cluster neighbors and to its neighbor-cluster). This resulted in a 4×2 hexagonal grid of prototypes or clusters. Each prototype's representative centroid was initially chosen from genes at random. The initial neighborhood size was set

such that each neighborhood contained 25% of prototypes. The total number of iterations was chosen such that each gene was repeatedly presented to the map 50 times. Each gene was clustered into a prototype to which it is closest in Euclidean (or $L2$) distance measurement. The average expression values of each individual clusters and corresponding $\pm$ two standard deviations were plotted in Figure 2.1c. We also estimated the centroids of each SOM cluster and correspond 95% confidence intervals using nonparametric bootstrap method without assuming normality (Efron, 1979). The derived centroids of clusters are almost identical and the confidence limits of these centroids are much smaller than and completely covered by the $\pm$ two standard deviations shown in Figure 2.1C.

Polar plot visualization of temporal expression pattern of a cluster The polar plots (Figureflu:somA) provide a different visualization of the differences between temporal gene expression profiles for Asx and Sx phenotypes. Each polar plot depicts the expression pattern shared by genes of a SOM cluster. Within a plot, the temporal expression of Asx resides on the top portion of the circle while Sx expression occupies the bottom half. Each phenotype's expression values are placed in time sequence, increasing in the counterclock-wise direction, inside its own half circle. Consequently, the expression profiles of Asx and Sx at any given time point can be compared at opposite ends of a radial line passing through the polar origin. Such symmetric arrangement facilitates visual examination of contrasts in phenotypic gene expression patterns. We emphasize that it is not adequate to only look at one phenotype alone or the ratio of Sx/Asx expression values. This is because of the fact that both Asx and Sx undergo significant changes in gene expression profiles, a consequence of universal protective immune response. In Figure 2.1b we show heatmaps for the top 5 genes from each SOM cluster having the most significant differential expression. The expression values from different time points are aligned

horizontally. Although this type of visualization arrangement is in line with traditional clustering results, we can see that it is less convenient to contrast the expression values of two phenotypes at any given time point as it requires visual search through the horizontal time line. The reader may find that segment plots of SOM clusters add interpretability to the heatmaps of temporal expression patterns, allowing more direct simultaneous comparisons between particular time points and phenotypes.

Biological pathway enrichment analysis To identify biological pathways that are enriched in each individual clusters, we used Ingenuity© Pathway Analysis (IPA) tool and queried their proprietary knowledge database of functional interactions between molecules. The representative pathways were shown in Table 2.1.

Testing monotonicity of expression pattern of clusters We used the non-parametric Jonckheere-Terpstra (JT) method (R package SAGx by Per Broberg) (Lehmann, 1975, Hero and Fleury, 2004) to quantitatively test whether a monotonic increase or decrease trend exists in a cluster expression centroid. For each of the eight clusters, the JT test was carried out to test the null hypothesis that there exists no monotonic trend in changes of gene expression over time. The alternative hypotheses are that the median gene expression of later time points was higher or lower than that of earlier time points:

$$M_t^i < M_{t+1}^i < \cdots < M_{t+n}^i \text{ or } M_t^i > M_{t+1}^i > \cdots > M_{t+n}^i \tag{2.2}$$

where M_t^i is the centroid of gene expression of cluster i at time t.

This test was performed for each one of two phenotypes separately with significance measures (p-values) shown in Table S2.3. The p-values were further adjusted with Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to correct for multiple hypothesis testing.

We note that JT test is different from EDGE test in that JT tests any monotonicity trend in temporal gene expression subject to the constraint that the change in expression has to be either monotonically increasing or monotonically decreasing. On the other hand, EDGE tests any phenotypic difference in expression over all time points. Therefore EDGE has more power in detecting differential expression over time without requiring monotonicity whereas JT has more power in detecting monotonic temporal expression within each Asx and Sx phenotype without requiring differential expression.

Associating disease phenotypes with canonical biological pathway To identify the canonical gene pathways in each SOM cluster that are highly associated with disease phenotypes, we applied Globaltest (Goeman et al., 2004) using all pathways included in MsigDB database (v2.5) (Subramanian et al., 2005). Briefly, all significant genes from a SOM cluster were first mapped onto individual MsigDB pathways. Then we carried out the group testing procedure outlined in (Goeman et al., 2004) to test the association between each pathway and phenotypes using logistic regression via hierarchical generalized linear model fitting. Specifically, the model is formulated as

$$\mathbf{E}(Y_i|\beta) = \text{logit}^{-1}(\alpha + \sum j = 1^m x_{ij}\beta_j) \tag{2.3}$$

$$\boldsymbol{H_0} : \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

where $Y_i \in (0,1)$ for the Asx and Sx phenotypes, respectively, and x_{ij} is the expression value for j-th gene of i-th subject and $j \in C$ for a predetermined pathway C.

The null hypothesis is simply that all β's equal to zero, corresponding to no correlation between the pathway C (includes all j genes) and disease phenotypes. The significance of association was assessed using permutation test and we further adjusted the p-values to account for multiple testing according to the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). This analysis was conducted for each SOM cluster individually

and for all 16 time points in the challenge study where the gene expression profiles were examined.

Correlating disease symptom scores with temporal expression values of clusters

We estimated the correlation between clinically determined symptom scores and the temporal gene expression of SOM clusters using a standard linear mixed model. Specifically, for each one of the 10 categories of symptom scores, we regressed the scores onto the expression value vector of every one of eight SOM clusters, separately, with a random-effects term accounting for within-subject temporal correlation. For each symptom and cluster prototype the model is

$$\mathbf{y_i}^{(t)} = \mathbf{x_i}^{(t)}\beta + \mathbf{1}b_i + \boldsymbol{\varepsilon}_\mathbf{i} \tag{2.4}$$

where

- $\mathbf{y_i}^{(t)}$ is a $t \times 1$ vector of measures on a symptom category for subject i over t time points
- $\mathbf{x_i}^{(t)}$ is a $t \times 1$ vector of average gene expression of a cluster for subject i over t time points
- β is a scalar coefficient of fixed-effects of expression values
- b_i is a scalar coefficient of random-effects for subject i

The goodness-of-fit of the mixed model was assessed using the *signed coefficient of determination* (Faraway, 2004, Hssjer, 2008) defined as $r_i = \text{sign}(\hat{\beta})\sqrt{\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}}$ where

- r_i is a correlation coefficient taking values between minus and plus one.
- $\hat{\beta}$ is the estimated fixed-effects coefficient in the fitted model (see above).

- $\sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2$ is the *unsigned coefficient of determination*, which is equivalent to the squared correlation between the response variable and the predictor in the fitted mixed model regression.

The quantity r_i is an approximation to the correlation coefficient between symptom scores and cluster expression values.

Unsupervised detection of disease signature with Bayesian Linear Unmixing (BLU)

All the above analysis methods involved the use of clinical labels, namely Sx and Asx. BLU extends and validates the SOM results using an unsupervised learning technique that does not involve these labels. In particular, without the benefit of ground truth clinical symptom scores, BLU discovered many of the same genes as SOM and specified a gene expression factor that separated asymptomatic from symptomatic subjects. Within the symptomatic group, the genes that principally contributed to this factor exhibited temporal expression patterns correlated highly with symptom patterns and differentiate the time samples between pre- and post-onset time of the inflammatory response.

Originally developed for unmixing composite spectra in hyperspectral imaging (Dobigeon et al., 2009), Bayesian Linear Unmixing (BLU) is a Bayesian factor analysis method. Like other unsupervised Bayesian factor analysis methods, BLU finds a decomposition of the data matrix $\mathbf{Y}$, here a $\boldsymbol{p}$ by $\boldsymbol{n}$ matrix of abundances of the $\boldsymbol{p}$ mRNA transcripts for each of $\boldsymbol{n}$ gene expression profiles, into a matrix product $\mathbf{MA}$ where each column of $\mathbf{M}$ is a factor and each column of $\mathbf{A}$ is a set of factor loadings corresponding to individual factors in $\mathbf{M}$ for a given chip:

$$\mathbf{Y} = \mathbf{MA} + \mathbf{N} \tag{2.5}$$

In essence, BLU populates the matrix $\mathbf{Y}$ as with all the gene expression profiles (microarray chips) over subject, time, and probe index. It estimates two latent variables $\mathbf{M}$

and **A**, whose product best approximates the most important information contained in the observation **Y** while minimizing the residual model fitting error (denoted as **N** in the formula above). However, unlike other factor analysis methods, BLU decomposes the data into relative proportions, the columns of **M** and the columns of **A** are non-negative and the columns of **A** sum to one. More specifically, each BLU-discovered factor can be viewed as a gene expression profile, whose amplitudes represent the relative contribution of each gene present in that factor, and the factor loadings are the proportions of these factors that are present in each chip. Such positivity constraints are natural in gene microarray analysis as the expression intensity measurements of genes are always non-negative.

Before applying BLU, we first performed ANOVA to screen genes with most significant time-varying expression. At FDR q-value < 0.01 significance level, we obtained a list of $\boldsymbol{p} = 935$ genes. This pre-selection of genes is completely independent of the Sx and Asx labeling and eliminates all genes except those having the strongest temporal dynamics. Subsequently, BLU was run on this smaller set of gene expression profiles of 935 genes. A total of three major BLU factors were extracted and the principal factor that is most highly correlated with disease symptom scores is shown in Figure 2.6A. A simple thresholding of the score of this principal factor then clusters all gene expression chips into two groups: those with strong factor scores and those without. Each row of Figure 2.6A shows the factor score associated with each gene expression chip with respect to the principal factor. The chips with the smallest scores can be found in asymptomatic and pre-symptom regions, 2 and 3 respectively, while those having maximum scores close to 1 are in post-symptom region 4 (Figure 2.6A). Strikingly, the subjects corresponding to those chips with high factor scores are exactly those who were later confirmed to have developed clinical overt symptoms. Of note, the boundary between regions 3 and 4 (Figure 2.6A) of the Sx subjects is the abrupt symptom onset time detected by this principal

BLU factor. Furthermore, a list of most dominant genes that are enriched in this factor was determined by using a simple dominance ranking on the factor loading of each gene over the three factors. Specifically, a gene is said to enrich the principal factor if its loading in this factor is the highest among the three factors.

Based on results of BLU, we defined 4 classes: Class 1 (pre-inoculation) corresponds to chips acquired from subjects before the inoculation time; Class 2 (post-inoculation Asx) corresponds to chips from asymptomatic subjects acquired after-inoculation; Class 3 (post-inoculation and pre-symptom) corresponds to chips from symptomatic subjects acquired after inoculation but before symptoms occur; and Class 4 (post-symptom) corresponds to chips from symptomatic subjects after symptoms occur (Figure 2.6A). Among these, discrimination between Class 3 and Class 2 is of particular interest since they essentially separate those who will become symptomatic from those who will not.

Early and late phase disease stratification using a logistic boosting model Using the 4-class designation defined by BLU, we employed a state-of-the-art machine learning method, LogitBoost (also called BinomialBoosting) (Bhlmann, 2006, Bhlmann and Yu, 2003, Bhlmann and Hothorn, 2007), to construct the risk stratification model. The objective of using this model was two-fold: 1) to determine the principal genes that contribute to the highest discrimination capacity over different pairs of these classes; and 2) to quantify the difficulty or uncertainty in discriminating between these classes in terms of receiver-operating-characteristic (ROC) curves. An ensemble of weak classifiers, called base learners, was constructed for each pair of classes. Each base learner constructs decision boundaries for classification on the basis of the gene chips falling within the region defined by that class pair. To further simplify the model and make it less prone to overfitting, we specialized to a simple univariate logistic regression decision rule for each

base learner. Model fitting was carried out with a functional gradient descent algorithm (Bhlmann, 2006, Bhlmann and Yu, 2003, Bhlmann and Hothorn, 2007).

During training, only one predictor variable enters the model at an iteration of the fitting procedure. Such component-wise univariate addition of predictor provided an implicit variable selection mechanism that supplements the classifier with the next best predictor gene at each step. The final estimate (after M boosting iterations) takes a form of a linear combination of base procedures ($\hat{f}_m$) fitted at an iteration m

$$\hat{f}(x) = \nu \sum_{m=1}^{M} \hat{f}_m(x) \tag{2.6}$$

with ν being the *shrinkage factor* ($0 < \nu \leq 1$). This *shrinkage factor* determines the amount of contribution to the final model by each fitted base leaner. A smaller value of ν results in a slower rate of increase in the overall mis-classification risk, thus making it less prone to overfitting. A disadvantage of choosing a smaller ν is that more boosting iterations are needed to reduce the same amount of classification risk as compared to faster learning rate. This results in more computation time.

In order to balance the tradeoff between lower mis-classification risk and speedy computation time, we introduced a two stage model fitting procedure. In the first-pass fitting step, a larger weight ($\nu = 1$) was used to identify those variables (genes) that never entered the model. They were thus removed from the training data as they are not required for this classification task, thus effectively reducing the search space for model fitting. We note that the removal of these genes do not necessarily imply that these genes have no discrimination capacity. On the contrary, some of these removed genes may well be discriminatory as they are all significantly differentially expressed between asymptomatic versus symptomatic. The reason that they were not selected by the model was simply because that other gene discriminants having better predicting power under the given model

selection criteria. In many cases, the selected gene surrogates were often highly correlated with the ones left out of the model (e.g., the genes belong to the same SOM cluster). In the second step, we used a smaller weight ($\nu = 0.01$) to construct the final classifier on this pre-filtered dataset. Furthermore, the Akaike Information Criterion (*AIC*) was employed to determine the optimal number of boosting iterations

$$\boldsymbol{AIC}(M) = -2\, max\, \text{loglikelihood} + 2\, df_M \tag{2.7}$$

To assess the performance of the boosted classification model and the selected predictor genes, we used a bootstrap resampling technique (Efron, 1979) to *post hoc* cross-validate the boosting classifiers. The training set ($2/3$ of data) was used to construct the boosting ensemble while the test set ($1/3$ of data) was used for testing. This strategy is similar to the *out-of-bag* unbiased error estimation used in Random Forests (Breiman and Friedman, 2001). We report the performance with ROC curves using only the test set and chose the area under the curve (AUC) as a measure of classification performance. We also computed 95% confidence intervals on the ROC, based on the bootstrap resampled data, for both true positive prediction and false negative prediction at each threshold point on the ROC. This essentially quantifies the intrinsic difficulty or uncertainty in discriminating each class pair. A total of $5,000$ bootstrap copies of data were generated to perform this analysis. Of note, by using tree-based ensemble classification technique Random Forests, we were able to obtain similar classification performance (data not shown).

We want to point out that no additional constraint on subject selection was enforced during model fitting. In other words, expression profiles from all individual volunteers were subject to bootstrap selection in an equal and independent manner. An alternative analysis strategy would be to specifically hold out samples from a few subjects during training and test on the held-out samples. However, given that the total number of subjects

in one phenotype can be as few as 8, such constraint would likely result in underestimating performance.

2.7 SUPPLEMENTARY DISCUSSION

Comparison of this study with the study reported by Zaas et al. (Zaas et al., 2009). This work probes the temporal nature of the host genomic response as compared to Zaas et al that looked at a single time point (peak infection). The question we are addressing here is whether and how the data evolve over time and whether the asymptomatic state represents a passive or active response to pathogens. Our data not only show that the peak Sx response from Zaas appears to be manifest as an evolving signature of two gene clusters (2 and 3 which are predominantly inflammatory response), but additionally that mechanisms characterized by clusters 1 and 4–8 are temporally activated as well. Our data further shows that there is an active temporal response in Asx that differs from the Sx response and is particularly strong in clusters 2, 6, and 8.

Zaas et al found 30 biomarkers that best discriminated between symptomatic and asymptomatic individuals at peak symptom time. The vast majority (29 out of 30) of these biomarkers are found in clusters 3 (A^{nc},S^{up}_{mid}) and cluster 2 (A^{dw}_{early},S^{up}_{mid}) of our paper (Table S2.5). Not surprisingly, these two clusters are those that show the largest contrast between Asx and Sx expression levels near peak symptom time (Figure 2.1). Other clusters reported in our paper correspond to genes that respond differently in Asx and Sx at earlier time points. Thus, while being completely consistent with the results from Zaas et al (Zaas et al., 2009), our analysis goes beyond peak symptom time and establish striking temporal differences in host response programs between Asx and Sx subjects.

In Zaas et al, the performance of the peak time Sx vs Asx classifier was validated on an independent dataset presented in Ramilo et al (Ramilo et al., 2007). As 29 of the

30 genes identified in Zaas et al are in our clusters this also validates our results at peak symptom time. Unfortunately, as the Ramilo dataset only consists of clinical samples taken near peak symptom time it cannot be used to validate our temporal analysis at other time points.

The nature of asymptomatic phenotype We performed several tests to rule out the possibility that our results may merely reflect failed inoculation in the asymptomatic subjects instead of innate differences in host response. Although it is difficult to rule this out with 100% certainty, our data suggests that it is highly unlikely that the inoculation failed to establish productive infection in Asx hosts. The evidence against failed inoculation is a follows.

1. The temporal gene expression analysis presented in the main text has shown that Asx transcription state is not passive. Instead, it actively evolves in response to viral challenge. As presented in the manuscript, the viral inoculation elicited a strong molecular host response in the Asx subjects. When the expression profiles from asymptomatic subjects were studied alone, a total of more than $3,000$ genes showed statistically significant post-infection expression changes. In particular, such expression change does not correlate with viral detection. For example, two subjects **#3** and **#17** never yield detectable virus (< 1.25) in their nasal wash (Table S2.2). However, the Asx-specific temporal suppression of gene NLRP3, a key factor involved in activating *inflammasome* protein complex, is among the most significant for these two subjects (Figure S2.14).

Moreover, the gene expression responses of the two seroconverted Asx subjects (#2 and #3), according to haemagglutination inhibition (HAI) assay, are not significantly different from those of other asymptomatic individuals (Figure S2.14). As additional evidence, Figure S2.15 shows individual Asx subjects' temporal expression of RPL3 (refer-

enced in Figure S2.5 of the paper). The overall average of the Asx profiles is temporally changing at FDR level of significance (q-value) 0.0002 and again the subjects #2 and #3 do not appear to have atypical trajectories. Considering the fact that these subjects tested negative for binding antibody (Ab) to HA prior to inoculation, this indicates that the non-passive transcriptional responses we observed in Asx hosts are not directly related to their serum binding Ab activity. We also found no significant correlation between serological conversion and the final disease outcome (p-value $= 0.27$), suggesting that host gene expression signature serves as a better marker for symptomatic infection than serology measures do.

2. In the paper we presented a set of predictor variables (Table S2.4 first column - labeled 1v2) that differentiate between pre-inoculation baseline samples and asymptomatic post-inoculation samples. Their high level of discrimination performance (ROC curve in Figure 2.6C; 1vs2) suggests that a robust immune response program was indeed activated in the Asx subjects. A few of these predictor variables, e.g. GM2A, IRS2, and FOXO3, have been previously implicated in innate immunity and insulin receptor signaling. We think that these variables represent potential new targets for studying viral control mechanisms in Asx host response.

3. The viral shedding rates observed in our study are not inconsistent with that of previous studies. Specifically, 50% (4 out of 8) of the Asx subjects had evident viral shedding and this is on par with that of "subclinical" or "secondary" infections reported by Lau et al (25). The level of shedding is has been referred to in the literature as "asymptomatic infection". Also, 75% (6 out of 8) of the Asx subjects reported some symptoms during the study. This provides further support for our clinical determination of the Asx subjects as "asymptomatic" (Lau et al., 2010).

4. We can also rule out any possible dosage effect as the inoculation dosage was found

to be un-related to the infection outcome (Figure S2.13). Subjects who received relatively lower amount of inoculation do not necessarily become more ill than the ones who received higher dose of virus and there is no statistically significant dependence between disease outcome and inoculation dosage. The test for dosage effect failed according to two standard statistical test of significance: Fisher's exact test (resulting in rejection of the correlation hypothesis at any level less than the p-value of 0.2299); and very low R value ($R^2 = -0.0662$) of a linear regression of the disease outcome on dosage level.

5. Assay of the serum neutralizing antibody (nAb) titre was performed on all samples at early time points and many subjects showed relatively high level of nAb (≥ 40) at the time of challenge. However, there is no significant difference between the nAb titers found in Asx or Sx subjects (Figure S2.12A,B). Because these subjects were recruited from a natural population and thus had likely prior exposure to viral pathogens such level of nAb activity is not surprising.

More concretely, a Wilcoxin rank test generated a p-value of 0.80 at day 0 and 0.82 at day 7 on the hypothesis that there is no difference between nAb levels in the Asx and Sx groups. Furthermore, the pre-inoculation nAb does not have significant effect (R squared of linear regression line in black equal to 0.06) on disease symptom severity as measured by the clinical Jackson scores (Figure S2.12C). Most importantly, the nAb titer is observed to increase over time in both Asx and Sx individuals (Figure S2.12D). At minimum, this indicates a boosting effect of immunity, and suggests that even if viral replication was inhibited, enough virus was detected by the Asx host immune system to cause expansion of Ab producing cells.

6. The reported attack rate in our study is consistent with other similar studies reported in the literature, e.g., Turner et al (Turner et al., 2005) and Carrat et al (Carrat et al., 2008). On the basis of an extensive survey of 56 human influenza challenge studies with $1,280$

health volunteers, Carrat et al reported that the frequency of symptomatic infection was 66.9% (95% confidence interval: 58.3, 74.5).

Taken together, we think that this provides strong evidence that the inoculation did elicit a unique and robust host molecular response in Asx subjects. This response is significantly different from that of Sx individuals. The fact that some Asx subjects were not infected does not render this group of subjects any less interesting than their symptomatic counterparts. We believe that there exist important biological and immunological reasons that some volunteers can withstand considerable amount of viral insult and show no severe disease symptoms and we hope that the findings will lead to additional studies that clarify the apparent immunity of asymptomatic responders.

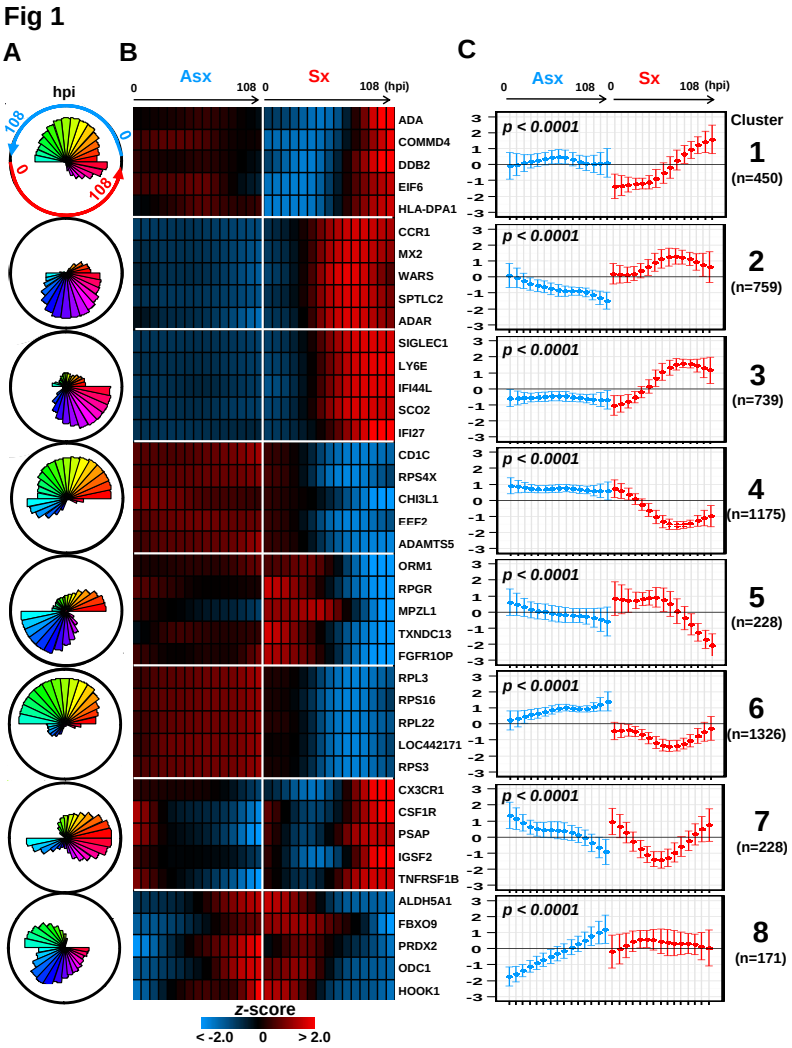
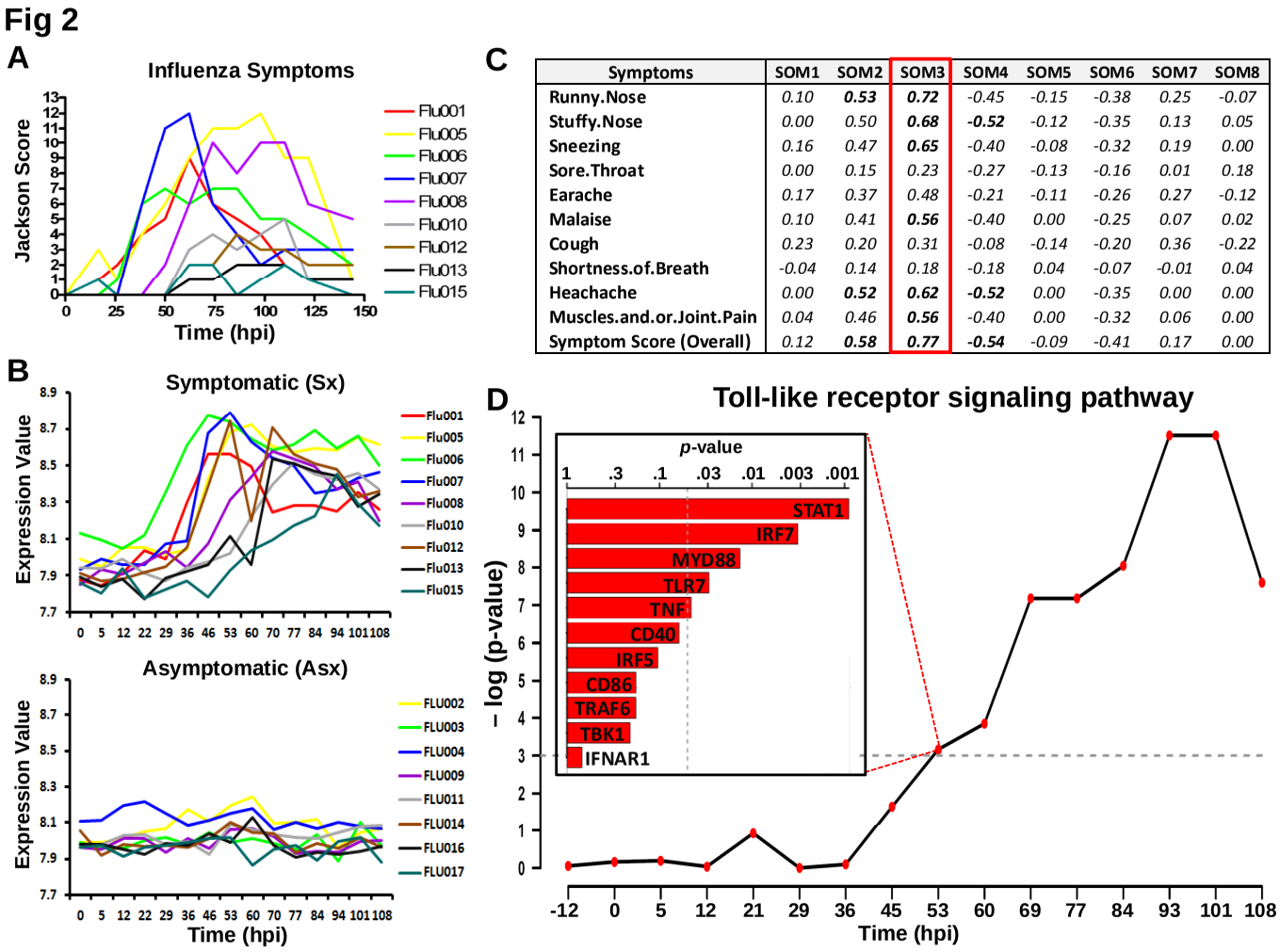


Figure 2.1: Distinct transcriptional dynamics between Asx and Sx subjects. (A) Polar plots of the 8 SOM clusters and their associated gene expression patterns. Each segment plot represents the prototype of a cluster. Individual time points are scaled and ordered in sequence and phenotype around the circle. Specifically, the temporal expression of Asx resides on the top portion of the circle while Sx expression occupies the bottom half. Each phenotype's expression values are placed in time sequence, with time increasing in the counterclock-wise direction, inside its own half circle. The degrees of angle are equally divided among segments within the circular plot. The different lengths of radii of the segments represent the deviation of a time point from the average expression level of the complete time course. (B) Heatmap of top 5 genes from each cluster. Genes are ordered within cluster according to their significance level. (C) Centroids of each SOM cluster show individual cluster average expression profile and corresponding ± 2 standard deviations. The total number of genes and significance (p-values) of differential expression between phenotypes are shown at the top left corner. The statistical significance of phenotype-specific trend of expression monotonicity can be found in (Table S2.1). hpi: hours post inoculation.



Symptoms	SOM1	SOM2	SOM3	SOM4	SOM5	SOM6	SOM7	SOM8
Runny.Nose	*0.10*	***0.53***	***0.72***	*-0.45*	*-0.15*	*-0.38*	*0.25*	*-0.07*
Stuffy.Nose	*0.00*	*0.50*	***0.68***	***-0.52***	*-0.12*	*-0.35*	*0.13*	*0.05*
Sneezing	*0.16*	*0.47*	***0.65***	*-0.40*	*-0.08*	*-0.32*	*0.19*	*0.00*
Sore.Throat	*0.00*	*0.15*	*0.23*	*-0.27*	*-0.13*	*-0.16*	*0.01*	*0.18*
Earache	*0.17*	*0.37*	*0.48*	*-0.21*	*-0.11*	*-0.26*	*0.27*	*-0.12*
Malaise	*0.10*	*0.41*	***0.56***	*-0.40*	*0.00*	*-0.25*	*0.07*	*0.02*
Cough	*0.23*	*0.20*	*0.31*	*-0.08*	*-0.14*	*-0.20*	*0.36*	*-0.22*
Shortness.of.Breath	*-0.04*	*0.14*	*0.18*	*-0.18*	*0.04*	*-0.07*	*-0.01*	*0.04*
Heachache	*0.00*	***0.52***	***0.62***	***-0.52***	*0.00*	*-0.35*	*0.00*	*0.00*
Muscles.and.or.Joint.Pain	*0.04*	*0.46*	***0.56***	*-0.40*	*0.00*	*-0.32*	*0.06*	*0.00*
Symptom Score (Overall)	*0.12*	***0.58***	***0.77***	***-0.54***	*-0.09*	*-0.41*	*0.17*	*0.00*

Figure 2.2: Cluster 3 molecular signature is most significantly correlated with clinical symptom scores. (A) Clinical symptom scores of symptomatic subjects with individuals represented by curves in different colors. (B) Cluster 3 gene expression of symptomatic subjects (top) and asymptomatic subjects (bottom). (C) Table of correlation coefficients (total variance explained) between standardized symptom scores and SOM clusters. (D) Significance of association (p-value) between Toll-like receptor (TLR) pathway and overall symptom severity. Significant positive association between TLR-pathway genes and symptom severity is shown at 53hpi (top left).

Fig. 3

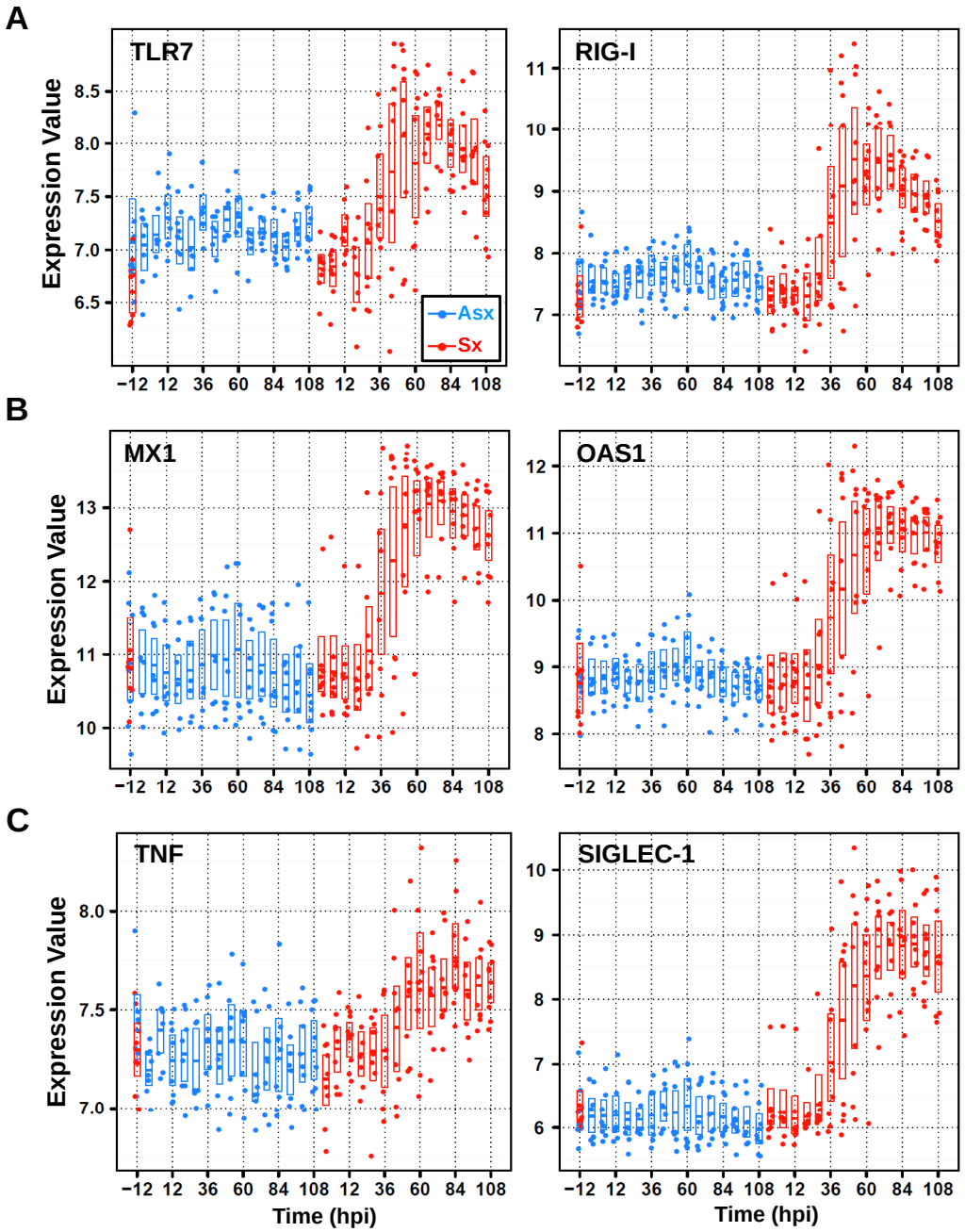


Figure 2.3: Similar expression dynamics of TLR7-pathway effector genes in cluster 3. Temporal expression patterns of representative significant genes on TLR-mediated signaling pathways that are related to the function of (A) pattern recognition and signaling regulation: TLR7 and retinoic acid inducible gene I (RIG-I) (B) antiviral: myxovirus resistant 1 (MX1) and 2',5'-oligoadenylate synthetase 1 (OAS1); (C) pro-inflammatory: tumor necrosis factor (TNF) and sialic acid binding Ig-like lectin 1 (SIGLEC-1). The expression intensities are plotted on a log base 2 scale and all genes are differentially expressed between Asx and Sx at significance level ≤ 0.0001.

Fig. 4

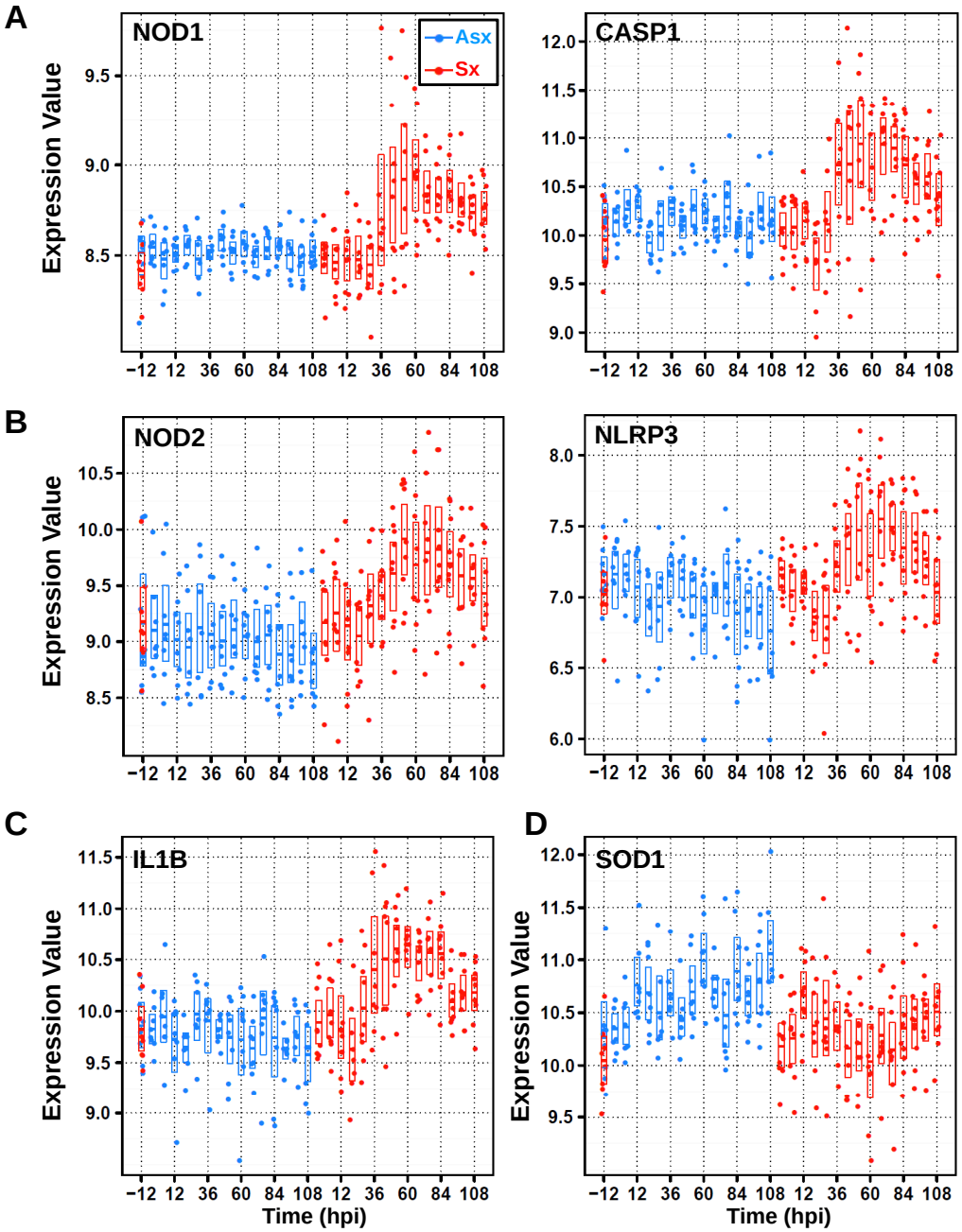


Figure 2.4: Divergent expression patterns of Nod/NACHT-LRR (NLRs) family of genes from cluster 2 and cluster 3 with contrasting expression of anti-oxidant/stress genes SOD1 and STK25 (or SOK1). (A) SOM cluster 3 genes nucleotide]binding oligomerization domain containing 1 (NOD1) and caspase 1 (CASP1) displays strong temporal upregulation in symptomatic subjects. (B) SOM cluster 2 genes NOD2 and NLRP3 exhibits downregulation in asymptomatic hosts and upregulation in symptomatic subjects. (C) SOM cluster 2 gene interleukin 1 beta (IL1B) shows symptomatic-specific upregulation versus asymptomatic-specific downregulation over time. (D) SOM cluster 6 genes superoxide dismutase (SOD1) shows upregulation versus downregulation in asymptomatic and symptomatic hosts, respectively. The expression values are plotted on a log base 2 scale and all genes are differentially expressed between Asx and Sx at significance level ≤ 0.0001.

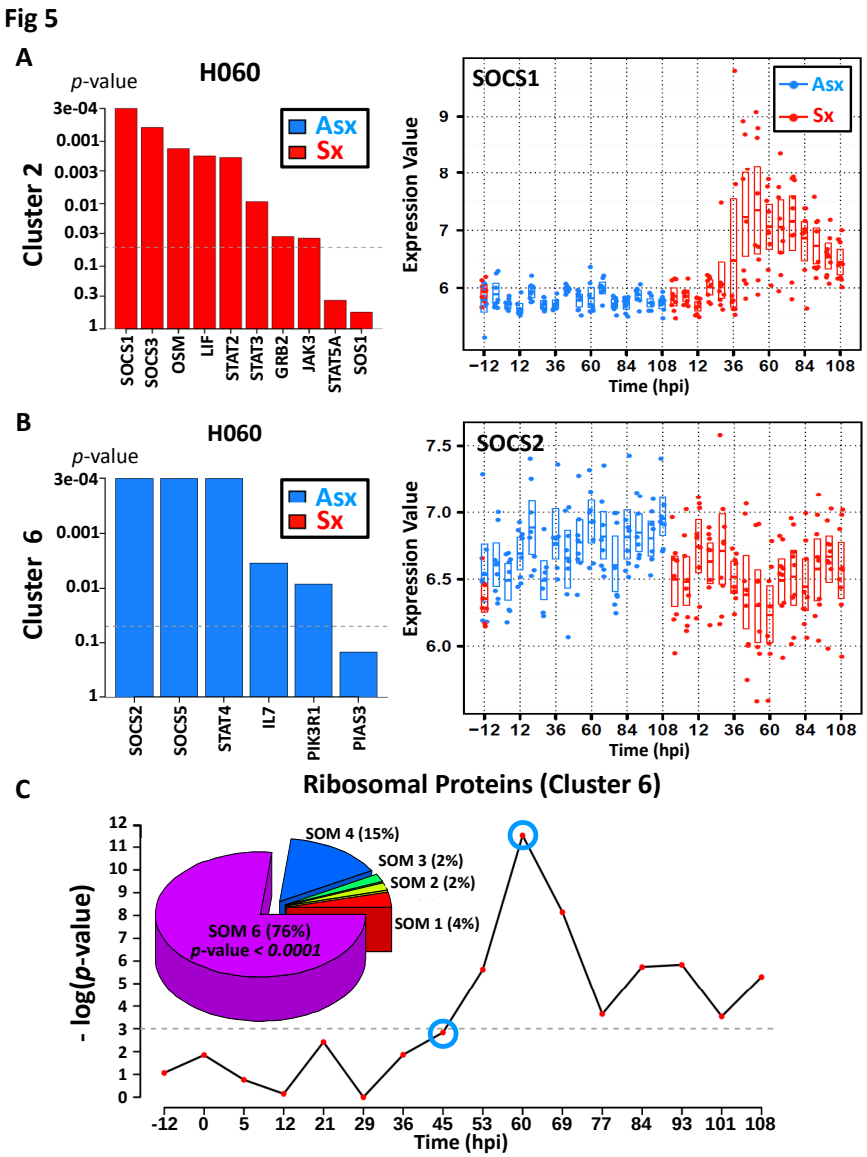


Figure 2.5: Asymptomatic hosts showed unique temporal expression kinetics of cluster 6 genes related to JAK-STAT signaling transduction and protein biosynthesis. (A,B) Distinct expression pattern of gene members in JAK-STAT pathway and their association with symptom severity. (A) Significant positive association between genes and disease severity is shown for 60hpi (left); temporal gene expression pattern of suppressor of cytokine signaling 1 (SOCS1) shows upregulation in symptomatic hosts. (B) Significant negative association between genes and disease severity is shown for 60hpi (left); temporal gene expression pattern of SOCS2 shows upregulation in asymptomatic hosts versus downregulation in symptomatic hosts. (C) Significance of negative association (p-value) between ribosomal protein synthesis (RPS)-related genes and overall disease severity; Pie chart (top left) shows a high degree of enrichment of significant RPS genes in SOM cluster 6, which is characterized by a trend of upregulation (in asymptomatic hosts) versus downregulation (in symptomatic hosts) over time.

Fig 6

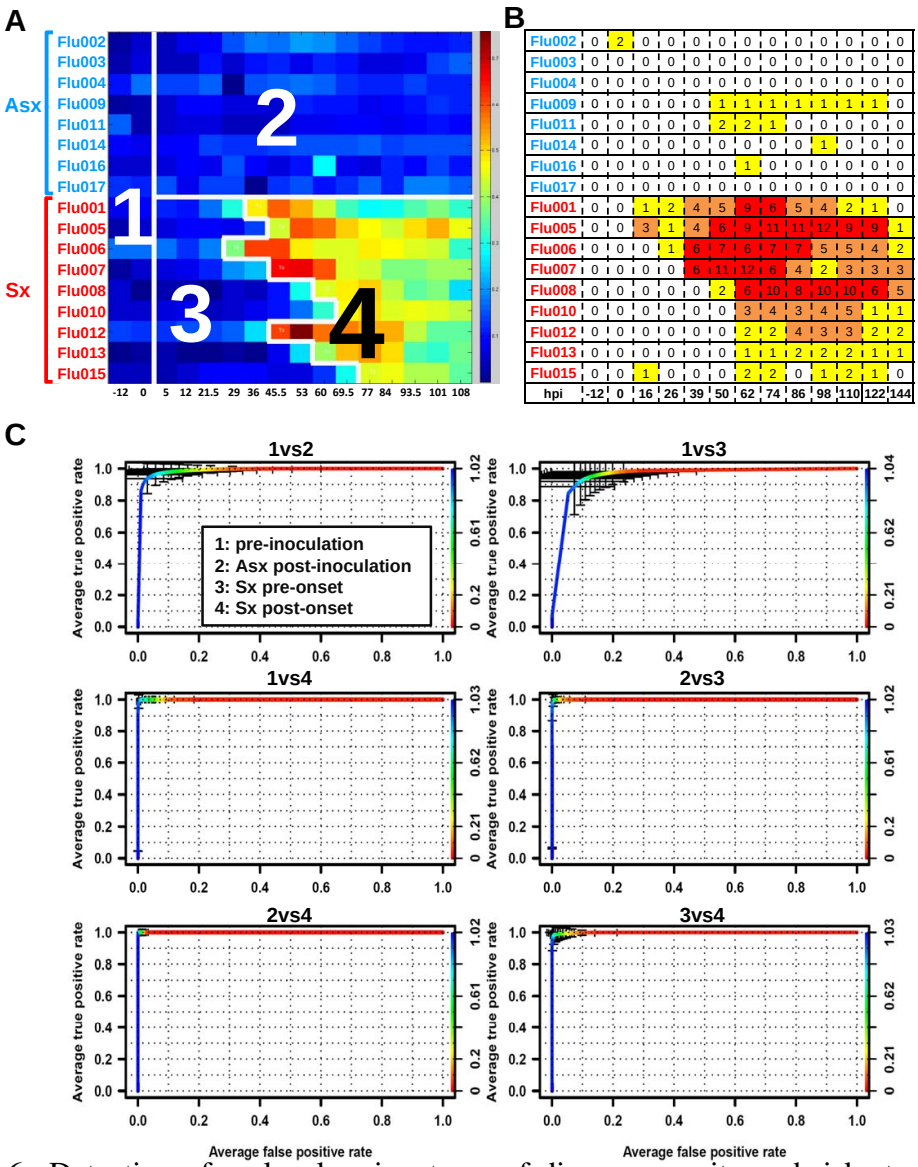


Figure 2.6: Detection of molecular signatures of disease severity and risk stratification models. (A) The scores of the top ranked factor detected by the unsupervised Bayesian linear unmixing (BLU) factor analysis method. Each microarray sample is represented by one square cell of the image and ordered by phenotype and subject (row-wise) and increasing time (column-wise). Color palette is coded according to the enrichment factor score determined by BLU. The higher the score, the warmer (red) color representation of the sample. The numbers (1 to 4) are the disease state (class) designation determined by BLU (3 and 4) and inoculation time (1 and 2). The boundary between 3 and 4 occurs at samples that are labeled (T_0) denoting the critical transition point (onset time) of Sx subject transcriptome profiles. The darkest blue color (absolute 0 loading) corresponds to samples that were not assayed. (B) Clinical symptom chart of corresponding subjects (rows) and times (columns) that are ordered in the same manner as A. (C) Performance of disease state model consisting of 52 distinct genes. Each individual plot shows the performance of a pairwise LogitBoost classifier. The ROC curve represents the average classification performance on the test set (hold-out set) in 2,000 bootstrapped copies of data. The error bar represents the 95% confidence interval (± 2 SEM) for each threshold point of the boosting classifier.

Table 1. Canonical pathways and representative genes enriched in individual SOM clusters.

SOM Cluster	# of Genes	Pathway	Representative Genes
1 (A^{nc}, S^{up}_{late})	450	immune cell trafficking; antigen presentation	CD74, HLA-DMA, HLA-DPA1, HLA-DPB1, CCR5, CCL4, TBX21, IL10RA, CD244, ICAM2
2 $(A^{dw}_{early}, S^{up}_{mid})$	759	inflmmation; chemotaxis of macrophage, neutrophils, and dendritic cells; antigen presentation, JAK-STAT signaling	SOCS1, SOCS3, NOD2, NLRP3, CASP5, IL1B, STAT3, ADM,C5, CCL2/7/8/11, CCR1, CCR4, CD14, CD59, CD163, CD209, CEACAM3, CXCL9, CXCL10, CXCL11, FAS, HLA-B, ICAM1, IL17RB, IL18R1, IL18RAP, LILRA2, LTBR, MX2, TGFB1, TLR1, TLR2, TLR4, TLR5, TLR8, TREM2, TRIM21, SERPINA1, CASP4, IFITM2
3 (A^{nc}, S^{up}_{mid})	739	inflammatory response; dendritic cell and neutrophil activation; IFN-signaling	TLR7, MYD88, IRF7, IRF5, IRF9, TNF, JAK2, PSMB8, STAT1, DDX58, IFIH1, IL18, IL10, MX1, RSAD2, OAS1, SIGLEC1, NOD1, CASP1, PKR, TRIM22, LILRB1, ISG20, IFNAR1, IFI44, CD86, CD40, CD63, C1QA, IL10RB, TNFRSF14, TNFSF10, TNFSF12, BTK, RNASE2; C3AR1, CYBB, FASLG, APOL3, ANXA2, IFI35, IFIT1, IFIT3, IFITM1, IFITM3
4 (A^{nc}, S^{dw}_{mid})	1175	oxidative stress; ca+ induced T cell apoptosis; iCOS signaling;	CCL5, RPS6KA5, ACTG1, CUL3, PRKC GENES, C-JUN, PIK3 family, MAP2K4, CD3E, CD247, CD40LG, CAMK4M, IL2RB, ITK, ITPR1, ITPR3, LAT, NFATC1, NFATC3, ICOS, FYN
5 $(A^{dw}_{mid}, S^{dw}_{late})$	228	antigen presentation; innate immune response	CD97, THBD, DDX17, IL1R2, ORM1, TREM1, AOC3, FOXO3, IL1R1, IL1RAP, AQP9, CA4, CAMK1D
6 $(A^{up}_{early}, S^{dw}_{mid})$	1326	protein synthesis; oxidative stress; RNA trafficking; JAK-STAT signaling	SOCS2, SOCS5, SOD1, SOK1, RPL3, EIF3 FAMILY GENES, CCR7, RPS9, RPS14, RPL22, C1QBP, DDX21, DDX50, ICOS
7 $(A^{dw}_{mid}, S^{dw}_{early})$	228	natural killer cell signaling; cell apoptosis	SIGLEC7, ASC, SHC1, MAPK7, KIR2DL1, KIR2DS4, KIR3DL1, SERPINF1, RAC1, CD4, CX3CR1, HLA-G, TNFRSF1B, ITGB2, CTSD
8 (A^{up}_{mid}, S^{nc})	171	cell morphology; cell signaling	EIF2AK1, LY96, BCL2L1, KRAS, PIM1, TGM2, RGS1, PKN2

CHARACTERISTICs	ALL SUBJECTs	ASYMPTOMATIC (Asx)	SYMPTOMATIC (Sx)
No. of subjects	17	8	9
Female sex (%)	8 (47%)	3 (38%)	5 (56%)
Age (min, max)	27 (22, 41)	28 (22, 41)	27 (22, 35)
White race (%)	14 (82%)	7 (88%)	7 (78%)
Average symptom score (min, max)	2 (0, 15)	(0, 1)	4 (0, 15)
Average peak symptom (min, max)	8 (0, 15)	0 (0, 1)	12 (8, 16)

Table S1. Subject Demographic and Clinical Characteristics of Viral Challenge Cohort.

A

	Virus Isolation (from nasal wash)										
Pheno	Unique ID	Day -2	Day -1	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Asx	flu002	none detected	none detected	none detected	<1.25	<1.25	<1.25	<1.25	1.75	<1.25	1.75
Asx	flu003	none detected	none detected	none detected	<1.25	<1.25	<1.25	<1.25	<1.25	<1.25	<1.25
Asx	flu004	none detected	none detected	none detected	<1.25	<1.25	<1.25	1.5	<1.25	<1.25	1.75
Asx	flu009	none detected	none detected	none detected	<1.25	<1.25	<1.25	<1.25	<1.25	<1.25	<1.25
Asx	flu 011	none detected	none detected	none detected	<1.25	<1.25	<1.25	<1.25	<1.25	<1.25	<1.25
Asx	flu014	none detected	none detected	none detected	<1.25	<1.25	<1.25	<1.25	<1.25	1.75	<1.25
Asx	flu016	none detected	none detected	none detected	<1.25	<1.25	<1.25	<1.25	<1.25	1.75	<1.25
Asx	flu017	none detected	none detected	none detected	<1.25	<1.25	<1.25	<1.25	<1.25	<1.25	<1.25
Sx	flu001	none detected	none detected	none detected	4.25	4.25	<1.25	<1.25	<1.25	<1.25	1.75
Sx	flu005	none detected	none detected	none detected	<1.25	4.5	3.5	<1.25	<1.25	<1.25	1.75
Sx	flu006	none detected	none detected	none detected	3.75	5	3.25	<1.25	<1.25	1.75	1.75
Sx	flu007	none detected	none detected	none detected	<1.25	6.25	2.75	<1.5	<1.25	<1.25	1.75
Sx	flu008	none detected	none detected	none detected	<1.25	4.75	1.75	<1.25	<1.25	<1.25	<1.25
Sx	flu010	none detected	none detected	none detected	<1.25	<1.25	3.75	<1.25	2.75	<1.25	<1.25
Sx	flu012	none detected	none detected	none detected	<1.25	5.01	5.01	<1.25	<1.25	2.75	1.75
Sx	flu013	none detected	none detected	none detected	<1.25	3.51	5.5	2.5	<1.25	<1.25	<1.25
Sx	flu015	none detected	none detected	none detected	<1.25	<1.25	<1.25	3.75	4.5	4	<1.25

B

	Serology				
Pheno	Unique ID	Pre-Screening visit	Day -1	Day 28 (convalescent)	Seroconversion
Asx	flu002	No detectable antibody	No detectable antibody	40	Yes
Asx	flu003	No detectable antibody	No detectable antibody	160	Yes
Asx	flu004	No detectable antibody	No detectable antibody	No detectable antibody	No
Asx	flu009	No detectable antibody	No detectable antibody	No detectable antibody	No
Asx	flu 011	No detectable antibody	No detectable antibody	Did not attend Day +28 visit	No
Asx	flu014	No detectable antibody	No detectable antibody	No detecable antibody	No
Asx	flu016	No detectable antibody	No detectable antibody	<20	No
Asx	flu017	No detectable antibody	No detectable antibody	No detectable antibody	No
Sx	flu001	No detectable antibody	No detectable antibody	320	Yes
Sx	flu005	No detectable antibody	No detectable antibody	320	Yes
Sx	flu006	No detectable antibody	No detectable antibody	57	Yes
Sx	flu007	No detectable antibody	No detectable antibody	<20	No
Sx	flu008	No detectable antibody	No detectable antibody	Did not attend Day +28 visit	No
Sx	flu010	No detectable antibody	No detectable antibody	Did not attend day +28 visit	No
Sx	flu012	No detectable antibody	No detectable antibody	80	Yes
Sx	flu013	No detectable antibody	No detectable antibody	40	Yes
Sx	flu015	No detectable antibody	No detectable antibody	20	No

Table S2: Viral shedding and serological testing data for all human volunteers (n=17) challenged with Influenza H3N2 viruses. a) Measure of viral titre isolated from nasal wash over a total of 9 days. **b**) Serological data on pre-screening, -24 hpi, and +28 days.

SOM Cluster	Asymptomatic (ASX)	Symptomatic (SX)
FLU1	*0.2964*	***< 0.0001***
FLU2	***< 0.0001***	***< 0.0001***
FLU3	*0.3924*	***< 0.0001***
FLU4	***< 0.0001***	***< 0.0001***
FLU5	***0.0002***	***< 0.0001***
FLU6	***0.0002***	***< 0.0001***
FLU7	***< 0.0001***	*0.1594*
FLU8	***< 0.0001***	*0.8264*

Table S3: Significance of monotonic trend of gene expression in SOM clusters. For the genes in each SOM cluster (Figure 1), we implemented the Jonkheere-Terpstra (JT) test (supl methods) of significance on Asx and Sx subjects, respectively, to test for monotonic increase or decrease of gene expression over time. Columns 2 and 3 show *p*-values associated with the null hypothesis that genes in the cluster have no monotonic trend. Red colored entries indicate clusters having highly significant monotonic expression profiles for a particular phenotype.

N	1vs2	1vs3	1vs4	2vs3	2vs4	3vs4
1	GM2A	IRS2	IRS2	RETN	RTP4	SMAD1
2	SLC35F2	FOXO3	SMAD1	C3AR1	GNG7	IFI44L
3	PITPNC1	APBA2	IFI44L	JUP	OSBPL10	RTP4
4	PIK3IP1	RPP30	GBP1	115648_at	OAS1	GRAMD1C
5	TPST1	NCKIPSD	BLVRA	RAB8A	CHI3L1	IRS2
6	PLAC8	ABCC3	IQGAP1	TPM2	CD1C	DDX17
7	RPP30	SLC20A1	GNG7	LAPTM4B	IFI27	
8	NID1	ADFP		LILRB2	MS4A1	
9	APOLD1	TBC1D4		CPNE1		
10	APBA2			HRASLS3		
11	TLK2			CLUAP1		
12	IRS2			GM2A		
13	PKN2			SLC12A9		
14	FOXO3			GNG7		
15	CENTA2			ENOSF1		
16				CDKN1C		
17				AP3S2		
18				DCHS1		
19				HBG2		

Table S4. Discriminatory genes selected by each logistic boosting model.
Genes are listed in decreasing order based on their discriminatory power in each model.

Influenza predictor genes (Zaas, 2009)	Cluster designation in this manuscript
RSAD2	3
IFI44L	3
SIGLEC1	3
LAMP3	3
IFIT1	3
IFI44	3
SERPING1	3
IFI27	3
ISG15	3
IFI44	3
HERC5	3
LOC26010	3
IFI6	3
LOC727996	N/A*
IFIT3	3
OAS3	3
OASL	3
4-Sep	2
XAF1	3
OAS1	3
LY6E	3
MS4A4A	3
SIGLEC1	3
TNFAIP6	3
CCL2	2
OAS1	3
MX1	3
TNFAIP6	3
RTP4	3
OASL	3

* This gene cannot be mapped due to public gene annotation issue.

Table S5. Comparison of genes identified by Aimee et al with significant genes in the present manuscript.

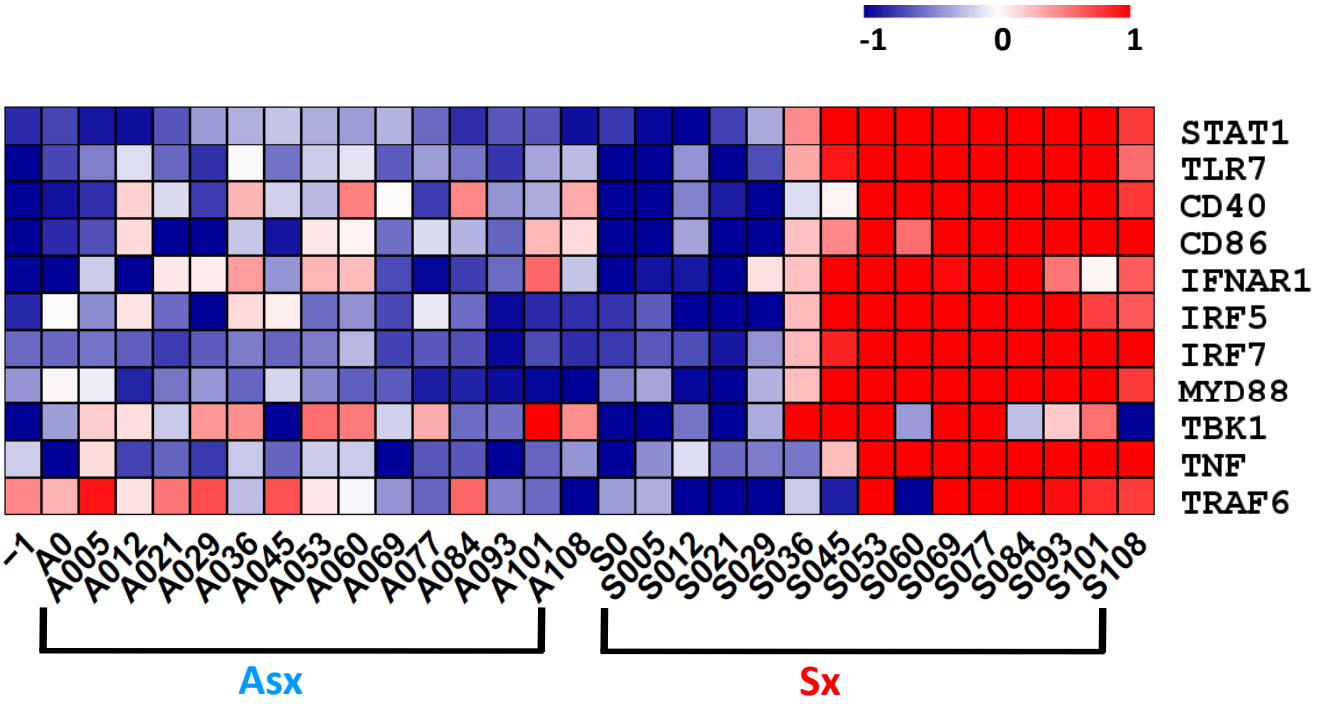


Figure S2.1: Temporal expression of Toll-like receptor 7 pathway member genes. Accompanying Figure 2.2c, temporal expression are shown for TLR7-pathways genes (n= 11) including STAT1, IRF7, MyD88, TLR7, TNF, CD40, IRF5, CD86, TRAF6, TBK1, and IFNAR1. The expression intensities are averaged over subjects in Asx and Sx phenotypes and plotted on a log base 2 scale.

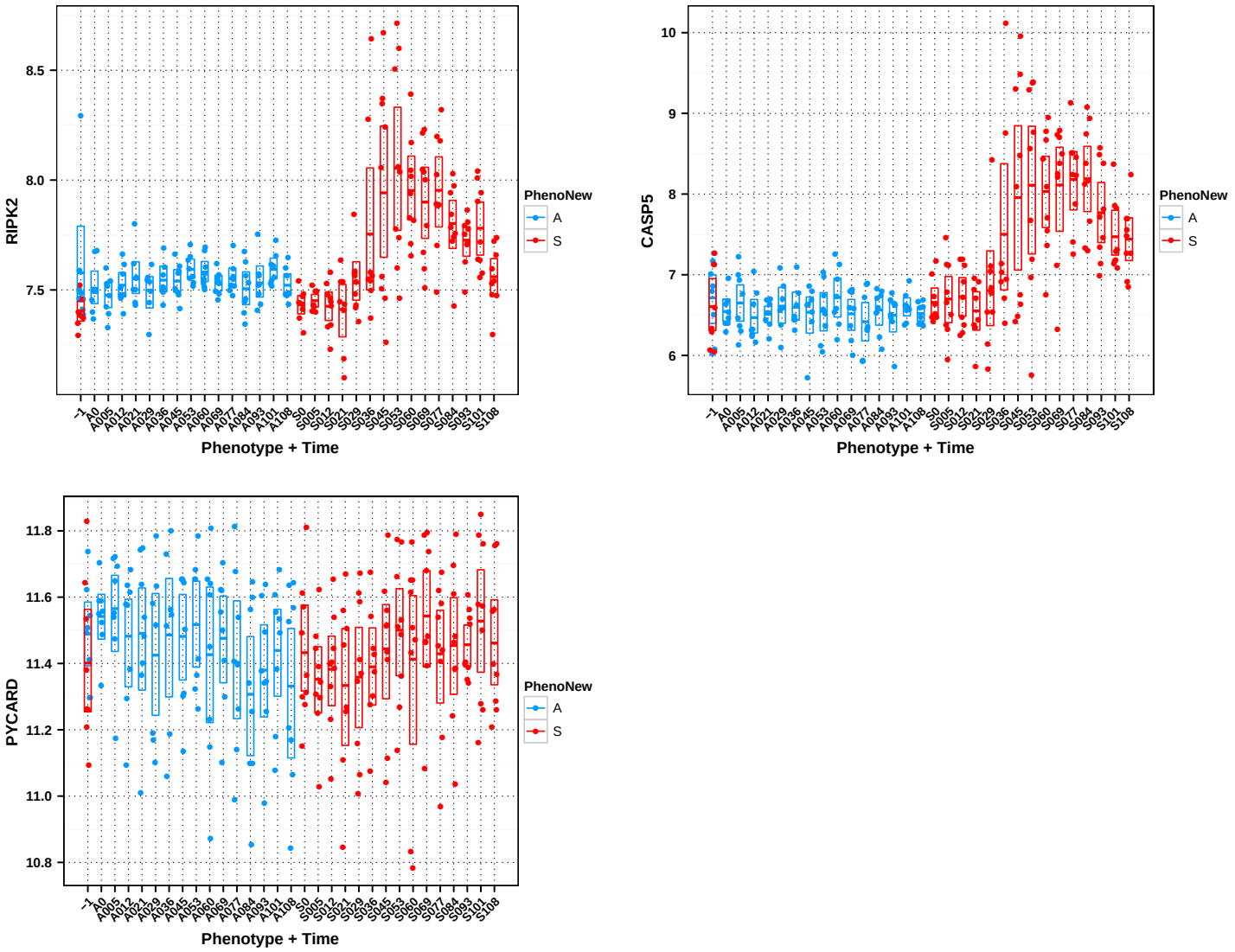


Figure S2.2: Temporal expression of NLR family genes. 1) cluster 7 gene PYD and CARD domain containing (PYCARD or ASC); 2) cluster 3 gene receptor-interacting serine-threonine kinase 2 (RIPK2); 3) cluster 2 gene caspase 5 (CASP5). The expression intensities are plotted on a log base 2 scale.

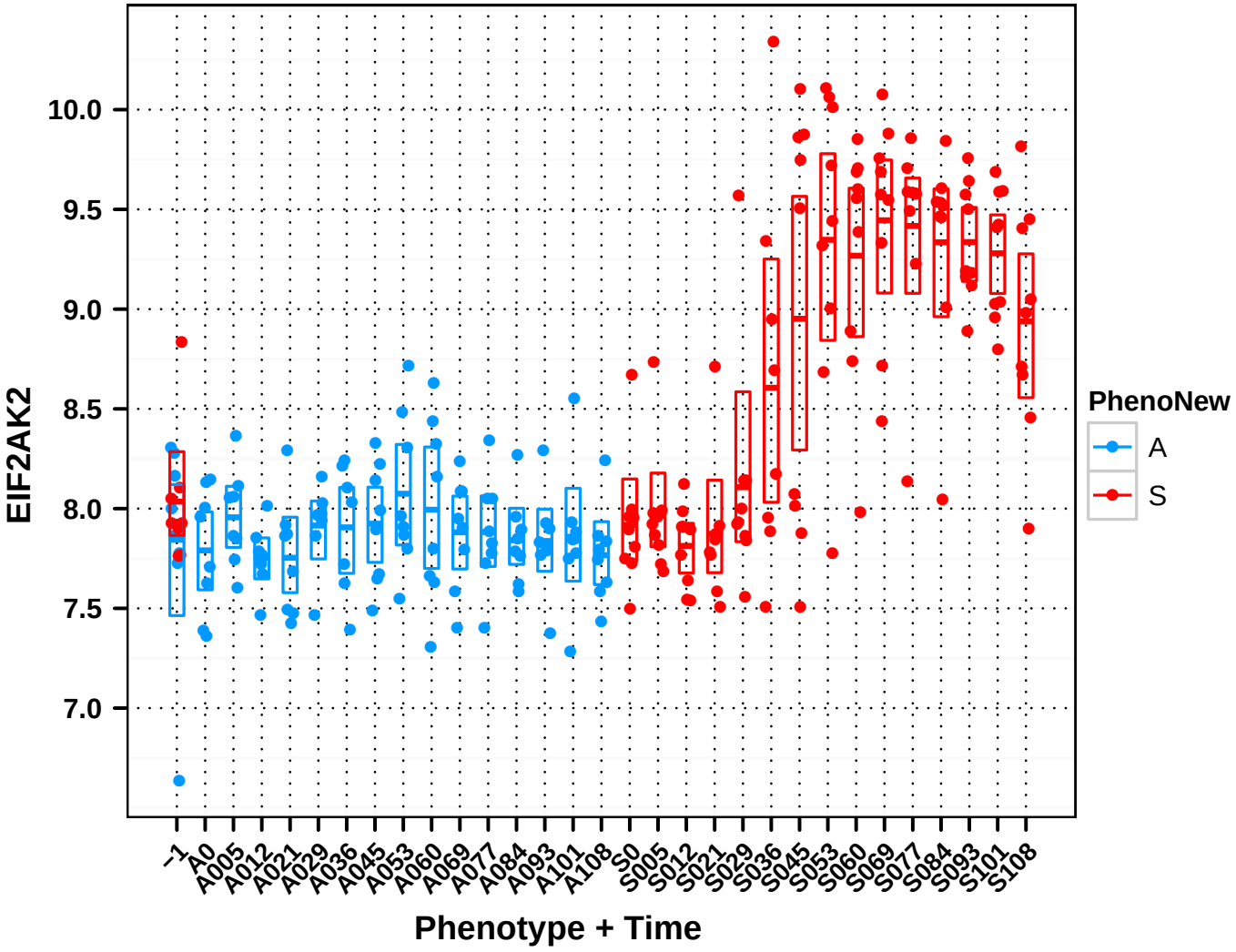


Figure S2.3: Increased temporal expression of antiviral RNA-dependent eIF-2 alpha protein kinase (EIF2AK2 or PKR) in cluster 3. The expression intensities are plotted on a log base 2 scale.

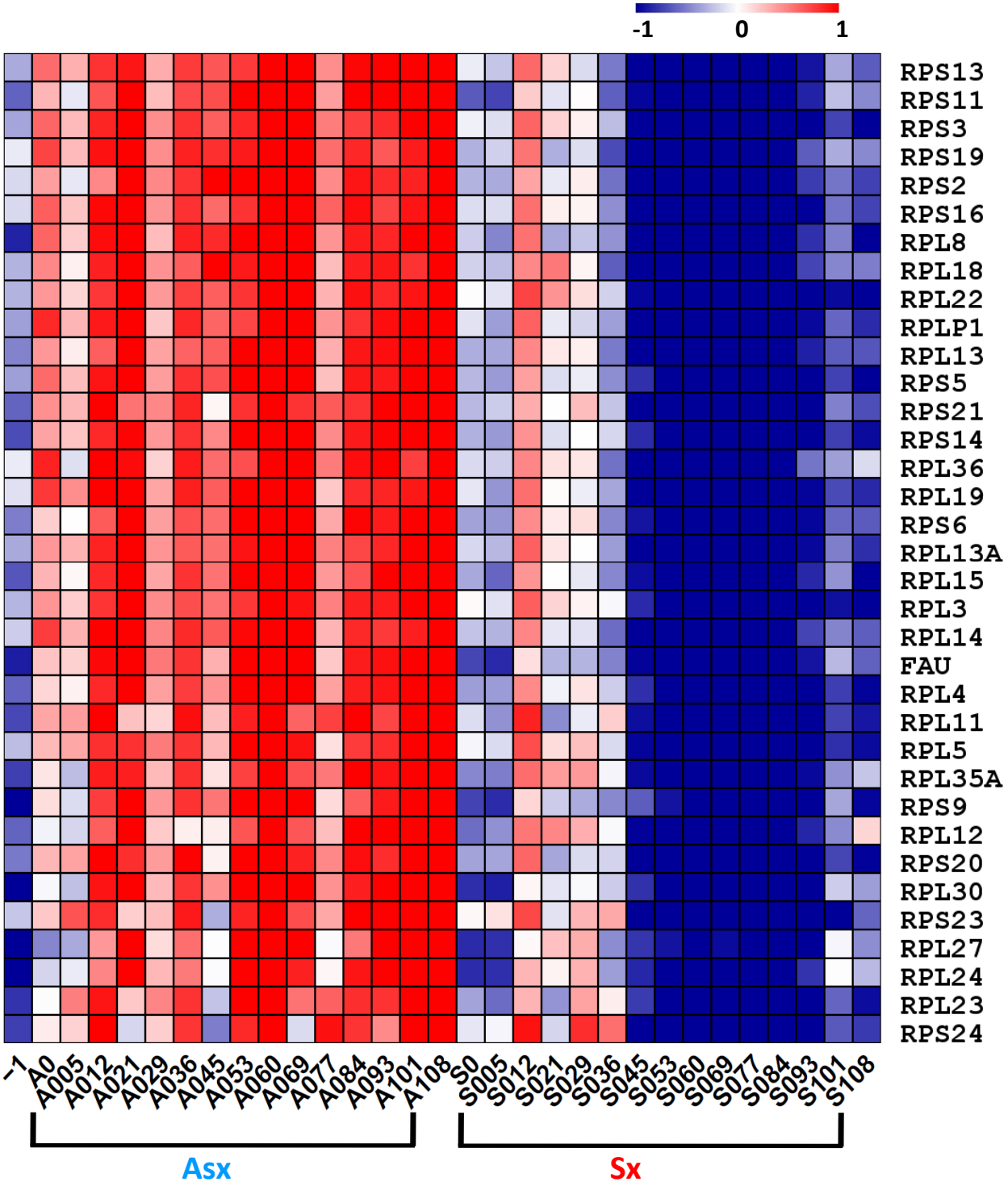


Figure S2.4: Phenotypically contrasting expression dynamics ribosomal protein synthesisrelated genes (n= 35) in cluster 6. The expression intensities are averaged over subjects in Asx and Sx phenotypes and normalized to have zero mean and unit standard deviation.

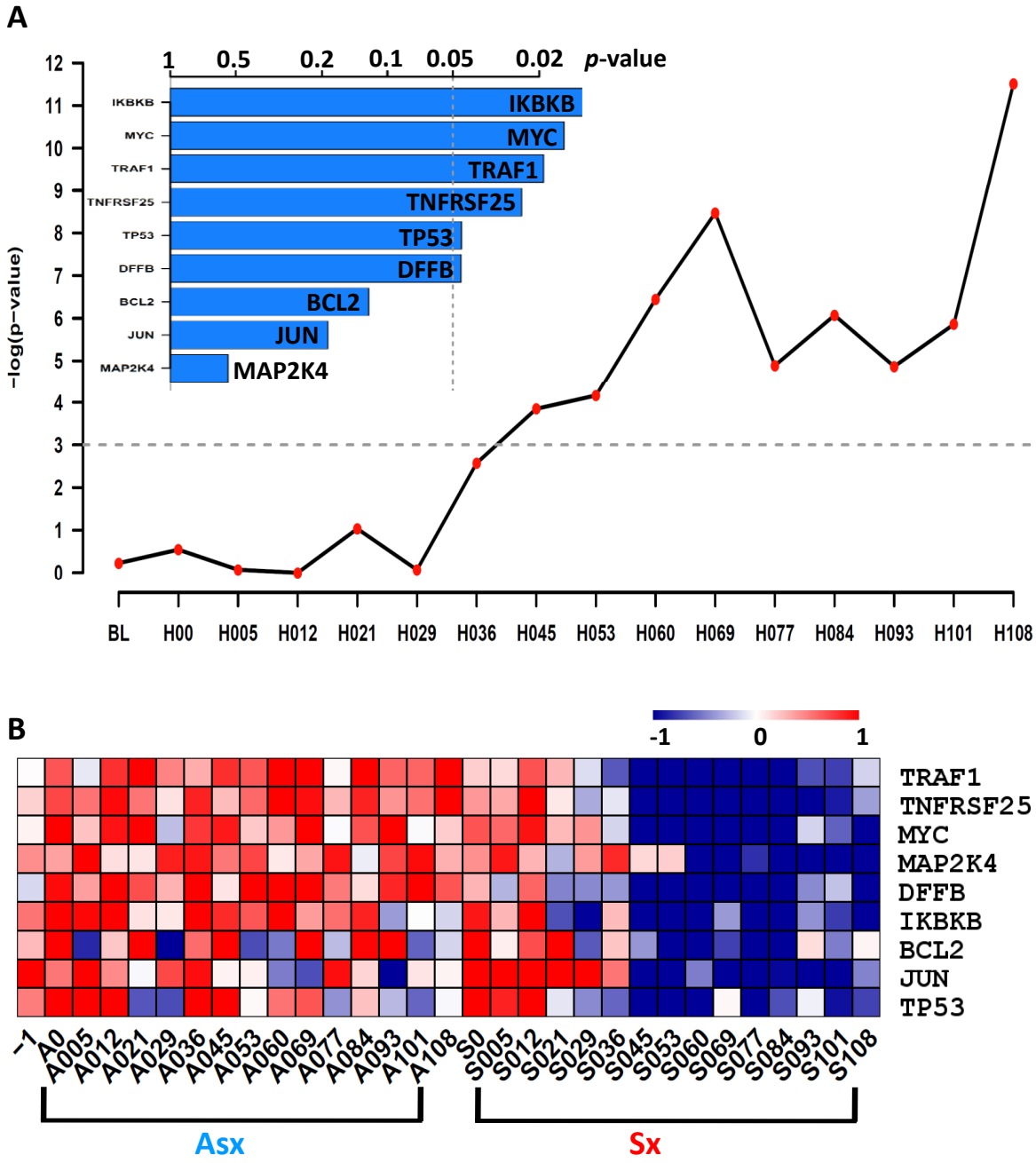


Figure S2.5: Symptomatic-specific temporal downregulation of cluster 4 genes (n=9) that regulate programmed cell death (apoptosis). A) Significance (*p*-value) of association between phenotypes and the whole group of genes at all time points and at time 45 hpi (top left panel). B) Average temporal expression intensities are computed on subjects in Asx and Sx phenotypes and normalized to have zero mean and unit standard deviation.

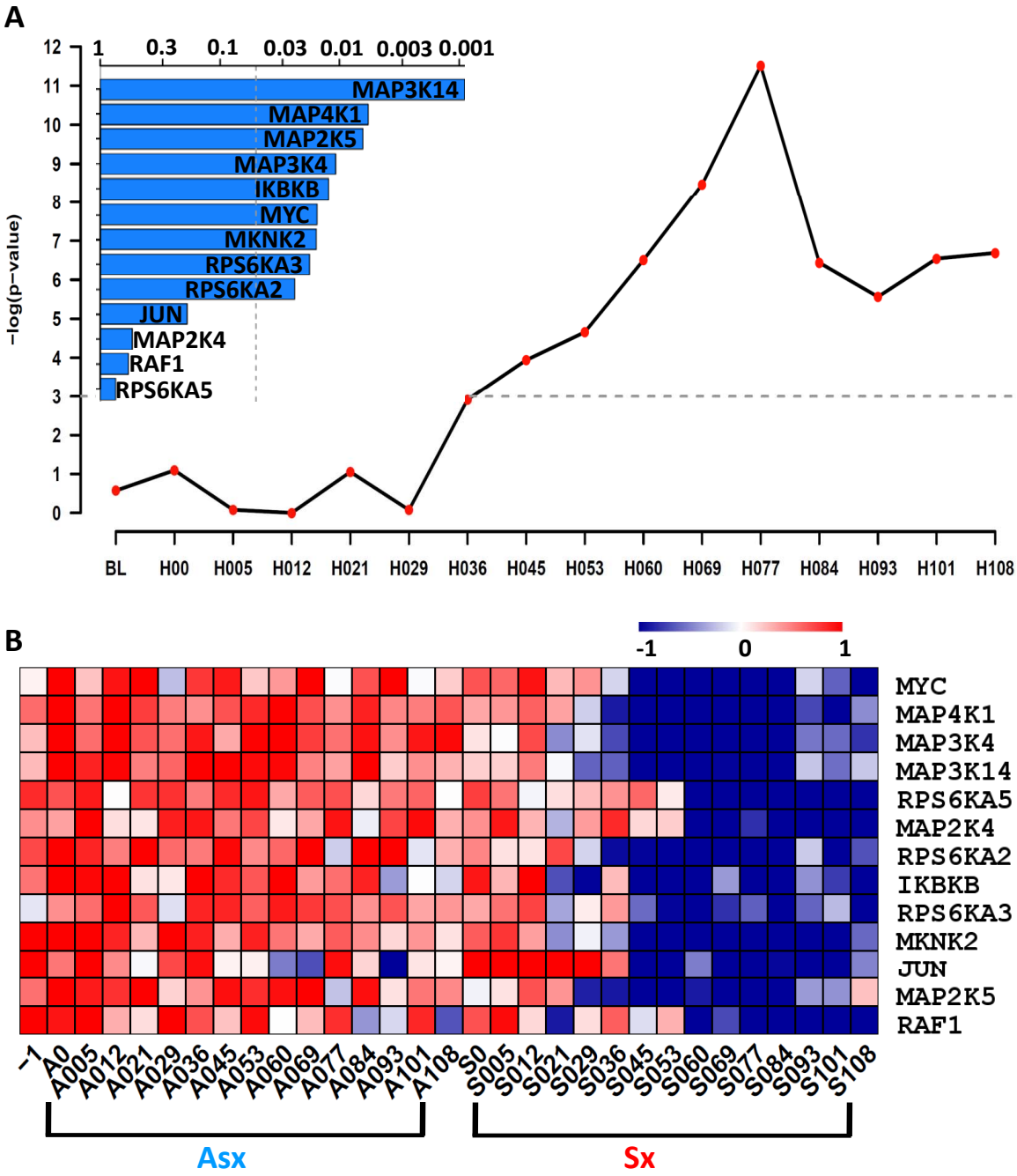


Figure S2.6: Symptomatic-specific temporal downregulation of cluster 4 genes (n= 13) that are related to mitogen-activated protein (MAP) kinase cascades. A) Significance (*p*-value) of association between phenotypes and the whole group of genes at all time points and at time 45 hpi (top left panel) . B) Average temporal expression intensities were computed on subjects in Asx and Sx and normalized to have zero mean and unit standard deviation.

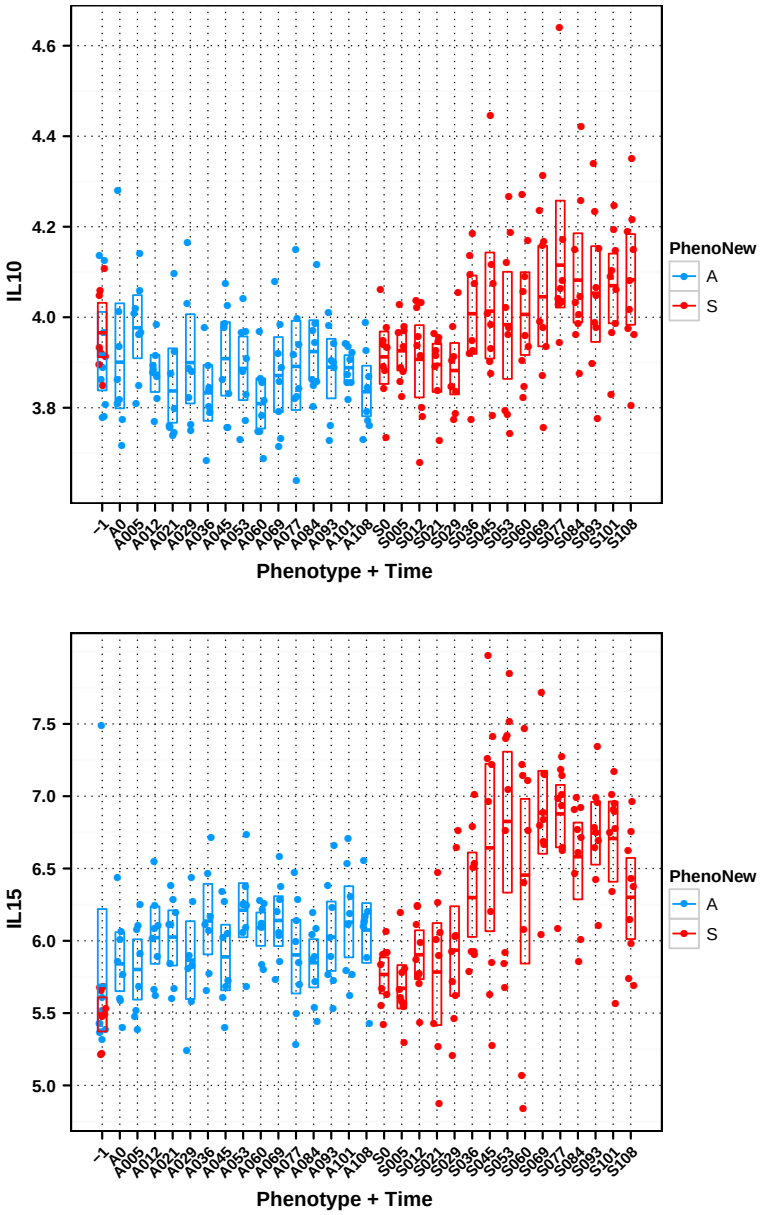


Figure S2.7: Increased temporal expression of inflammatory response regulators (cluster 3), interleukin 15 and interleukin 10. The expression intensities are plotted on a log base 2 scale.

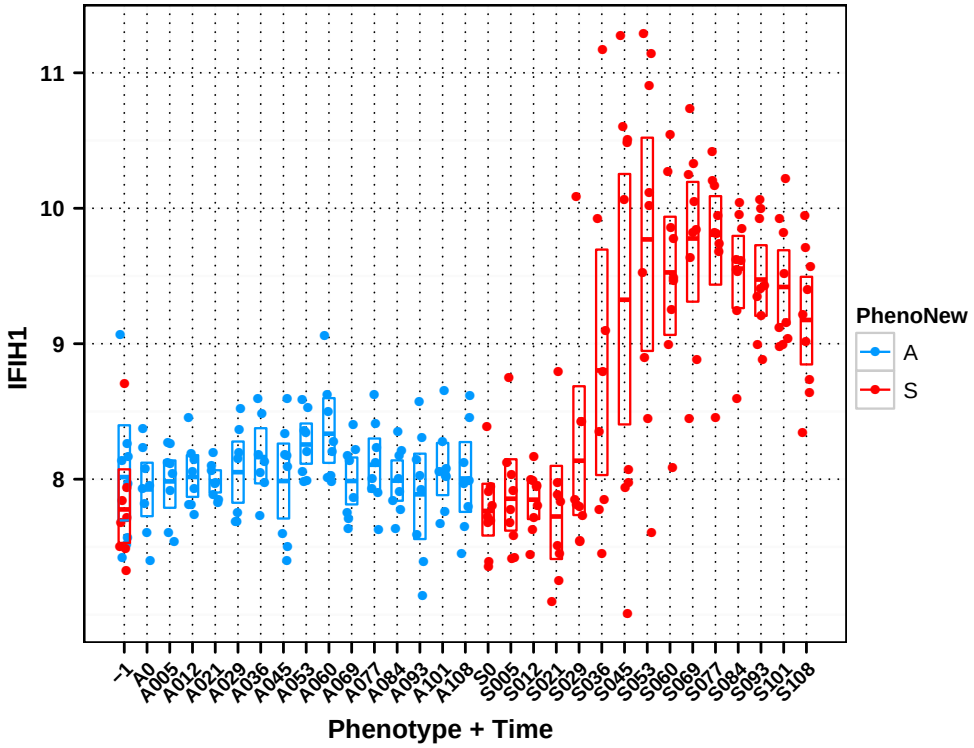


Figure S2.8: Temporal gene expression of cluster 3 gene cytoplasmic double-strand viral RNA sensor IFIH1 (interferon induced with helicase C domain 1). The expression intensities are plotted on a log base 2 scale.

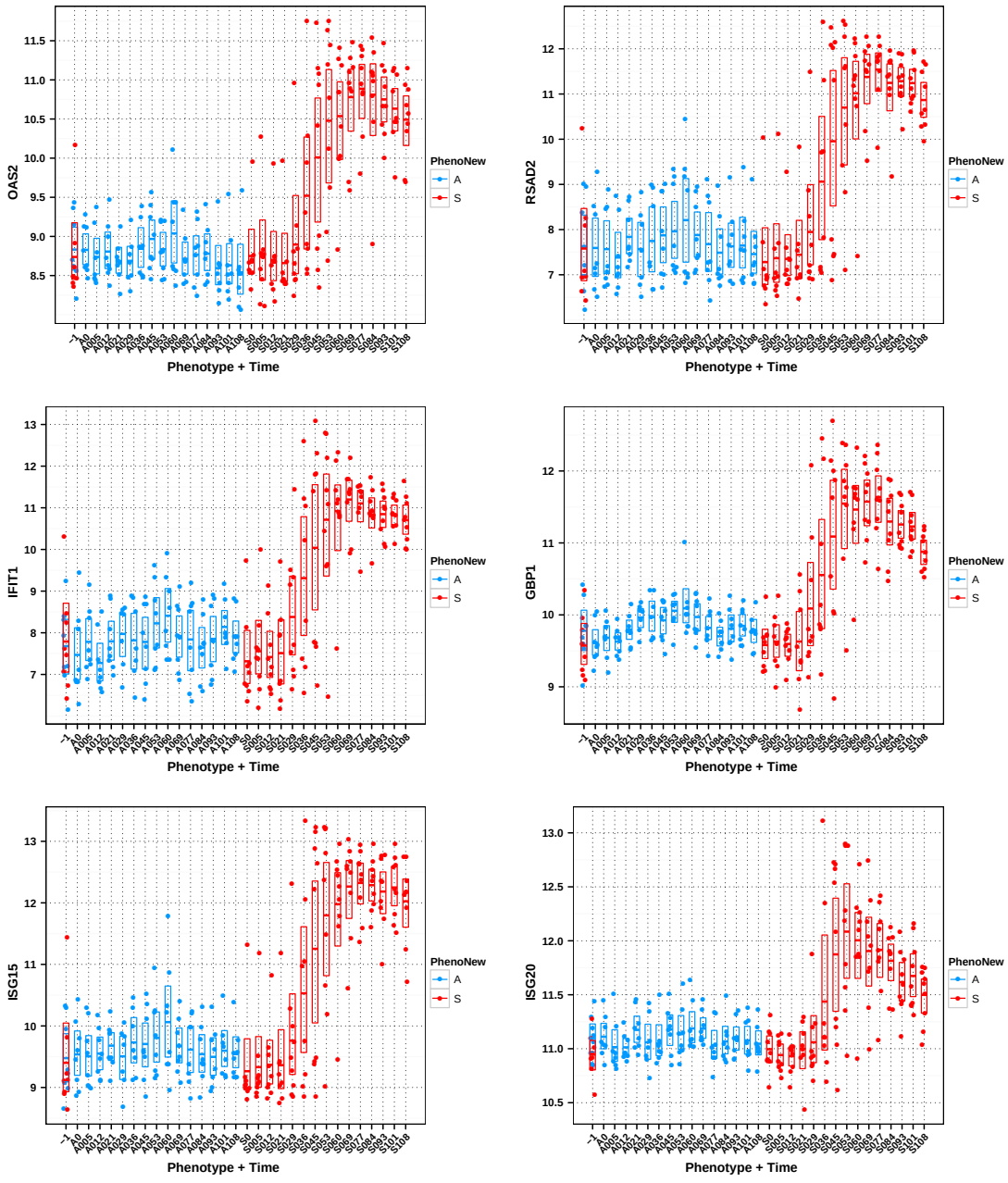


Figure S2.9: Temporal expression of interferon inducible anti-viral genes from cluster 3. The expression intensities are plotted on a log base 2 scale.

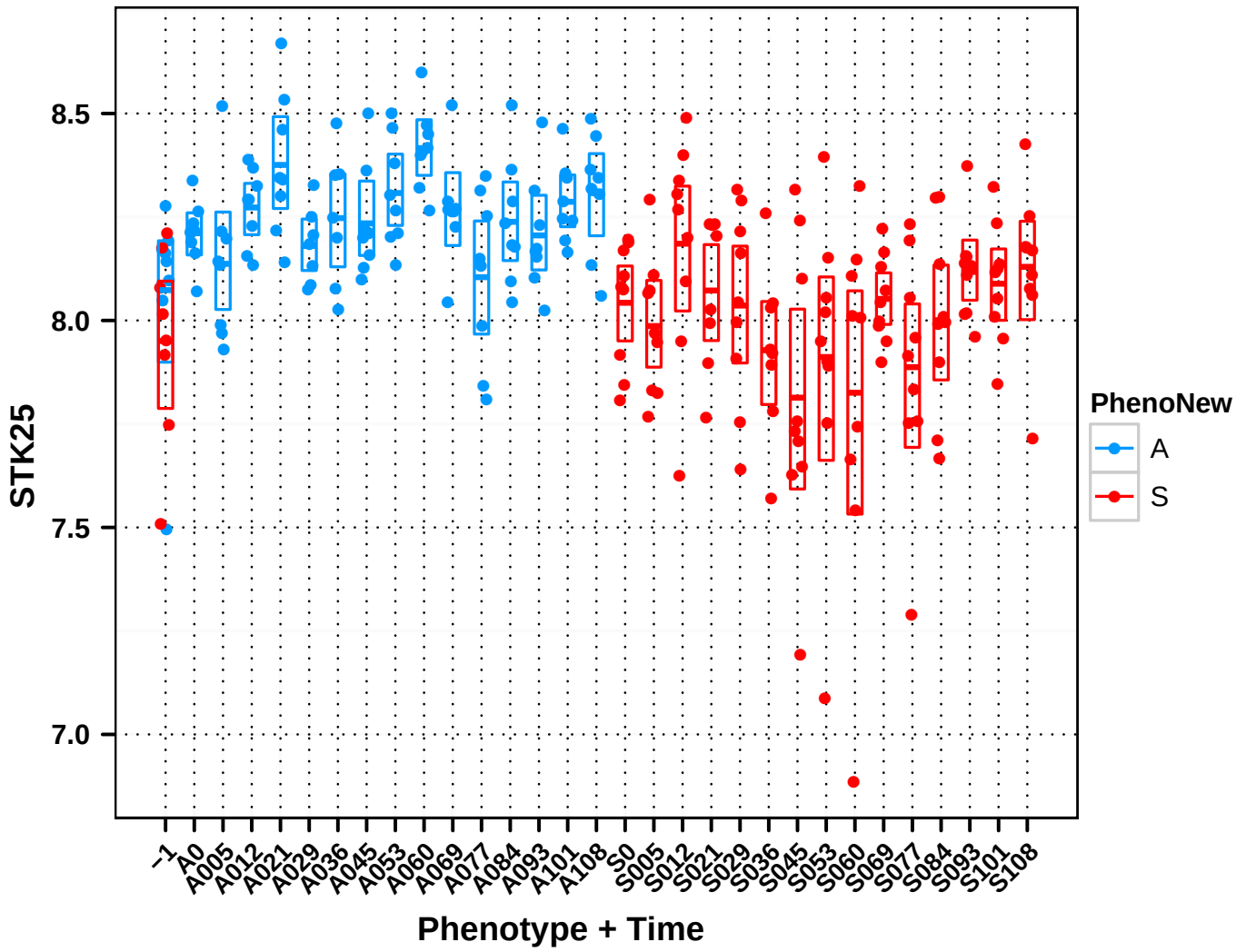


Figure S2.10: Temporal gene expression of cluster 6 gene serine/threonin kinase 25 (STK25 or SOK1). The expression intensities are plotted on a log base 2 scale.

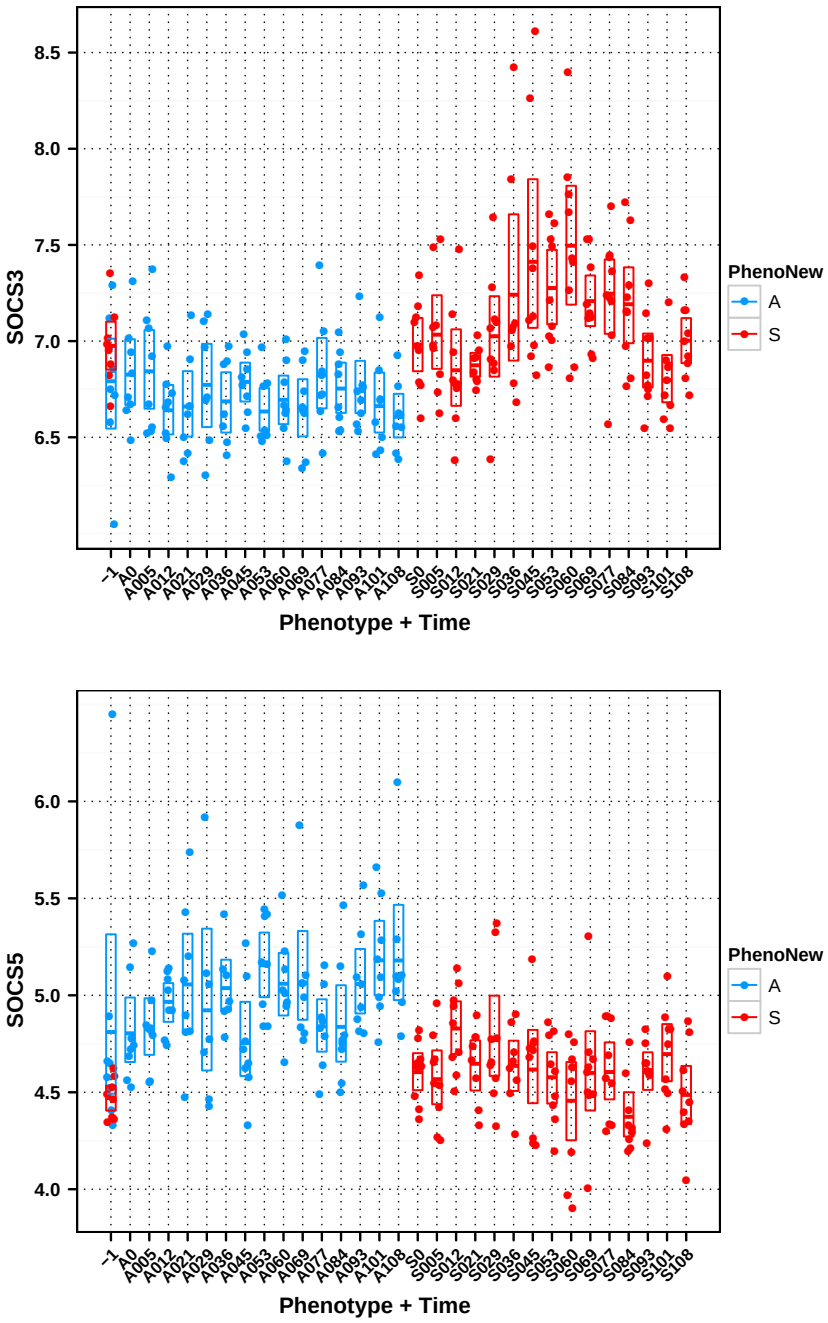


Figure S2.11: Temporal expression of genes from the family of suppressor of cytokine signaling (SOCS), including cluster 2 gene SOCS3 and cluster 6 gene SOCS5. The expression intensities are plotted on a log base 2 scale.

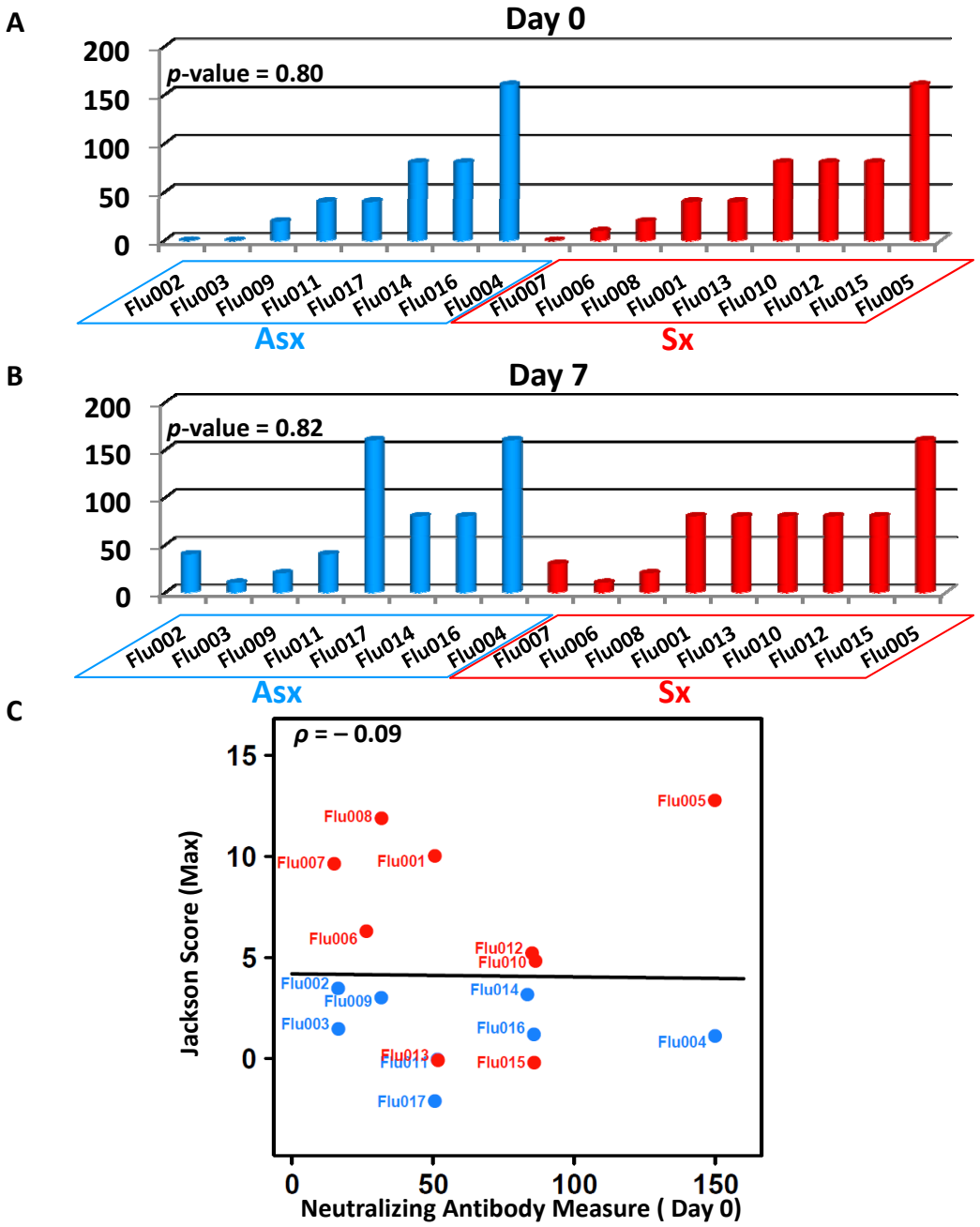


Figure S2.12: Neutralizing antibody (nAb) measure prior to inoculation shows no significant phenotypic difference and is not correlated with disease outcome. A, B) nAb of all subjects at Day 0 (A) and day 7 (B). No difference were observed between Asx and Sx on both days (non-parametric rank test). C) No evident correlation between nAb on Day 0 and maximum Jackson standardized score. A linear regression fit of score on nAb readings is shown in dark black line. Correlation test was performed using Spearman test. D) nAb increased in both Asx and Sx subjects from day 0 to day 28. ⋆No sample available on day 28.

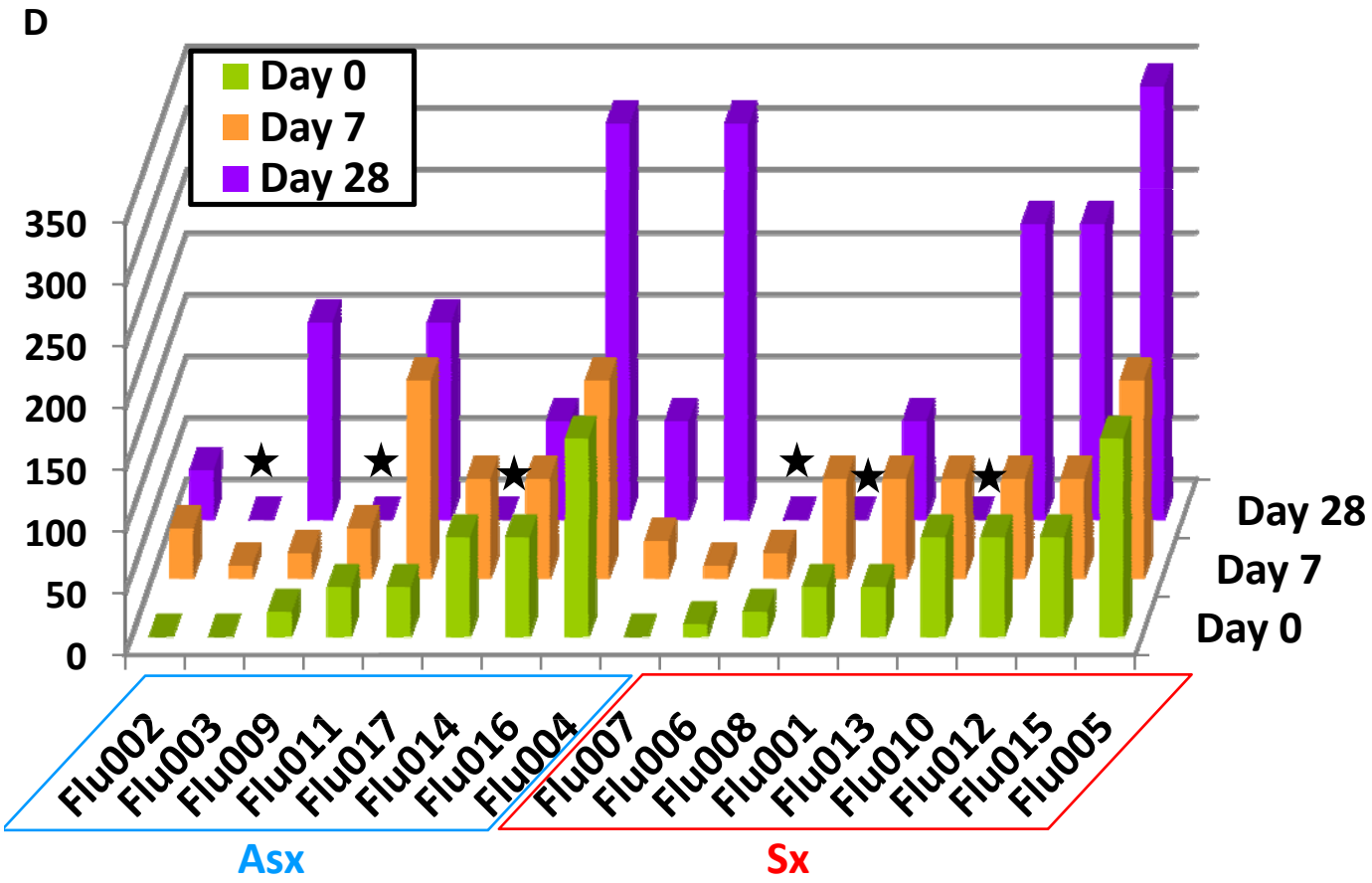


Figure S2.12: Neutralizing antibody (nAb) measure prior to inoculation shows no significant phenotypic difference and is not correlated with disease outcome (Ctd). D) nAb increased in both Asx and Sx subjects from day 0 to day 28. ⋆No sample available on day 28.

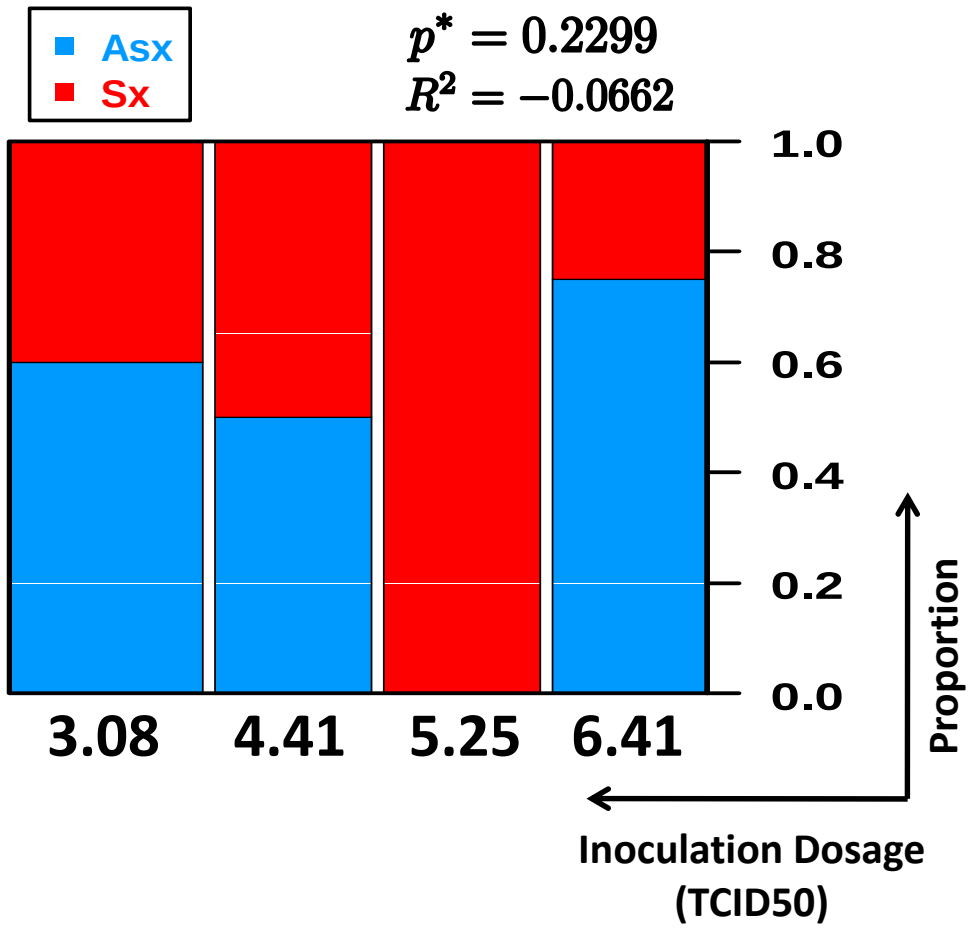


Figure S2.13: The infection outcome is independent of the dosage of viral inoculation. Each bar represents a randomized group of four to five subjects receiving a varying dose of Influenza A virus inoculation (Supplementary Materials) at day 0. Within each group, the subjects are divided into either Sx (red) or Asx (blue) subgroups based on clinically determined disease outcome. Fisher exact test was performed to test whether the dosage has any effect on disease development.

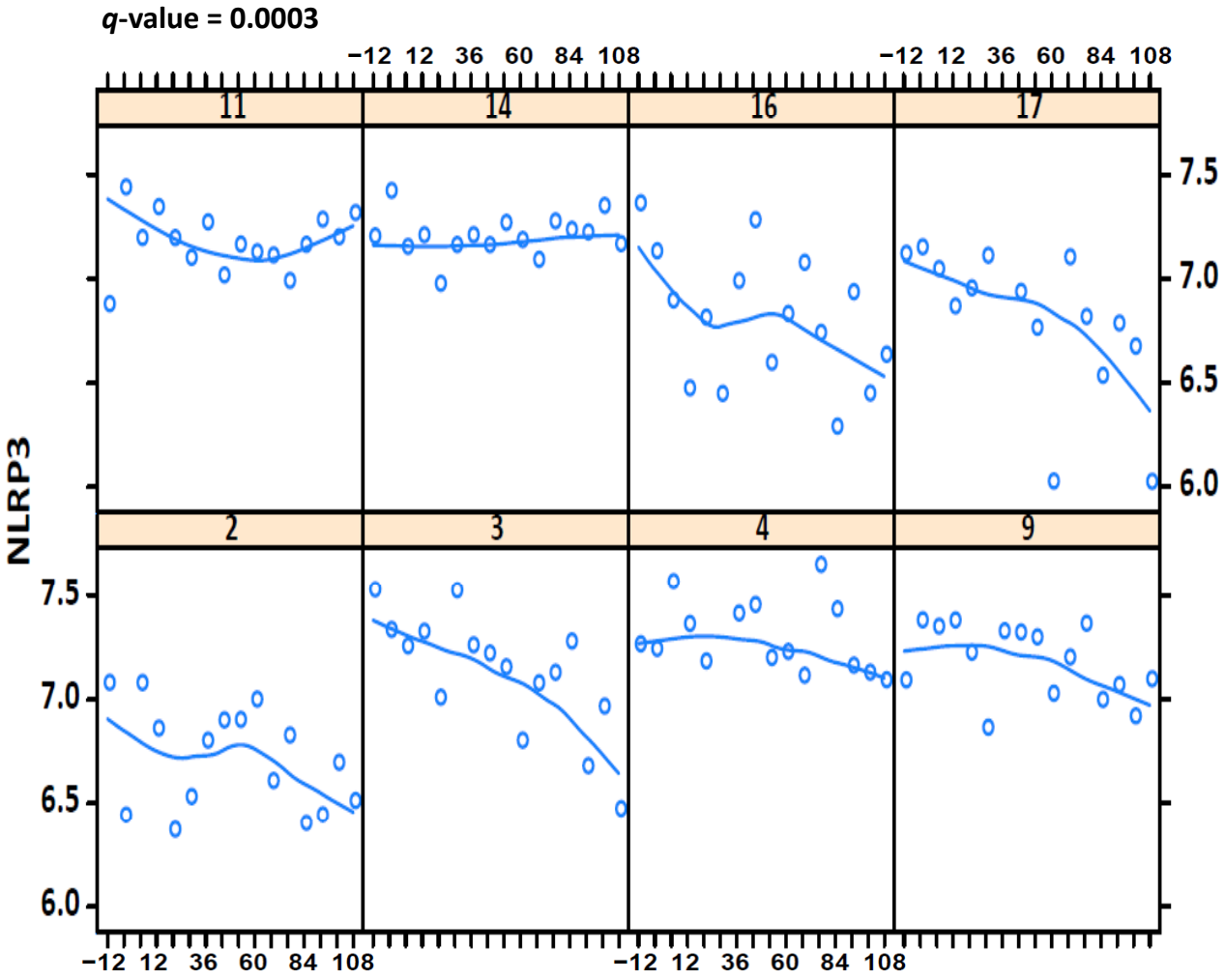


Figure S2.14: Asymptomatic subjects demonstrated non-passive transcriptional response program. As an example, we show a significant temporal expression decrease of the inflammasome related gene NLRP3 in eight individual asymptomatic subjects. Each subpanel depicts the temporal expression of one individual asymptomatic subject. The y-axis is the log base 2 signal intensity y of NLRP3 and the x-axis is the time from -12hpi to 108hpi (hour post inoculation). A polynomial fitting of expression values (solid line) was fitted using LOESS model and significance of temporal trend was assessed with EDGE. Subjects #3 and #17 never showed detectable amount of virus (< 1.25) in their nasal wash (Table S2.2).

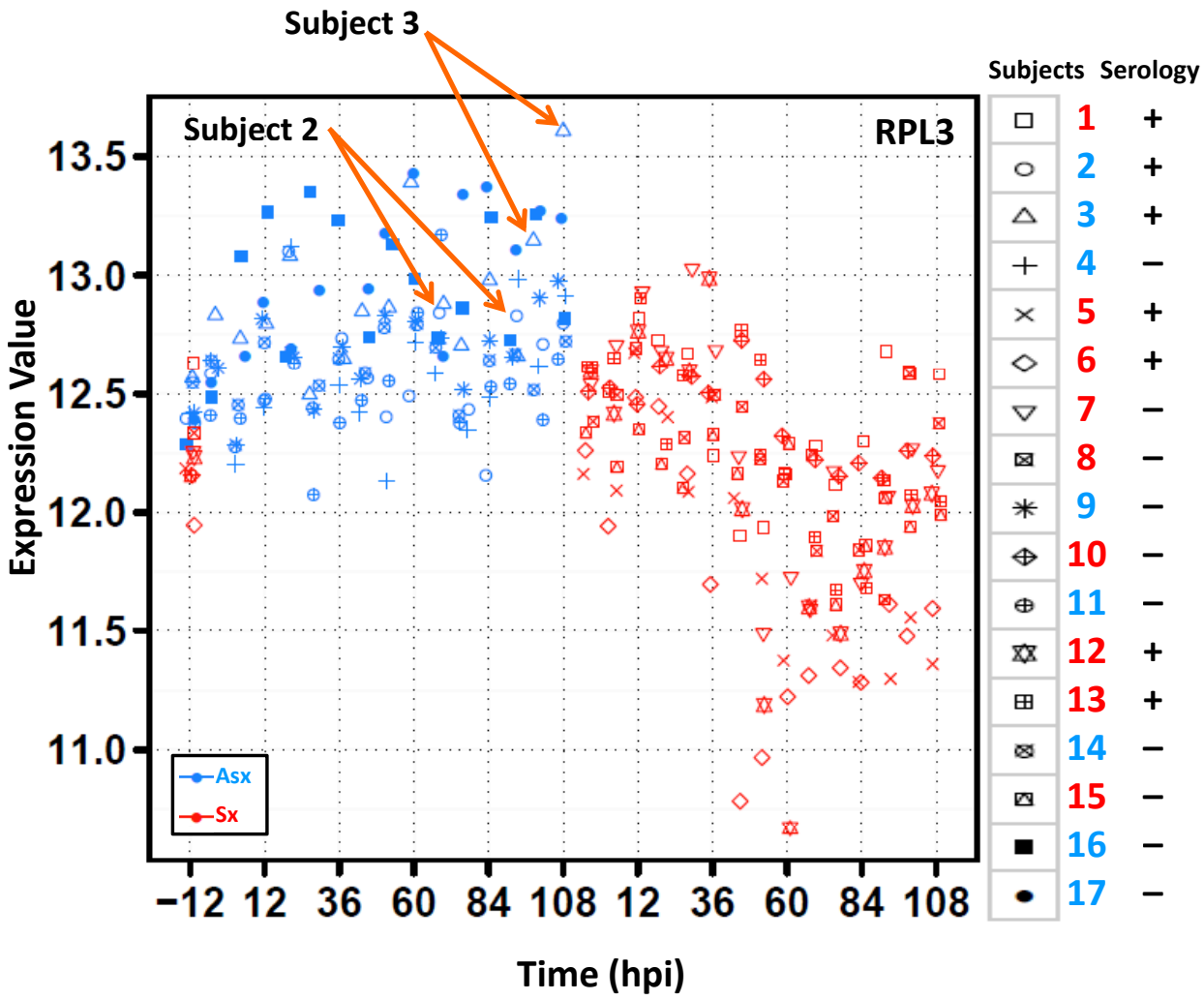


Figure S2.15: Serological conversion versus clinical symptom outcome and gene expression. The RPL3 gene expression trajectories for Asx (blue) and Sx (red) are representative of SOM cluster 6. Legend at right gives the character encoding of each subject along with their disease outcome ('blue' Asx and 'red' Sx) and their serologic conversion outcome ('+' converted and '−' no converted). There is no significant relation between disease outcome and serological conversion (p-value of 0.27 according to likelihood ratio test of dependency between these two outcomes). The two seroconverted asymptomatic individuals (subject #2 and #3) are called out by orange arrows in the gene expression trajectory plot. The RPL3 expression profiles of these two subjects are not significantly different from those of the other asymptomatic hosts.

CHAPTER III

Towards Early Detection: Temporal Spectrum of Host Response in Symptomatic Respiratory Viral Infection

3.1 INTRODUCTION

Exposure to pathogenic viral agents such as Influenza viruses (FLU), human rhinoviruses (HRV), and respiratory syncytial virus (RSV) are necessary, but not sufficient, for healthy human hosts to develop symptomatic respiratory illness. The interplay between hosts and pathogens, especially during the very early stage of infection, is thought to play a critical role in the development of overt symptoms (Hendley, 1983, Turner et al., 1982). The human system has evolved sophisticated immune mechanisms to fight against viruses. Yet viruses constantly prove themselves to be experts in devising effective counter measures to dodge host defense mechanisms. Individually, HRV accounts for 50-80% of upper respiratory tract infections (Gern and Busse, 1999). RSV remains to be the leading cause of lower respiratory disease in young children under age five. A recent study showed that reinfection of RSV in adults is more common than previously recognized and moderate severity can cause higher-than-estimated health issues in the elderly (Falsey et al., 2005). Seasonal FLU is highly contagious and infects 10 – 20% of the population each year (Belsey et al., 2006). Together, these three viruses contribute to the majority of respiratory tract infectious diseases. Furthermore, infections caused by them often compromise the

immune system and set the stage for subsequent development of other types of infections and diseases such as asthma, COPD etc (Gern and Busse, 2002). The prevalence of these infections and related complications, hospitalizations, and mortality poses serious issues for public health. Recent Influenza H1N1 outbreak, also known as Swine flu, further underscores the importance and emergency of a better understanding of disease dynamics. Enhanced capacity in early detection of pathogen exposure and forward prediction of disease progression is essential to successful disease management and pandemic planning.

In recent years, a plethora of seminal studies on pattern-recognition receptors (PRRs) and related signaling cascades have tremendously advanced our understanding of innate immunity (Kawai and Akira, 2007, Stetson and Medzhitov, 2006, Kawai et al., 2004, Honda et al., 2005, Takaoka et al., 2005, Yamamoto et al., 2003). Many elegant experimental analyses further elucidated the mechanistic activation and modulation of host response to invading pathogens (Ichinohe et al., 2009, Yoneyama et al., 2004, Huang et al., 2001, Zhu et al., 2008, Fenner et al., 2006, Ryo et al., 2008, Proud et al., 2008). By design, however, host responses in these experimental conditions are often characterized for individual cells via cell culture; or they represent a snapshot of the immune response pertaining to static or limited number of time points. Yet the components of host immune system are diverse and they act cooperatively in a complicated manner. Owing to both technical and ethical difficulties, we are unable to experimentally determine the full course of immune responses leading to symptom development in otherwise healthy hosts. Beginning with pathogen recognition, the detailed time sequence and orchestration of host responding events remain to be fully understood. Given that peripheral blood contains all key elements of the immune system, we hypothesize that it can be used to monitor the temporal dynamics of host-virus interactions and the temporal trajectory of host response may help address the pathogenic nature of viral infections. Analyzing genetic and

proteomic expression profiles of healthy human subjects challenged with FLU, HRV, and RSV viruses, we studied the full temporal spectrum of virus-mediated disease dynamics. This report offers an hour-by-hour detailed view of host immune response as a continuum, spanning the time from exposure to peak symptom manifestation. We present a robust classification model for estimating current disease state with implicit forward prediction capability for final outcome forecasting.

3.2 RESULTS

In this section we describe results of two different types of analysis. Both analysis methods can be interpreted as clustering methods. Two variants are used: similarity clustering and discriminatory clustering. The similarity clustering looks for clusters of genes whose temporal expression profiles are most similar, with respect to a given similarity measure; the discriminatory clustering looks for groups of genes that are collectively best for predicting or discriminating between different pairs of classes, e.g., Sx versus Asx or pre-onset Sx versus post-onset Sx.

3.2.1 Similarity clustering: analysis of differential expression for temporal profiling

Temporal differential gene expression analysis

Three cohorts of healthy human volunteers, each consisting of 20 subjects, received inoculation of live viruses FLU, HRV, and RSV. Respectively, 9, 10, and 9 in each challenge study developed mild to severe symptoms (supplementary methods). Gene expression profiles (GEP) were measured using whole blood drawn from 57 subjects (3 volunteers were taken off FLU study for safety concerns) at an interval of 4∼8 hours post challenge till peak symptom developed, which were roughly around 86, 72, and 108 hours (Table S3.1). Total of *267*, *198*, and *237* arrays were obtained and analyzed. We sought

to identify genes whose expression levels changed significantly between asymptomatic (Asx) and symptomatic (Sx) subjects during the course of the viral challenges. Using EDGE (Storey et al., 2005) with FDR adjusted p-value $< 1\%$ (supplementary methods), we determined 5007, 1210, and 632 (except in RSV nominal p-value was used due to less densely sampling) significant genes for FLU, HRV, and RSV respectively. The changes in gene expression are dramatic both in terms of sheer number of transcripts affected and the magnitude of such alterations. More importantly, these differential genes showed significant consistency among three challenge studies with $>90\%$ of genes from RSV and $>83\%$ of HRV genes overlapping with FLU genes (Figure 3.1C). This suggests that a similar process of symptom development is shared by all three viral pathogens; thus there exists a pan-viral process. It is worth noting that more than half of these significant genes also showed marked time course changes in Asx subjects alone. Yet the changes observed in Asx phenotype differ from those in Sx in two important aspects - the magnitude of the changes and the particular time point when such changes occur. It is our goal to focus on the phenotypic differences that could potentially contribute to the pathogenesis of symptomatic illness

Similarity clustering using Kohonen Self Organizing Maps.

As in previous study of immune host response (Huang et al., 2001), Kohonen's Self-Organizing Maps (SOM) was applied to cluster significant genes according to their temporal kinetics (supplementary methods). The resulted manifold consisting of units, or clusters, and their associated prototypes are shown in Figure 3.1. Each prototype represents the pattern of an individual SOM cluster. Figure 3.2 displays the average magnitude of the expression level of each SOM cluster and the corresponding error bars, providing

a direct visual assurance of the purity and variations of these clusters. The polar plots in Figure 3.1A provide a different visualization of the differences between temporal gene expression profiles for Asx and Sx phenotypes. Each polar plot depicts the expression pattern shared by genes of a SOM cluster. Within a plot, the temporal expression of Asx resides on the top portion of the circle while Sx expression occupies the bottom half. Each phenotype's expression values are placed in time sequence, increasing in the counterclock-wise direction, inside its own half circle. *Consequently, the expression profiles of Asx and Sx at any given time point can be compared at opposite ends of a radial line passing through the polar origin.* Such symmetric arrangement facilitates visual examination of contrasts in phenotypic gene expression patterns. For example, a fan shape extending over +/-90 degrees (FLU cluster 7) represents a cluster of genes that are expressed early in Asx and late in Sx. A fan shape over 0 to 180 degrees (FLU cluster 6) represents strong expression in Asx and weak expression in Sx at all time points. We emphasize that it is not adequate to only look at one phenotype alone or the ratio of Sx/Asx expression values. *This is because of the fact that both Asx and Sx undergo significant changes in gene expression profiles, a consequence of universal protective immune response.* In Figure 3.1B we show heatmaps for the top 20 genes from each SOM cluster having the most significant differential expression. The reader may find that segment plots of SOM clusters add interpretability to the heatmaps of temporal expression patterns, allowing more direct simultaneous comparisons between particular time points and phenotypes.

In the FLU study, we identified a total of 8 SOM clusters, each characterized by their particular prototypes (centroids). These prototypes differentiate different temporal signatures of host response associated with Sx versus Asx phenotype. The clusters include genes that show differential expression at early (4 – 18hpi), middle (18 – 48hpi), and late (>48hpi) stages of infection. The differential expression dynamics of these prototypes

either sustain throughout the entire challenges or extinguish after a short period of time. These clusters are analyzed separately below with biological relevance extracted from canonical pathway analysis (supplementary methods).

FLU Cluster 1 consists of genes that are relatively underexpressed in Sx and rapidly increase at 45hpi. The expression profile of the Asx group transiently increases between 36 – 84hpi. Some interesting genes in this cluster are CCL4, TBX21, CD74, CD244, HLA-DMA, and IL10RA, all are involved in regulating immune cell trafficking and antigen presentation.

FLU Cluster 2 includes genes exhibiting sustained downregulation unique to Asx phenotype. In Sx, the expression of these genes increase to their peak level at the middle of challenge (45 – 69hpi), followed by a rescinding trend. The most relevant biological function involved by this group of genes is chemotaxis and activation of critical components of innate immune response — macrophages, eosinophils, neutrophils, and plasmacytoid dendritic cells (pDCs). Activations of these cells result in strong inflammatory response. Some genes in this category are CD86, TGF-β, CCL11, CCL8, CEACAM3, CXCL9, CXCL10, IL1B, TRIM21, and many Toll-like receptors (TLR1/TLR2/TLR4/TLR5/TLR8). The transcript of a key TLR4 adaptor protein, TRIF, is also found in this cluster. The TRIF-dependent pathway activates type I interferons (IFNs) as well as CXCL10 (Zhu et al., 2008). Interestingly, two suppressors of cytokine signaling (SOCS) genes, SOCS1 and SOCS3, also reside in this cluster. They are known to negatively regulate the activation of macrophages and DCs (Fenner et al., 2006, Ryo et al., 2008).

FLU Cluster 3 is characterized by strong activation, in Sx phenotype, of genes responsible for proinflammatory responses. Compared to cluster 2, genes in cluster 3 remain overexpressed many hours after symptom peaked (60hpi). Roughly, these genes can be divided into three major categories. The first one includes genes in MyD88-dependent

TLR signaling pathway — TLR7, MyD88, IRF5, IRF7, and IRF9. Two other PRRs genes RIG-I and MDA-5 are also located in this cluster, whose recognition of dsRNA viruses activate type I IFN signaling (Yoneyama et al., 2004). The second category encompasses many core IFN-inducible antiviral genes such as PKR (EIF2AK2), ISG15, OAS1, OAS2, OAS3, OASL, RSAD2, MX1, IFIH1, IFI27, IFI44, IFIT1, IFIT3, IFITM, TRIM5/22/38, Ly6E, TRAIL, HERC5, SAMD9, and USP18. It is worth noting that none of the type I/II interferons showed significant phenotypic differences. The third category of genes includes proinflammatory cytokines TNF, IL15, and IL10, a known mediator of inflammatory response.

FLU Cluster 4 contains genes that show sustained downregulation in Sx phenotype in contrast to slight or no change in Asx. Major biological functions of genes in this cluster are iCOS signaling, oxidative stress, and calcium induced T cell apoptosis etc. Actg1, c-Jun (AP1), CD247, CD40LG, CAMK4M, protein kinase C family genes, and phosphoinositide-3-kinase family genes all belong to this cluster. CD27, the key receptor on TLR signaling pathway that activates NF-*k*B and MAPK8/JUN is also located in this cluster. A point of interest about this cluster is the presence of *SIGIRR*. An endogenous inhibitory member of TLR-IL-1R superfamily, SIGIRR was reported to regulate inflammatory response, especially TLR4-mediated signaling (Wald et al., 2003).

FLU Cluster 5 associates Sx phenotype with a delayed downregulation pattern at 36hpi in comparison to much earlier decline (5 – 8hpi) in Asx phenotype. Genes with significant phenotypic differences such as AOC3, IL1RAP, THBD, IL1R1, CAMK1D are known to function in antigen presentation and accordingly innate immune response.

FLU Cluster 6 is populated by genes that begin to show marked decrease at 29hpi in Sx phenotype, returning to baseline after 60hpi. In contrast, in Asx phenotype they steadily increased till 108hpi. This cluster represents genes involved in protein synthesis

(especially ribosomal proteins), immune response, and RNA trafficking. A few examples are RPS3, RPS6, RPS14, RPS16, RPL4, RPL13, iCOS, EIF3 family genes, CCR7, CD8A, CD8B, CD79A, DDX18, DDX47. The unusual saturation of genes related to ribosomal protein synthesis suggests a role by ribosomal proteins in limiting viral replication.

FLU Cluster 7 features a transient (0 – 36hpi) downexpression of its member genes in Sx phenotype versus a sustained high level of expression in Asx. One notable gene in this cluster is *Siglec7*. Expressed in NK cells, it functions as an inhibitory receptor (Falco et al., 1999). The primary biological functions implicated by this cluster of genes are NK cell signaling and cell death.

FLU Cluster 8 contains genes related to cell morphology and cell signaling. A few examples are LY96, HOOK1, PRDX2, ODC1, EIF2AK1, BCL2L1, KRAS, and RGS1. In Sx subjects, these genes show no change or slight transient increase. On the other hand, in the Asx subjects these gene start from a relatively lower expression level and increase sharply after inoculation. Note that the variation of genes in this cluster is higher than that of the other clusters.

Pan-viral temporal gene expression patterns.

Following the same SOM clustering approach as used in FLU, discussed above, we identified 6 and 3 temporal gene clusters from the HRV and RSV challenge studies, respectively (Figure 3.1A-B). It appears that these clusters share characteristic resemblance to their FLU counterparts. To determine the temporal patterns common and distinct to individual viral entities, we aligned the prototypes of the three challenges on the same time scale. For each SOM cluster, expression values at every 6 hours are derived from a spline fit to the centroids. A total of 48 data points per cluster were generated within the time frame 0 –

144hpi, with 24 data points for each phenotype. Unsupervised hierarchical clustering was then performed on the prototypes and the dendrogram is shown in Figure 3.3A. Essentially, the SOM clusters can be divided into two major groups according to the contrast between Sx and Asx phenotypic expression patterns — downregulation (upper half) and upregulation (lower half). Pairwise between-cluster associations were quantified using temporal Pearson correlation coefficients between the clusters (Figure 3.3B). In the figure, the solid circles denote positive correlations and open circles negative correlations. Circles with red border indicates correlation magnitude greater than 0.9.

Upregulated pan-viral clusters:

Several clusters of upregulated genes showed strong similarities with each other. In particular, two pairs (FLU2, HRV5) and (FLU3, RSV3) exhibited nearly identical temporal features. *Such striking agreement of the temporal patterns from three distinct pathogenic agents clearly demonstrates the existence of a pathogen-independent host response signature.* In fact, out of the 395 common genes (Figure 3.1C), 310 (78%) of them are from FLU2 and FLU3 clusters. In contrast, these two clusters of genes only account for 29% of significant genes in FLU ($p<0.0001$). Similar conclusions can be drawn for HRV (74%) and RSV (83%).

The most prominent subset of these pan-viral genes is the group of IFN-inducible antiviral genes, including PKR, ISG15, ISG20, OAS1/2/3, MX1/2, IFI27, IFI44, IFITM1/3, IRF1/2/5/7/9. Interestingly, the family of four tripartite motif-containing (TRIM) genes, TRIM5/21/22/38, are all activated with similar kinetics. There has been data supporting their activation by interferon stimulating signals and their role in mediating antiviral response via E3 ubiquitin ligase activities is well known (Ozato et al., 2008, Carthagena

et al., 2009). Collectively, these effector genes create and maintain in symptomatic hosts an unusually high level of antiviral state. In parallel, our results show strong pan-viral induction of a panel of proinflammatory cytokines and factors such as TNF, IL1B, IL15, CASP1, CASP5, CCL2, CCL5, and STAT1. To our knowledge, upregulation of IL15 has not yet been identified as a pan-viral marker of symptomatic infections. A recent study has demonstrated a direct link between the upregulation of IL15 and the level of peripheral T and B cell immune activation in HIV/HCV coinfections (Allison et al., 2009). It is difficult to determine the exact source of these mediators, given the wide variety of cells in the peripheral blood. *Nevertheless, our results clearly demonstrate the Sx-specific convergence of signaling towards a pathogenic inflammatory response, possibly as a result of increased virus replication activities.*

Downregulated pan-viral clusters:

Among the clusters exhibiting a downregulation trend in Sx phenotype are: FLU4 and HRV1, whose gene expression prototypes have similar patterns. We did not directly observe a RSV cluster with an exactly same temporal pattern as that of the (FLU4, HRV1) pair. However, RSV1 has sufficient similarity to be considered a possible match. Unlike (FLU4, HRV1), the RSV1 cluster begins to decrease at 36hpi and continues declining until the end time of the RSV challenge. Considering that RSV has longer incubation time, we speculate that we would have observed a similar expression pattern as that of (FLU4, HRV1), had the available challenge study assays continued beyond 144 hours. In fact, RSV does appear to exhibit the slowest symptom onset time when compared to FLU and HRV.

Significant correlation between SOM clusters and symptom scores.

We next investigated the existence of associations between gene expression profiles and clinical disease severity, as determined by the Jackson criteria. Clinical symptoms were recorded twice daily using standardized symptom scoring and included measurements of viral tiers (supplementary methods). Among all the SOM clusters, we found that gene expression of FLU3, RSV2, and HRV5 have significant positive correlation with symptom (Figure 3.4). The positively correlated FLU3 cluster clearly mimics the disease progression measured by symptom scores (Figure 3.4 right panel). In contrast, we observed a time lag in HRV. The expression values of HRV5 genes did not peak until 42 – 48hpi, which is essentially 0.8T of symptom peak time. This could reflect the lower severity of symptoms in HRV as compared to RSV and FLU (Table S3.2). Notably, the observed associations are all significantly higher than by chance, as estimated by random permutation of expression values ($p<0.0001$). Such high degree of correlation further supports our hypothesis that gene expression pattern is strongly associated with symptom development and may be used to discriminating symptomatic from subclinical infections.

Corroboration with multimodal analyte assays.

To further our understanding of potential factors and signaling events involved in virus-triggered host immunity, we complemented our analysis with other modalities of high-throughput data, including data from protein array, mass-spectrometry, and clinical laboratory tests. Using a custom antigen array, we surveyed 90 known immune response related antibodies in 328 plasma samples. These samples were selected to capture six times points that are of critical clinical importance — baseline, prechallenge, 0.1T, 0.2T, 0.8T, and T

(where T denotes peak symptom time). Similar to gene expression analysis, we identified temporal differentially expressed proteins between phenotypes (supplementary methods). Perhaps not surprisingly, the strongest signals are from the later time points after inoculation (Figure 3.5A-B). In particular, B2M, CRP, IL10, and TNFRSF1B are pan-virally upregulated at late phase (0.8T and T) in the Sx phenotype. This is consistent with their roles as effectors and mediators of inflammatory responses or apoptosis. Their upregulation is also indicative of the host immune system's efforts to limit detrimental effects of inflammation. Notably, the relative lower expression level in Asx suggests their potential utility in subclinical diagnosis of infection.

Following a similar pattern, many virus-specific proteins showed late stage increase. Among them, inflammatory cytokines IL5 and CCL4 are shared by HRV and RSV. In addition, IL18 is upregulated in FLU and RSV, and this analyte is known to induce activation of TNF (Iannello et al., 2009). Interestingly, FLU and RSV also showed overexpression of alpha-1-antitrypsin (SERPINA1), which is known to function as an anti-inflammatory mediator (Janciauskiene et al., 2007). On the other hand, the increase of ICAM1 is FLU specific, suggesting an increased cell-cell interaction and adhesion activity in FLU. FLU and HRV both showed elevated expression of IL1R1 with its coding transcript also significantly upregulated in FLU5. It is in RSV that we noticed higher CCL11 expression level associated with Asx phenotype at as early as 0.1T. Since CCL11 is responsible for local recruitment of eosinophils, this suggests the engagement of eosinophilic inflammation at early stage in RSV. Yet the lack of symptoms is intriguing. When examined in the context of temporal gene expression, these proteins showed both positive (FLU2, FLU3, RSV2, RSV3, HRV4, HRV5) and negative (FLU4, FLU5, HRV1, and RSV1) associations with clusters we identified using microarrays (Figure 3.5C). Further analysis of genes in these clusters will likely provide more information about the sequence of signaling events

preceding protein synthesis during host response.

In parallel, we conducted unbiased proteomic analysis, focusing on FLU pre-inoculation and peak symptom time points (supplementary methods). A total of 16 differentially expressed protein compounds were verified with >2 peptides and they showed discriminatory value in separating Sx phenotype from Asx and baselines (data not shown). Among these 16 proteins, 7 of them are also present in microarray and antigen arrays. At peak symptom time, activated expression levels were seen in Serum amyloid A1 (SAA; FLU2), KIAA0748 (FLU6), and beta-2-microgolubin B2M (antigen array). Another 4 molecules showed downregulation, including cofilin (CFL1; FLU7), transgelin 2 (TAGLN2; FLU5), vinculin (VCL; FLU7), and gamma actin 1 (ACTG1; FLU4). Strikingly, the changing direction of these proteins measured by mass spectrometry are in complete agreement with the SOM cluster prototype of their coding mRNAs (Figure 3.1A, 3.5A-C).

Overall, the results obtained from multiple modalities of data are mutually consistent and corroboratory.

3.2.2 Discriminatory clustering: Analysis of differential expression for disease state prediction

The differential expression analysis performed in the previous section has the power to explain functional categories and elucidate related biological pathways. However, such explanatory power is not necessarily prescriptive, i.e., it may not specify genes that best discriminate or predict different states of health and disease. In this section we turn to the discovery of genes that are particularly effective in such prediction. Our goal is to construct a classifier to detect exposure to pathogens and to predict the clinical outcome in terms of symptom presentation. We accomplished this by combining the power of Bayesian factor analysis for unsupervised segmentation of the pre Sx and post Sx phases of infection and boosted ensemble classifiers for supervised classification of the gene microarray samples.

We applied a Bayesian factor analysis method called Mixed Component Analysis (MCA - supplementary methods) on the gene expression microarray data. In brief, MCA discovered a group of genes, called a factor, that clearly distinguished Sx from Asx subjects and delineated the onset time separating pre-symptom from post-symptom phases of the Sx subjects (Figures 3.6A-C). Based on this MCA factor, we identified four pivotal phases in the course of disease progression, namely pre-challenge, Asx post-challenge, Sx pre-onset, and Sx post-onset. An example of these four phases in FLU is shown in Figure 3.6A with sampling regions labeled as 1, 2, 3, 4, respectively, in the form of sample-time by subject design matrix. The color levels in Figures 3.6A represent the factor loadings that indicate the correlations of each sample (gene chip) to the MCA factor. In specific, the MCA factor is a relative expression profile of 268 genes that are represented in this factor (Table S3). The larger this factor loading is in a sample, the warmer (red) this sample's representing color is. With high resolution, this factor clearly reveals the critical onset transition point of the transcriptome profiles in each symptomatic individual. A direct correlation between the factor loading signature and the disease severity is further validated by the symptom score chart shown in Figure 3.6B.

Using the four class segmentation represented in the FLU matrix of Figure 3.6A, or the 10 class segmentation associated with the all-virus matrix of Figure 3.6C, we evaluate the intrinsic difficulty of discrimination between different pairs of regions using state-of-the-art boosted classifiers from machine learning. A boosted classifier selects a subset of genes that optimize the tradeoff between overfitting a training set and maintaining accuracy on a test set. Each pair of classes will have its own best boosted classifier along with the associated most discriminating subset of genes. Thus this application of boosted classifiers is a "discriminatory clustering method" which clusters groups of genes in terms of their class discrimination power. This discriminatory clustering method is in direct contrast to the

SOM method that clusters genes according to the similarity of their temporal expression profiles, as described in 3.2.1.

Two classification methods with different base classifiers were used: a *LogitBoost* classifier and a *random-forest* classifier (supplementary methods). These two classification methods yielded similar results. For lack of space we here only report results obtained from the *LogitBoost* classifier. Briefly, boosting of logistic linear model was applied in the following manner. Using bootstrap resampling we generated 1,000 bootstrapped copies from the original data. Each bootstrap copy was separated into training (70%) and test (30%) subsets. A classifier was built on the training data only, blindfolded to the test set. After 1,000 randomized runs, we selected for the final classifier those genes that entered the model at least one third of the time. We then measured and reported the performance of the classifier, which uses only the small set of genes, in terms of its receiver operating characteristic (ROC) curve.

The receiver operating characteristic (ROC) curves in Figures 3.7, 3.8 and 3.9, along with their 95% bootstrap confidence intervals, indicate both the relatively satisfactory performance and difficulty in discriminating between different disease states with gene expression: pre-inoculation (class 1), pre-onset (classes 3, 6, 9), post-onset (classes 4, 7, 10) and asymptomatic (classes 2, 5, 8). Figures 3.7, 3.8 and 3.9 also lists the predictor genes that accompany each ROC and define symptom-discriminatory clusters of genes. On the basis of these findings, we make several remarks.

Remark 1: Gene groups that discriminate between Asx and Sx early host response differ from Sx late host response discriminants.

The gene lists associated with high quality ROC curves (2,3), (5,6), and (8,9) discriminate between Asx post-inoculation and Sx pre-symptom, i.e., Asx and Sx early host response,

while those associated with ROC curves (2,4), (5,7) and (8,10) are best for discriminating between Asx and Sx late host response. Notably, the early Sx and late Sx discriminants include some well known immune response genes, such as OAS1 and CD177. However, there are differences as well, especially in FLU. For example, APOLD1 (2 vs 3 in FLU) is a gene transcript corresponding to an endothelial cell early response protein. Interestingly, it is also included in the list of discriminants between baseline and Asx post-inoculation (1 vs 2 in FLU) but is absent from the list for discriminating baseline and Sx post-inoculation samples (1 vs 3 in FLU).

Remark 2: Among all phases of host response discrimination, detection of early post-inoculation against baseline is the most difficult.

Discrimination between baseline and post-inoculation/pre-symptom classes generally has the worst performance (ROC pairs (1,2), (1,3), (1,5), (1,6), (1,7), (1,8)) for all viruses. Interestingly, an exception is FLU for which it is relatively easy to discriminate between baseline and Asx post-inoculation (1 vs 2) but more difficult to discriminate between baseline and Sx pre-onset (1 vs 3). Whether this is due to an enhanced virus-induced suppression of early host response in the Sx subjects as compared with Asx subjects is certainly of interest to the study of FLU infections.

Remark 3: Gene clusters that discriminate between Sx post-symptom and Asx post-inoculation classes are consistent with Zaas et al. (Zaas et al., 2009)

The ROC pairs (2,4), (5,7) and (8,10) are associated with genes that discriminate between Sx post-onset and Asx post-inoculation classes. The associated lists of gene discriminants include OAS1 (2 vs 4), RSAD2 (5 vs 7) and CD177 (8 vs 10), respectively, for FLU, HRV and RSV. Note also the presence of SMAD1 in the list of FLU discriminating genes be-

tween Sx pre-onset and Sx post-onset (3 vs 4).

Remark 4: Overall, FLU has the most powerful gene discriminants.

Among all challenge studies FLU has the best ROC curves, in terms area-under-the-curve (AUC), leading to more accurate discrimination between classes. This is consistent with the ANOVA analysis (results not shown) in that, as compared to HRV and RSV viral challenge studies, the FLU study produced an order of magnitude more temporally differentially expressed genes for a specified false discovery rate. The relatively better performance of FLU discriminants may also be a result of the relatively fewer missing samples in the FLU challenge data, possibly leading to better MCA segmentation of Sx onset time. The quality of the MCA segmentation can be inferred from the quality of the ROC for class pairs (3,4), (6,7), and (9,10). These class pairs are associated with discriminating Sx pre-onset from Sx post-onset: quality for the FLU pair (3,4) is somewhat better than that of the HRV pair (6,7) and RSV pair (9,10). We can eliminate this possible source of bias when the full complement of HRV and RSV samples becomes available in the follow-up study.

Figure 3.10 summarizes boosted quadratic classifier performance when the classifier is constrained at 10% false discovery rate. In some cases (e.g., FLU), fixing false discovery rate underestimates achievable performance. Nonetheless, it is useful for comparing across different viral entities. Overall, the prediction models showed good performance.

At average 10% level of FPR, the models attain an average TPR at 100% for Asx versus Sx postchallenge; 98% to 100% between Prechallenge versus Sx-postchallenge; 95% to 100% between Asx-postchallenge versus Sx-preonset; 97 to 100% between Sx-preonset versus Sx-postonset; 80% to 99% between Pre-challenge versus Asx-postchallenge; 65% to 91% for Asx-postchallenge versus Sx-preonset. Apparently, the prediction is most

challenging for separating prechallenge samples from either Asx-postchallenge or Sx-preonset samples. This could be due to the heterogeneous nature of these classes. Some Asx-postchallenge subjects might have effectively cleared the virus and fully recovered whereas some Sx-preonset subjects might experience slightly delayed response. Furthermore, the RSV model seems to be the least capable of accurate classification. This is likely due to less densely sampled data points in HRV and RSV that were available for training the classifier and performing variable selection. In terms of the predictor genes selected, three viral pathogens show little overlap (Figure 3.9B), which is in contrast to the significant extent of overlap among the full set of significant differentially expressed genes obtained using similarity clustering (Figure 3.1C). This suggests that while the transcriptional programs induced by three viruses are strong enough to elicit similar expression patterns (Figure 3.3A-B), these patterns are not sufficiently strong to be predictive between host response states. At the same time, our choice of classification algorithm may also play a role in that it only picks the most predictive variables and ignores their correlates. Some of the genes that are left out of the model can deliver similar level of prediction power, provided one accounts for their strong correlations to those included in the model (data not shown). The variables selected by random forest are in general consistent with boosting, as reflected by the Gini Index measure of variable importance (Table S4). Such coherent results from two independently implemented base classifiers demonstrate the robustness of our prediction model.

Remark 5: Differences and similarities between Influenza H1N1 and H3N2

It is worthnoting that despite the focus on influenza H3N2 viruses, the findings presented here are in good agreement with our preliminary results from influenza H1N1 viruses, especially in the aspects of disease signature detection and risk stratification (Figure S3.2,

S3.3). However, the model performance evaluated on H1N1 usin the same group of H3N2 discriminant genes does suggest that H1N1 and H3N2 infections result in a largely similar phenotype towards the late stage of infection. During the early infection stage, these two viruses seem to elicit rather different responses in the hosts as shown by he relatively poorer discrimination performance of the H3N2 model at early stages of H1N1 infection.

3.3 DISCUSSION

This report, to our knowledge, presents for the first time the analysis of a full temporal spectrum of pathogen-elicited host response in overt respiratory infections. Well-designed and meticulously organized, this study represents by far the most extensive in vivo human challenge models on multiple respiratory viral agents. Most important, the sophisticated human immune system can now be examined in a unified manner in which concerted immune responses can be studied as a whole. The temporal events and their associated genes found by our analysis are highly relevant to immunological responses. Many of them have been extensively studied and linked to host inflammatory processes. *Our results offer an opportunity to look beyond individual signaling events and into their collective modular effects on symptomatic disease pathogenecity.* For instance, it is TLR7, not TLR1 or TLR5, that directly detects viral components of ssRNA viruses. The pan-viral upregulation of TLR1 and TLR5 revealed by our analysis would then supports the notion of non-specific induction of cellular host response. This raises the question of whether they participate in reinforcing the production of MyD88-mediated proinflammatory cytokines. Relevantly, we also observed FLU-specific increase of TLR4 and TLR2 which are known to recognize RSV viral envelope proteins. A recent study further identified a synergistic cross-talk mechanism between the TLR4-MyD88-independent and MyD88-dependent pathways. In DCs, the synergistic act amplifies proinflammatory response and produces

a stronger cytokine profile (Zhu et al., 2008). Indeed, the pan-viral upregulation of cytokines and chemokines, TNF, CCL2 (MCP1), CCL5 (Rantes), and CXCL10 (IP-10), is strikingly similar to that of asthma. Such virus-mediated cytokine profile would result in selective recruitment and accumulation of eosinophil, monocyte, and neutrophils - known as inflammatory infiltrate - in the airways (Luster, 1998). Taken together, it warrants further investigations to ascertain the net effects of such simultaneous non-specific activation of multiple toll-like receptors.

Interferon regulatory factors (IRFs) have drawn great attention in various recent studies for their modulatory role in TLR signaling cascades. Out data showed pan-viral sustained overexpression of IRF1/2/5/7/9 in Sx phenotype. Among them, IRF7 regulates type I interferon signaling whereas IRF1/2/9 are involved in development and activation of NK cells by modulating the transcription of IL15 (Ogasawara et al., 1998, Lohoff et al., 2000). In accordance, IL15 showed similar dynamics as IRF1/2/9. *However, it is IRF5 that is of most relevance to the development of symptoms. As a downstream master regulator of TLR-MyD88 signaling pathway, IRF5 directly promotes the induction of proinflammatory cytokines.* IRF5-deficient mice were shown to be LPS-resistant (Takaoka et al., 2005). The pan-viral Sx-phenotypic overexpression of IRF5 found in our study suggests that further studies should be performed to clarify IRF5's contribution to the pathogenesis of symptomatic disease.

The circulating mRNA expression level of SOCS1 showed pan-viral early upregulation in Sx phenotype, contrasting to an Asx phenotypic suppression. SOCS1-deficiency has been associated *in vivo* with amplified type I interferon antiviral responses and reduced viral load. The SOCS1$^{-/-}$ IFNAR$^{-/-}$ mice also demonstrated prolonged survival without detrimental inflammatory damage in the host, suggesting a suppressive role of SCOS1 on pro-inflammatory influence by type I interferons (Fenner et al., 2006). Increased expres-

sion of SOCS1, resulted from a polymorphism in its promoter region, has been associated with the pathogenesis of adult asthma (Harada et al., 2007). Consistent with these studies, RSV induced increase of SOCS1 and SOC3 have also been linked to weaker antiviral response by the hosts (Ryo et al., 2008, Harada et al., 2007). *Therefore, it is tempting to speculate that Sx-specific upregulation of SOCS1 may interfere with interferon antiviral effects at very early times and cause increased viral replication.* This would then create a vicious cycle wherein unrestricted viral production further stimulates and intensifies host response as PRRs sense more viral components.

Virus-induced programmed cell deaths (apoptosis) suppose to prevent or ameliorate inflammation by removing infected cells during viral infections. In our study, strong t evidences of transcriptional apoptosis were observed concurrently with inflammatory responses in the Sx phenotype. They also closely correlate with disease severity. Besides death ligands TNF and TRAIL, which are common to all three viruses, FLU showed upregulation of FasL and protease caspase 10. Other related genes such as caspase 4/7 and Fas are observed in FLU and RSV whereas caspase 6 was seen in both FLU and HRV. It is clear that the three viral infections exhibit somewhat different caspase profiles related to apoptosis. In contrast, they agree unanimously on the two caspases that promote inflammation, caspase 1 and caspase 5. Simultaneously, NOD2 and RIP2 are co-activated with caspase 1 and 5 by all three viruses. Expressed in peripheral blood lymphocytes (monocytes), NOD2 belongs to another critical pattern-recognition family NACHT-LRRs (NLRs), which juxtaposes the role of TLRs in pathogen recognition. Together, NOD2 and RIP2 form a protein complex that drives production of proIL-1B. In the presence of activated caspase 1 and 5, this signal eventually leads to the maturation and release of proinflammatory cytokine IL-1B (Martinon and Tschopp, 2005), a plausible explanation of pan-viral upregulation of IL-1B in our data. In FLU, we further observed significant

changes in genes NLRP1/2/3 and ASC (HRV as well), key components of so-called *inflammasome* (Martinon and Tschopp, 2005). In agreement to our results, a new FLU study demonstrated additional NLRs-mediated viral recognition and caspase-1 inflammasome activation in hematopoietic cells (Ichinohe et al., 2009). *This, combined with our observations, support a hypothesis that NLRs are directly involved in recognizing RNA viruses FLU, HRV, and RSV in addition to TLRs.* Given the fact that these two pathways converge at NF-*k*B activation, this would *imply a hitherto underappreciated role of NLRs* in pathogenic respiratory infections. *Considering further the pan-viral upregulation of MDA-5 and RIG-I, we are essentially witnessing a full scale activation and engagement of PRRs from all known categories.* The overlapping or redundant stimulating signals to the immune system provided by these PRRs may potentially cause over-stimulated cellular innate immune response and relentless immune system activation. Their association with Sx phenotype would, at least partially, contribute to hyperactivated Sx host immune response.

Similarly, Siglec-1 (or Sialoadhesin), a macrophage-specific adhesion molecule, showed pan-viral sustained upregulation in Sx phenotype. Capable of increasing pathogen uptake in macrophages and promoting pathogen endocytosis, Siglec-1 could play a role in over-stimulating inflammatory response via increased antigen presentation. In addition, Siglec-7 was found to be FLU-specific. Located in cluster FLU7, it shows a transient sharp decline in Sx phenotype during early to middle phase of infection. With its inhibitory effect on controlling leukocyte expansion during inflammatory response, such transience decline might contribute also to the inflammation seen in Flu Sx cases. Most importantly, Siglecs are known to be recognized and exploited by HRV to limit complement activation and switch off antigen-specific immune responses (Kirchberger et al., 2005). The pan-viral activation of Siglec-1 in our data would directly follow this observation in HRV

with plausible extension to FLU and RSV. Additional studies are needed to determine whether virus-induced Siglecs expression is directly related to the production and release of inflammatory mediators.

In selecting genes to be included in the boosting classifier used to determine predictability of host response states, we imposed very strict conditions to keep as small a panel of predictors as possible to ensure model stability and generalization. A direct consequence of boosting simple linear base learners is that it breaks the correlation structure and penalizes the genes that are collinearly related, such as members from same SOM cluster. It is therefore not unexpected that the final set of predictor genes are from different biological pathways or with less known functions. Nevertheless, many of them are directly related to the subject of viral infection and immunological diseases. Examples include viral disassembly (CTSL1), hypersensitive reaction (FCER1G, GADD45A, IFI30, SOCS3, TYMS), antiviral (RSAD2, IRF9, IFIT3, OAS1, NOD2, C3AR1), cell killing and apoptosis (NCR3, TNF), immunological dysfunction (GM2A, GRM7, MICB, SLIT3), IL-10 signaling (HMOX1, BLVRA), and inflammatory responses (SELL, TRAF3, FCER1G, CD1C, CD200, TRAF3, LILRB2, LEPR, IL15RA). These predictors, along with other significant genes, may serve as biomarkers for more in-depth experimental studies and show previously unrecognized links to disease progression.

This paper represents a coherent pan-viral analysis of the temporal patterns of host response that differentiate symptomatic and asymptomatic phenotypes in an ambitious viral challenge study. Two methods of temporal expression analysis were applied: similarity clustering, where clusters are defined by similarities between temporal gene expression patterns, and discriminatory clustering, where clusters are defined by their ability to discriminate between various stages of host response between symptomatic and asymptomatic groups. The analysis resulted in identification of temporally modulated virus spe-

cific and pan-viral factors. Many of these factors are associated with well known host response pathways such as: activation of multiple pattern recognition receptors, antiviral response, inflammatory response, and cell apoptosis.

We emphasize that this paper's primary objective is not to supply comprehensive biological interpretations to the very many observations reported here; much more analysis remains to be performed. Nevertheless, our findings are presented in a coherent temporal setting wherein direct comparisons among viral pathogens can be carried out. It is our hope that these temporal sequences of common and pathogen-specific host responses can furnish new insights into the mechanism of immune and inflammatory responses. Further studies of the complex interplay of these temporal events may improve our understanding of pathogenic infections and symptomatic disease progression, which is greatly needed in order to better combat viral pandemics such as recent flu incidents.

3.4 MATERIALS AND METHODS

Three individual challenge studies were carried out for HRV, RSV, and FLU. Total of 20 healthy adult human volunteers were recruited for each study and underwent subsequent live virus inoculation. Blood specimens were collected over a course of 108 hours at an interval roughly 4 hours during the day. Gene expression, custom protein array using blood specimen were performed.

All exposures were approved by the relevant institutional review boards and conducted according to the Declaration of Helsinki. Funding for this study was provided by the US Defense Advanced Research Projects Agency (DARPA) through contract N66001-07-C-2024.

Human viral challenges

Human Rhinovirus Cohort ($n = 20$): We recruited healthy volunteers via advertisement to participate in the rhinovirus challenge study through an active screening protocol at the University of Virginia (Charlottesville, VA). The protocol was approved by the Human Investigations Committee of the University of Virginia, the Institutional Review Board of Duke University Medical Center and the SSC-SD Institutional Review Board (US Department of Defense; Washington, D.C.). Subjects who met inclusion criteria underwent informed consent and pre-screening for serotype-specific anti-rhinovirus approximately two weeks prior to study start date. On the day prior to inoculation, subjects underwent repeat rhinovirus antibody testing as well as baseline laboratory studies, including complete blood count, serum chemistries and hepatic enzymes. On day of inoculation, $10^6 TCID_{50}$ GMP rhinovirus (Johnson and Johnson) was inoculated intranasally according to previously published methods (Turner et al., 1982, Turner, 2001). Subjects were admitted to the quarantine facility for 48 hours following rhinovirus inoculation and remained in the facility for 48 hours following inoculation. Blood was sampled into PAXGene^{TM} blood collection tubes (PreAnalytix; Franklin Lakes, NJ) at pre-determined intervals post inoculation. Nasal lavage samples were obtained from each subject daily for rhinovirus titers to accurately gauge the success and timing of the rhinovirus inoculation. Following the 48^{th} hour post inoculation, subjects were released from quarantine and returned for three consecutive mornings for sample acquisition and symptom score ascertainment.

Human RSV Cohort ($n = 20$): A healthy volunteer intranasal challenge with RSV A was performed in a manner similar to the rhinovirus intranasal challenge. A healthy volunteer intranasal challenge with RSV was performed in a manner similar to the rhinovirus intranasal challenge. The protocol was approved by the East London and City Research Ethics Committee 1 (London, England), an independent institutional review board (WIRB: Western Institutional Review Board, Olympia WA), the Institutional Review Board of

Duke University Medical Center (Durham, NC), and the SSC-SD Institutional Review Board (US Department of Defense, Washington, D.C.). The RSV challenge was performed at Retroscreen Virology, Ltd (Brentwood, UK) in 20 pre-screened volunteers who provided informed consent. All subjects underwent informed consent. On day of inoculation, a dose of 10^4 $TCID_{50}$ respiratory syncytial virus (RSV; serotype A) manufactured and processed under current good manufacturing practices (cGMP) by Meridian Life Sciences, Inc. (Memphis, TN USA) was inoculated intranasally per standard methods. Blood and nasal lavage collection methods were similar to the rhinovirus cohort, but continued throughout the duration of the quarantine. Symptoms were recorded twice daily using standardized symptom scoring (Jackson Score, a combined measure of five symptoms of respiratory infection) (Jackson et al., 1958). Standardized symptom scores were recorded by trained study personnel. Due to the longer incubation period of RSV A, subjects were not released from quarantine until after the 165^{th} hour AND were negative by rapid RSV antigen detection (BinaxNow Rapid RSV Antigen; Inverness Medical Innovations, Inc).

Influenza Cohort ($n = 17$): A healthy volunteer intranasal challenge with influenza A

A/Wisconsin/67/2005 (H3N2) was performed at Retroscreen Virology, LTD (Brentwood, UK) in 17 pre-screened volunteers who provided informed consent. On day of inoculation, a dose of 10^6 $TCID_{50}$ Influenza A manufactured and processed under current good manufacturing practices (cGMP) by Bayer Life Sciences, Vienna, Austria) was inoculated intranasally per standard methods at a varying dose (1:10, 1:100, 1:1000, 1:10000) with four to five subjects receiving each dose. Due to the longer incubation period of influenza as compared to rhinovirus, subjects were not released from quarantine until after the 216^{th} hour. Blood and nasal lavage collection continued throughout the duration of the quarantine. All subjects received oral oseltamivir (Roche Pharmaceuticals) 75 mg by mouth twice daily prophylaxis at day 6 following inoculation. All patients were nega-

tive by rapid antigen detection (BinaxNow Rapid Influenza Antigen; Inverness Medical Innovations, Inc) at time of discharge.

Case Definitions: Symptoms were recorded twice daily using standardized symptom scoring (Jackson et al., 1958). The modified Jackson Score requires subjects to rank symptoms of upper respiratory infection (stuffy nose, scratchy throat, headache, cough, etc) on a scale of 0 – 3 of "no symptoms", "just noticeable", "bothersome but can still do activities" and "bothersome and cannot do daily activities". For all cohorts, modified Jackson scores were tabulated to determine if subjects became symptomatic from the respiratory viral challenge. A modified Jackson score of ≥ 6 over the quarantine period was the primary indicator of successful viral infection (Turner, 2001) and subjects with this score were denoted as "SYMPTOMATIC, INFECTED". Viral titers from daily nasopharyngeal washes were used as corroborative evidence of successful infection using quantitative culture (rhinovirus, RSV, influenza) and/or quantitative PCR (RSV and influenza) (Barrett et al., 2006, Jackson et al., 1958, Turner, 2001).

Subjects were classified as "ASYMPTOMATIC, NOT INFECTED (healthy)" if the Jackson score was less than 6 over the five days of observation and viral shedding was not documented after the first 24 hours subsequent to inoculation. Standardized symptom scores tabulated at the end of each study to determine attack rate and time of maximal symptoms (time "T").

Biological Sample Collections: For each viral challenge, subjects had the following samples taken 24 hours prior to inoculation with virus (baseline), immediately prior to inoculation (pre-challenge) and at set intervals following challenge: peripheral blood for serum, peripheral blood for PAXgene^{TM}, nasal wash for viral culture/PCR, urine, and exhaled breath condensate. For the rhinovirus challenge, peripheral blood was taken at baseline,

then at 4 hour intervals for the first 24 hours, then 6 hour intervals for the next 24 hours, then 8 hour intervals for the next 24 hours and then 24 hour intervals for the remaining 3 days of the study. For the RSV and influenza challenges, peripheral blood was taken at baseline, then at 8 hour intervals for the initial 120 hours and then 24 hours for the remaining 2 days of the study. For all challenge cohorts, nasopharyngeal washes, urine and exhaled breath condensates were taken at baseline and every 24 hours. Samples were aliquoted and frozen at -80^{o}C immediately. This study is focused on comparison of baseline samples with PAXgene^{TM} samples taken at time of peak symptoms. Paxgene^{TM} RNA from the timepoint of maximal symptoms was chosen for hybridization to Affymetrix U133a human microarrays for further analysis. For all results reported, gene expression signatures were evaluated at the time of maximal symptoms following viral inoculation for symptomatic subjects and a matched timepoint for asymptomatic subjects. Baseline (pre-inoculation) samples were also analyzed for all subjects.

Community influenza and bacterial infection cohort: Raw data from (Ramilo et al., 2007) was obtained from the public domain database GEO (GSE6269) and were analyzed independently using methods described below. RNA purification and microarray analysis: RNA was extracted at Expression Analysis (Durham, NC) from whole blood using the PAXgene^{TM} 96 Blood RNA Kit (PreAnalytiX, Valencia, CA) employing the manufacturer's recommended protocol. Complete methodology can be viewed in the Supplementary Methods. Hybridization and microarray data collection was performed at Expression Analysis (Durham, NC) using the GeneChip Human Genome U133A 2.0 Array (Affymetrix, Santa Clara, CA).

Statistical analysis

Differential expression analysis of gene and protein expression data. Raw gene expression profile was preprocessed using RMA (Bolstad et al., 2003) with batch effects removed/reduced using an empirical Bayes method (Johnson et al., 2007). Temporal gene expression was analyzed using EDGE (Storey et al., 2005). Briefly, a gene-wise natural cubic smoother was fit on the temporal expression profiles for each individual subject. To prevent overfitting, we fixed the number of spline knots such that there is three time points available for each knot. Subsequently group cubic spline was summarized for asymptomatic and symptomatic subjects, upon which across group gene expression is compared for differences. Statistical significance is assessed using F-test with simultaneous multiple testing FDR control. The final set of candidate genes are deemed as significantly differentially expressed genes with FDR adjusted p-value $< 1\%$, except for RSV we used the nominal p-value since there is less number of genes which is likely to be a results of less samples.

We used similar setting in analyzing custom antigen array data. In order to avoid overfitting from interpolation of data, we fix the degrees of freedom at two for natural cubic spline fitting. We chose q-value $<$ 20% cutoff. Considering the higher specificity of the custom antigen array, such choice of threshold is deemed to fairly reasonable.

Cluster significant genes using Self-Organizing Map (Kohonen, 1995). In principle, SOM algorithm presents complex high-dimensional relationships between data items in a low-dimensional display, while preserving their most important topological and metric relationships (Kohonen, 1995). In our analysis, we aim to place in the same region of a 2D grid layout those genes that are similar in temporal expression profiles, measured by their Euclidean distances. Prior to clustering, a natural cubic spline was fitted on the temporal expression values of each gene using smoothing spline method (Hastie and Tibshirani,

1990). We fixed the degrees of freedom at four (df=2 for RSV given only 5 time points available), yielding a more conservative model fitting in terms of the amount of smoothing. This is in concordance to the parameter setting we used in determining significance level of genes in EDGE (Storey et al., 2005). We consider such choice is critical to avoid possible overfitting of our model. The fitted values were subsequently z-score normalized. For SOM clustering, a 4×2 hexagonal grid of prototypes was used for FLU, 3×2 for HRV, and 1×3 for RSV. Since there is no golden standard in choosing the "best" map configuration among all possible maps, we proposed an analytical selection procedure in which we balance the complexity of the map (number of prototypes), the distances between genes and their prototypes, and the silhouette values of genes (Rousseeuw, 1987) (measure of the closeness of a gene to its within cluster members w.r.t its neighboring prototype members). Each prototype's representative centroid is initially chosen by random among genes. Initial neighborhood size was set to 0.5 such that 25% of prototypes began within each other's neighborhood. It decreases linearly over all iterations. Each gene was presented to the map 50 times.

Association study using Pearon correlation. In analyzing antigen array proteins and SOM centroids, we studied their temporal correlation using pearson correlation with permutation for generated random null distribution of random correlation. For each time point corresponding to an antigen array sample, gene expression values were derived from temporal splines fitted on individual SOM centroids. In HRV, median symptom T is 72hrs, time 0, 7, 14, 58, 72 hours are derived. Similarly, for RSV with median T at +142hours, corresponding gep sampling time at 0, 14, 28, 114, 142 hours are computed. For FLU with median T of +80hours, we generated gep time on the curve at 0, 8, 16, 64, 80 hours. Pearson correlation coefficients are used to measure association.

Build prediction model using model-based boosting. All samples are classified into four classes using Mixed Component Analysis (MCA). Class 1 pre-inoculation; class 2 post-inoculation asymptomatic; class 3 post-inoculation pre symptom; and class 4 post-inoculation symptomatic. Among these, class 2 and class 3 are of particular interest since they poses the biggest challenge in discriminating those who are infected yet show no symptoms from those who eventually display symptomatic. We used a state-of-art machine learning method, LogitBoosting (Bhlmann, 2006), to construct the prediction model for its substantial power and resistance to model overfitting. Adopting an one-versus-one strategy, we extended the boosting procedure with multi-class classification capability. In essence, an ensemble of weak classifiers is constructed for each pair of classes, which define the two decision regions that the ensemble is operating within. The final prediction is determined by a majority voting mechanism aggregating all pair-wise classification ensembles, with ties as misclassification. To further simplify the model making it less prone to overfitting, we carefully selected simple univariate least square as the base learner. Model fitting was carried out with a functional gradient descent algorithm (Bhlmann, 2006), minimizing the negative binomial log-likelihood function. Among the reasons in choosing negative log-likelihood loss function, we value the most its robustness and capability to provide probability estimates. Of note, there is only one univariate least square classifier enters the model at each iteration during the overall fitting procedure. Such univariate componentwise addition of predictor implies an implicit variable selection in that only the best predictor gene is chosen at each step. The Akaike information criteria (AIC) is used to determine the number of boosting iterations. Furthermore, a two stage model fitting procedure was conducted, largely for the reason of fast computation. Smaller weighs take more time for the algorithm to converge, which was an issue since our dataset is unusually large in dimension of both observations and variables. Therefore,

a bigger weight was used in first-pass fitting to remove those variables that never entered the model. This gives us a smaller dataset to operate on. In the following step, we used a smaller weight for the classifiers so that overfitting becomes even less likely.

The boosting method, by and large, is considered resistant to overfitting the data, a critical issue in classification problem. Nonetheless we aim at fitting an honest model that may not necessarily yield the best prediction performance yet it is likely to be the least overfitting one among all possible models. In order to assess the true prediction power of the overall model and its predictor genes, bootstrap resampling technique (Efron, 1979) is used to generate random copies of data upon which boosting classifiers were fit. Each bootstrap copy of data was then dividend into training (70%) and test set (30%) with balanced number of observations from each class. The training set was used to construct the boosting ensemble while the test set was used for prediction. We then investigated the area under the curve (AUC) using receiver operating curve (ROC) plot based on the predicted probability estimates of samples in test set. In order to further reveal the estimating uncertainties of the model, we compute the 95% confidence interval of prediction based on bootstrapped data for both true positive prediction and false negative prediction at each threshold point. Total of 1,000 bootstrap copies of data was generated.

Supplementary tables In Rhinovirus infection, the peak symptom is observed at $\sim$72hrs after infection. This is consistent with the report where the onset of common cold symptoms typically occurs 1 – 2 days after viral infection and the time to peak symptoms is generally 2 – 4 days (Tyrrell et al., 1993).

Protein data from NPW, EBC, and Urine. In addition, we examined protein expression levels in nasopharyngeal washing (NPW), urine specimen, and exhaled breath

condensate (EBC) on the same antigen array platform. For direct comparison, we once again chose the baseline and peak symptom time points. Overall, these proteins expression levels showed clear sign of immune response against viral pathogens regardless of their clinical outcome that is consistent with their corresponding mRNA expression. The heatmaps in Figure 7A show the ratio of protein expression levels at peak versus baseline time. The proteins listed are the ones with significant changes at peak symptom time using pair *t*-test statistic. All Sx and Asx samples are pooled together. More specific, NPW showed stronger upregulation signature in RSV whereas HRV dominates the upregulated protein list in EBC. It is in Urine that we observe some inconsistent expression pattern. Different from RSV and HRV, most FLU-specific proteins are downregulated at peak time showed. In terms of phenotypic differences, we did not observe exceptionally better discriminating power from these platforms except in RSV. NPW from RSV appears to be the only one that bears discriminating capacity. Two proteins, alpha-1-antitrypsin (SERPINA1) and myoglobin (MB), can separate 7 of 9 Sx samples from Asx phenotype (Figure S3.1C). Myoglobin in circulating blood clearly indicates significant tissue damage at peak symptom time, which correlates with a protective role of SERPINA1 in limiting excessive inflammation. Sex hormone-binding globulin (SHBG) also showed relative discriminatory power between phenotypes (data not shown). None of these showed significant predictive power in RSV microarray data.

Interestingly, EBC showed a marginally significant (p-value= 0.03) lower pH measures in Sx phenotype when data from three viral challenge studies were analyzed together (Figure S3.1D). Although the increased detection power is likely due to more available data points, there was report confirming the inhibitory effect of low pH values on the replication of respiratory viruses such as HRV and FLU (Gern et al., 2007).

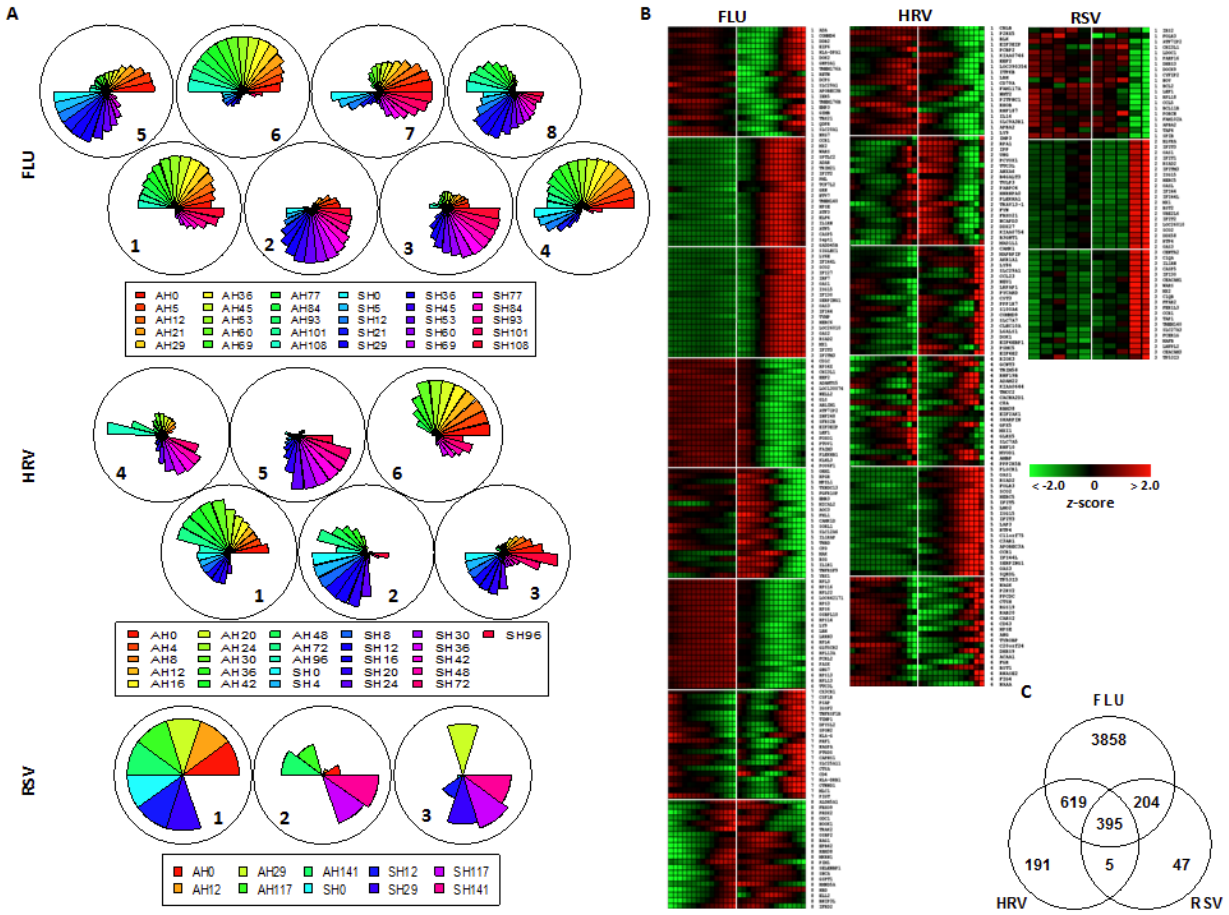


Figure 3.1: **Clustering of temporal significant genes comparing Sx versus Asx in HRV, RSV, and FLU challenge studies. A)** Polar plots of SOM clusters and their associated gene expression patterns. Each polar plot represents the codebook or prototype of a cluster. Individual time points are scaled and ordered in sequence and phenotype around the circle. More specifically, the temporal expression of Asx resides on the top portion of the circle while Sx expression occupies the bottom half. Each phenotypes expression values are placed in time sequence, counterclockwise, inside its own half circle. The degrees of angle are equally divided among segments within the circular plot. The different lengths of radii (therefore the area) of the segments represent the magnitude of gene expression of individual time points, relative to the average expression level of the complete temporal course. **B)** Heatmap of top 20 genes from each cluster. Genes are ordered within clusters according to their corresponding significance level. **C)** Intersection of significant genes from three different viral challenges.

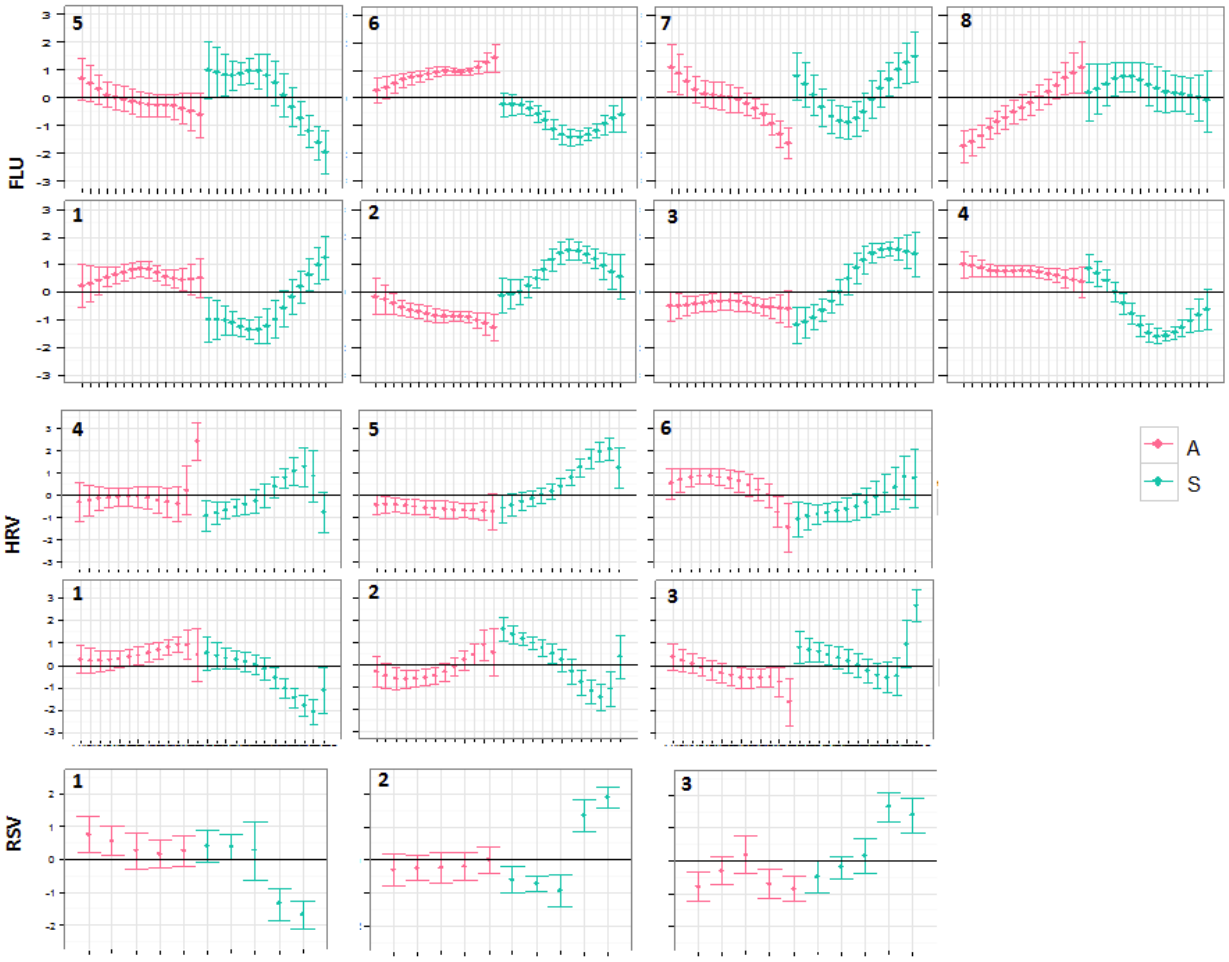
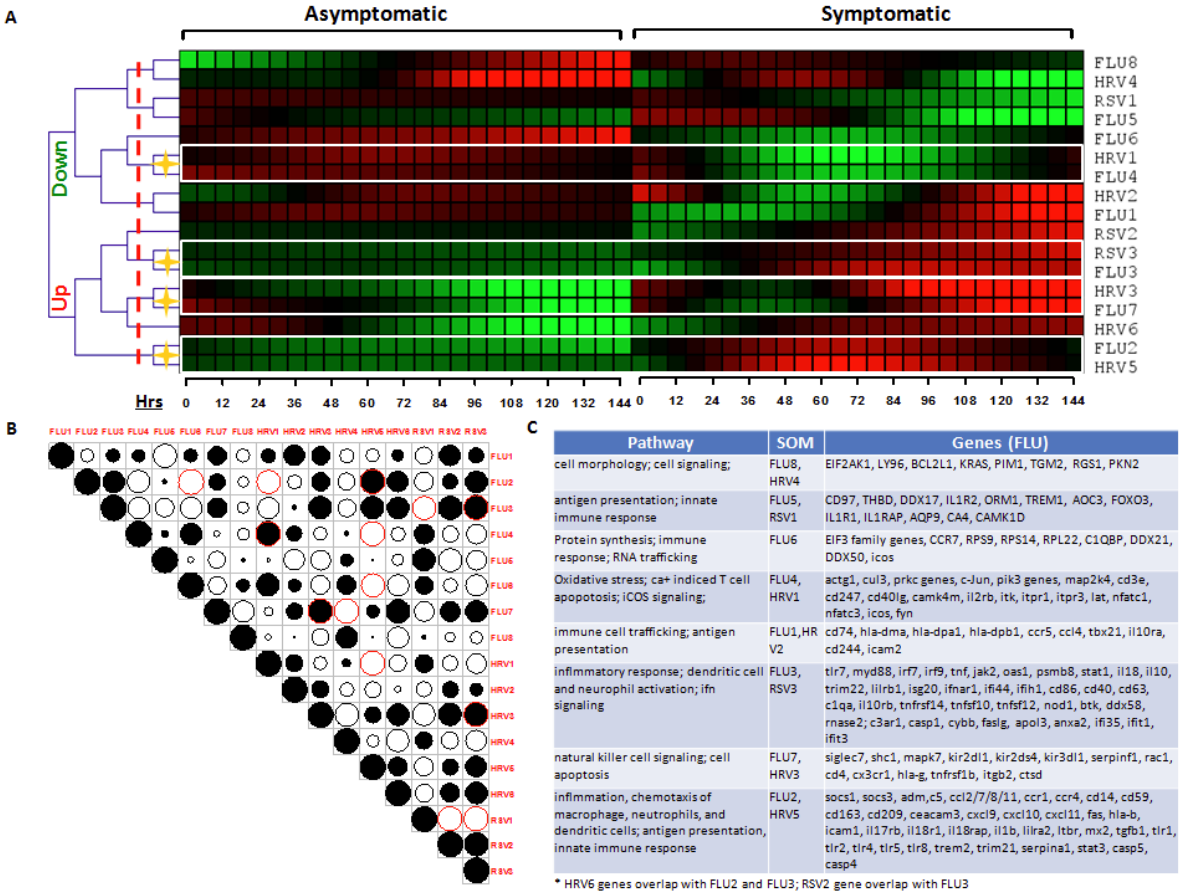


Figure 3.2: **Centroids of each SOM cluster.** Individual cluster average expression and corresponding +/- two standard deviations in FLU, HRV, and RSV. In addition, we also computed 95% confidence interval using nonparametric bootstrap method without assuming normality (Wald et al., 2003). The derived confidence limits of the mean expression of each cluster is much smaller than (and therefore completely covered by) +/- two standard deviations shown here (data not shown). This assures us the within cluster variation is well captured by the standard deviation as the summary statistics.



Pathway	SOM	Genes (FLU)
cell morphology; cell signaling;	FLU8, HRV4	EIF2AK1, LY96, BCL2L1, KRAS, PIM1, TGM2, RGS1, PKN2
antigen presentation; innate immune response	FLU5, RSV1	CD97, THBD, DDX17, IL1R2, ORM1, TREM1, AOC3, FOXO3, IL1R1, IL1RAP, AQP9, CA4, CAMK1D
Protein synthesis; immune response; RNA trafficking	FLU6	EIF3 family genes, CCR7, RPS9, RPS14, RPL22, C1QBP, DDX21, DDX50, icos
Oxidative stress; ca+ indiced T cell apopotosis; iCOS signaling;	FLU4, HRV1	actg1, cul3, prkc genes, c-Jun, pik3 genes, map2k4, cd3e, cd247, cd40lg, camk4m, il2rb, itk, itpr1, itpr3, lat, nfatc1, nfatc3, icos, fyn
immune cell trafficking; antigen presentation	FLU1,HRV2	cd74, hla-dma, hla-dpa1, hla-dpb1, ccr5, ccl4, tbx21, il10ra, cd244, icam2
inflmmatory response; dendritic cell and neurophil activation; ifn signaling	FLU3, RSV3	tlr7, myd88, irf7, irf9, tnf, jak2, oas1, psmb8, stat1, il18, il10, trim22, lilrb1, isg20, ifnar1, ifi44, ifih1, cd86, cd40, cd63, c1qa, il10rb, tnfrsf14, tnfsf10, tnfsf12, nod1, btk, ddx58, rnase2; c3ar1, casp1, cybb, faslg, apol3, anxa2, ifi35, ifit1, ifit3
natural killer cell signaling; cell apoptosis	FLU7, HRV3	siglec7, shc1, mapk7, kir2dl1, kir2ds4, kir3dl1, serpinf1, rac1, cd4, cx3cr1, hla-g, tnfrsf1b, itgb2, ctsd
inflmmation, chemotaxis of macrophage, neutrophils, and dendritic cells; antigen presentation, innate immune response	FLU2, HRV5	socs1, socs3, adm,c5, ccl2/7/8/11, ccr1, ccr4, cd14, cd59, cd163, cd209, ceacam3, cxcl9, cxcl10, cxcl11, fas, hla-b, icam1, il17rb, il18r1, il18rap, il1b, lilra2, ltbr, mx2, tgfb1, tlr1, tlr2, tlr4, tlr5, tlr8, trem2, trim21, serpina1, stat3, casp5, casp4

* HRV6 genes overlap with FLU2 and FLU3; RSV2 gene overlap with FLU3

Figure 3.3: **Common and unique temporal expression patterns across HRV/RSV/FLU challenge studies. A)** Unsupervised hierarchical clustering of SOM centroids. Each SOM centroid (prototype) is placed in a row and the sampling time points are indexed by the columns. On the left are the Asx prototypes and on the right are the corresponding Sx prototypes. Since the sampling time points are not identical across the three viral challenge studies, the expression data was interpolated to a common uniformly spaced time grid prior to hierarchical clustering of the prototypes. A total of 24 time points at 6-hour interval were fitted using a cubic spline. **B)** Plot of pairwise correlations among the prototypes. The area of the circles corresponds to the magnitude of correlation with solid circles encoding positive correlation and open circles negative correlations. Circles with a red border represent correlation coefficients greater than 0.9 or less than -0.9. **C)** A table of common and distinct pathways described by different clusters of significant FLU genes shown as examples.

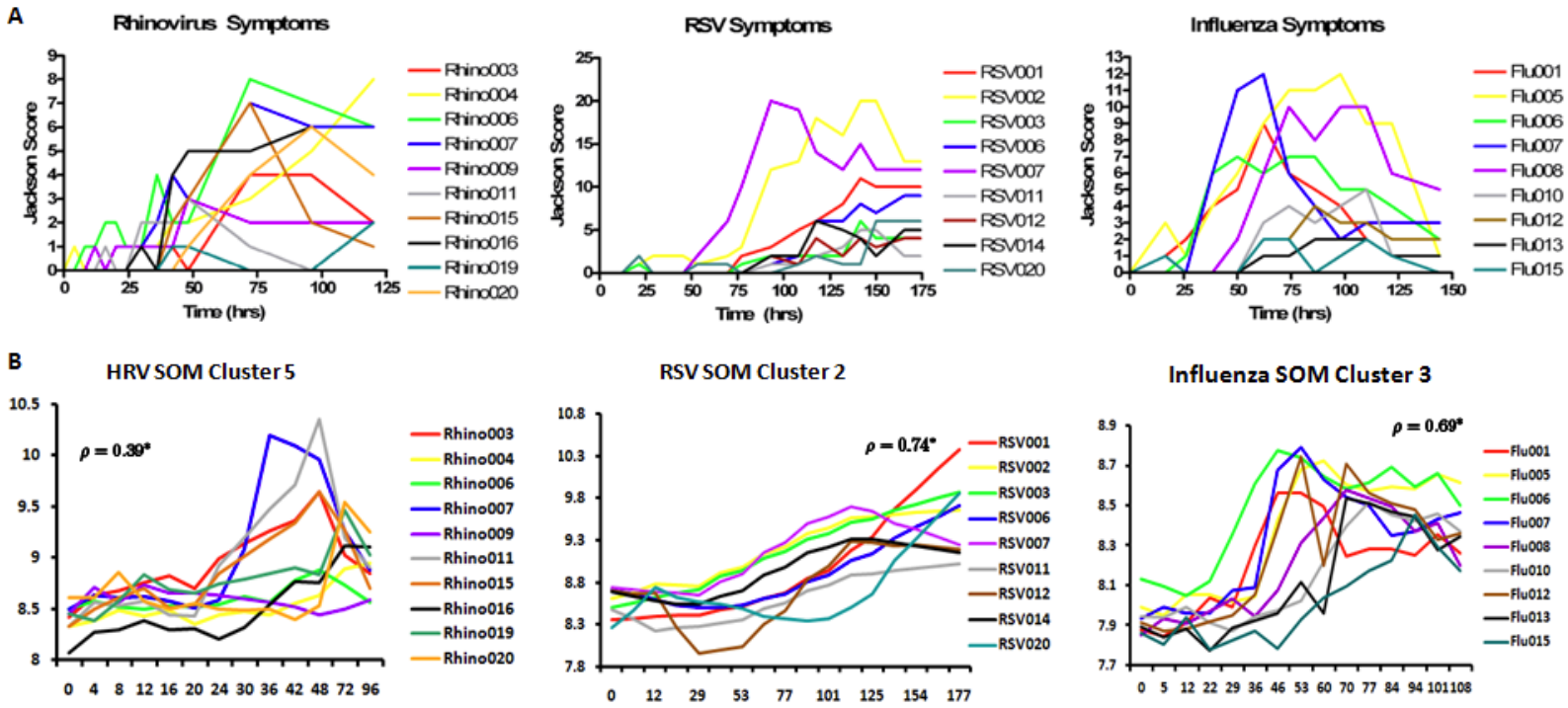


Figure 3.4: **Significant correlation between SOM prototypes and clinical symptom scores.** Individual subjects are represented by lines in different colors. The expression values of different genes within SOM cluster is averaged at each time point by individual subjects. **A)** Clinical symptom diary for HRV, RSV, and FLU. **B)** Expression kinetics of most positively correlated SOM clusters. *p-value $< .0001$. We define p-value as the probability of obtaining a correlation coefficient in random samples as large as observed. The random correlation coefficients are computed by 1,000 permutations of the gene expression values. The p-value is then. Footnote that we have more time points in symptom diary than GEP in FLU. Furthermore, HRV time points are not exactly the same between the two modalities (GEP and symptom scores). We chose the closest ones (instead of smoother fitting of data).

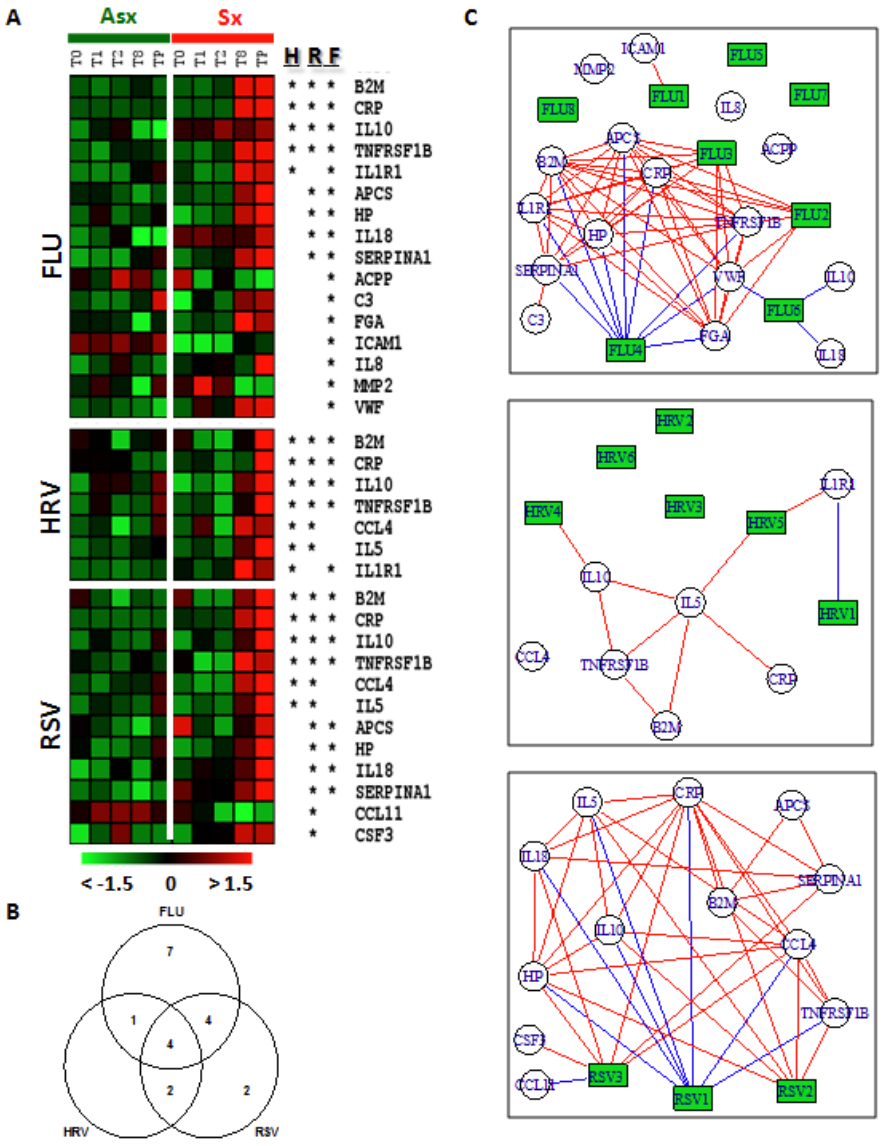


Figure 3.5: **Temporal expression of plasma proteins. A)** Heatmap of temporal expression of significant proteins identified in HRV, RSV, and FLU. Protein names are shown on the right side of the heatmap with * representing significance (q-value < 0.1) in H(RV), R(SV), and F(LU). **B)** Venn diagram showing the overlap among differentially expression proteins in three viral challenges. **C)** Graphical model representation of temporal relevance between proteins and SOM centroids (supplementary methods). Lines denote strong association with Pearson correlation coefficient >0.8 in magnitude. Positive correlations are represented by red lines while negative correlations are in blue color. SOM centroids are depicted by green rectangular boxes while significant proteins are denoted by empty circles.

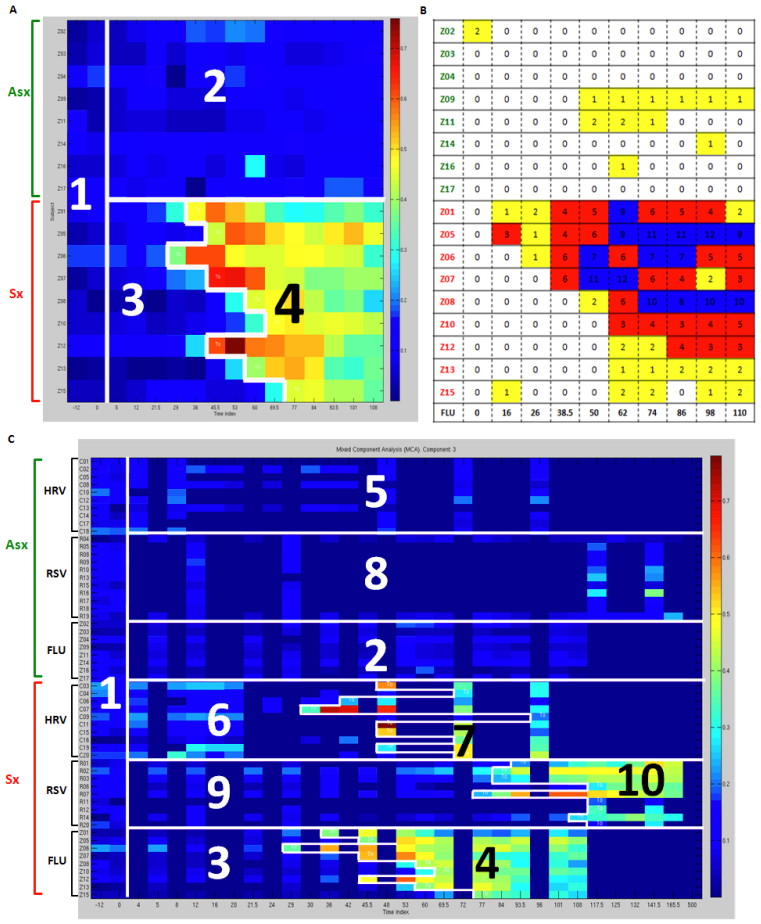
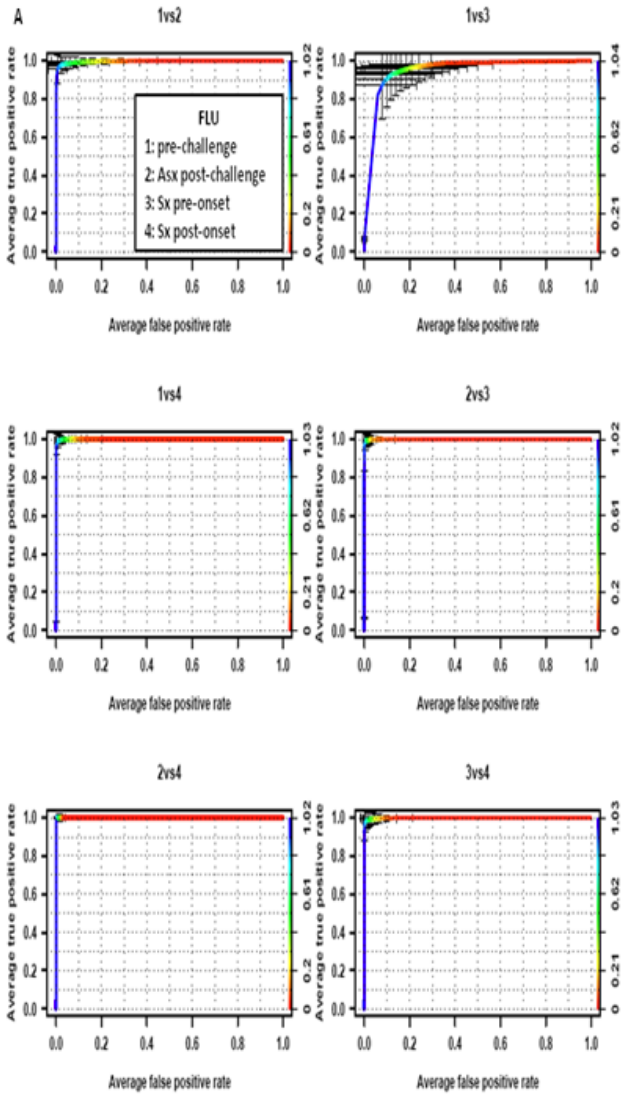


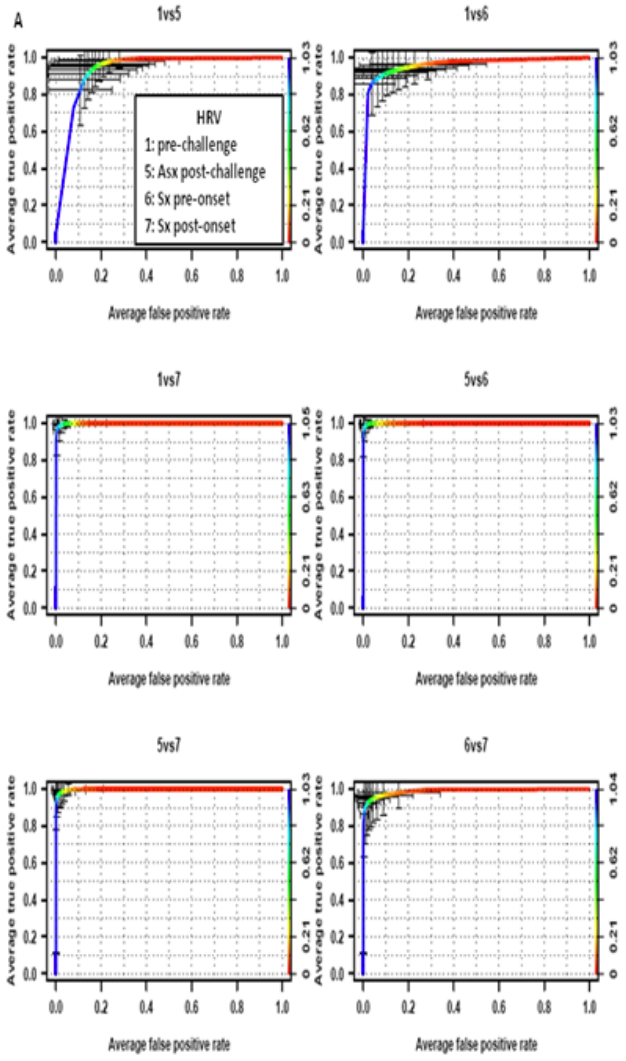
Figure 3.6: **MCA detection of critical transition point of transcriptom profiles. A)** Influenza (FLU) factor loading-coded color representation of all individual's expression profiles. Each microarrary is represented by one square cell of the image and ordered by phenotype and subject (row-wise) then increasing time (column-wise). Color palette is coded according to the factor loading in each single gene expression profile determine by MCA. The higher the loading, the warmer (red) color representation of the sample. The numbers (1 to 4) are the class designation determined by MCA (3 and 4) and inoculation time (1 and 2). The boundary between 3 and 4 occurs at samples that are labeled (**To**) denoting the critical transition point (onset time) of Sx subject transcriptome profiles. The darkest blue color (absolute 0 loading) corresponds to samples that were not assayed. **B)** Clinical symptom chart of corresponding subject and time ordered in the same manner as Figure 6A. **C)** MCA factor loading matrix for all three viruses with 10 associated classes.



B

1vs2 (n=13)	1vs3 (n=9)
FKBP5, ST14, POLD2, PLAC8, S100PBP, ACOT8, GM2A, TPST1, LEPR, APOLD1, MT1F, PITPNC1, RNF13	KIF1B, ST14, IRS2, FKBP5, APBA2, LYK5, TOMM34, INPP4B, TMEM149
1vs4 (n=9)	**2vs3 (n=20)**
IFI44L, IRS2, SMAD1, GBP1, PSME2, BLVRA, GRAMD1C, GNG7, NLRP1	LPAR1, C3AR1, RAB8A, ZNF493, GM2A, TPM2, GNG7, JUP, ENOSF1, RETN, HHLA3, CTH, BAIAP3, SLC12A9, APOLD1, LILRB2, BTN3A2, TOMM34, ZMAT3, LAPTM4B
2vs4 (n=7)	**3vs4 (n=6)**
RTP4, GNG7, OSBPL10, OAS1, CD1C, IFI27, CTSL1	IFI44L, SMAD1, RTP4, IRS2, GRAMD1C, DDX17

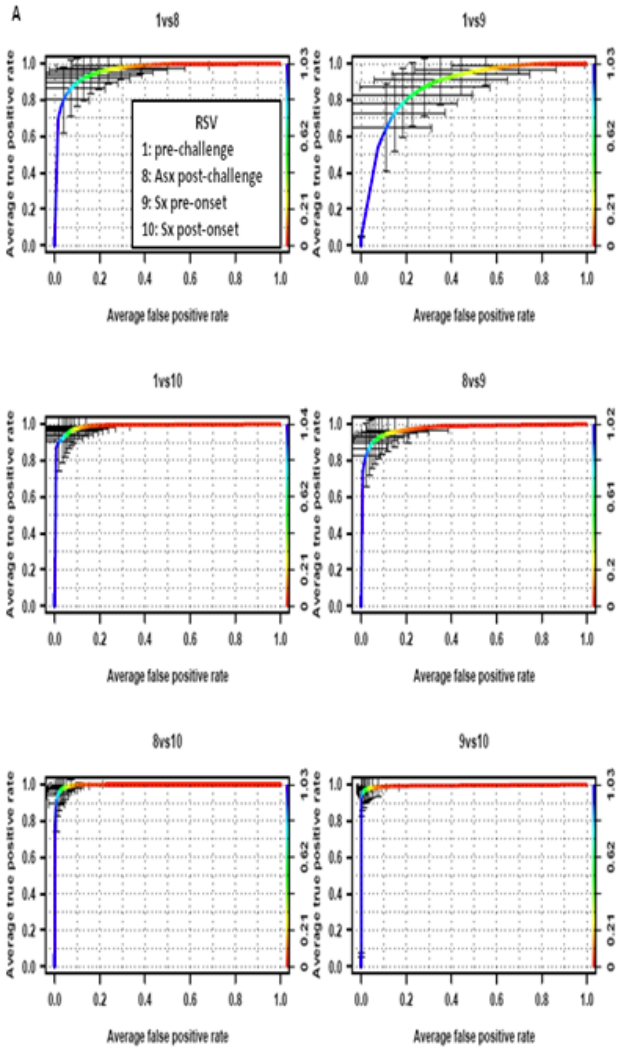
Figure 3.7: **Influenza: Performance of boosting classifier consisting of 51 distinct genes. A)** Each individual plot shows the performance of a pairwise boosted classifier. The ROC curve represents the average performance on the test set (hold-out set) in 1,000 bootstrapped copies of data. The error bar represents 95% confidence interval coverage estimation ($\pm$ two standard deviation under normality assumption) for each thresholding point for both detection and false alarm rate. **B)** Predictor genes selected during the bootstrapped boosting phase of each classifier with genes listes according to their relative importance.



B

1vs5 (n=15)	1vs6 (n=13)
MFSD7, FOLR3, EDF1, NCR3, DOCK4, TTC26, TXNDC14, MICAL2, ARPC5, HYAL2, CCRL2, THAP10, SLIT3, CR2, RFPL3	MICB, NUP214, LOC26010, NSUN5, C14orf124, AFM, SORT1, CALM1, COX4NB, ETFB, C19orf56, SGK1, DHRS2
1vs7 (n=7)	**5vs6 (n=14)**
RSAD2, BLK, SORT1, PTPRO, CD36, P2RX5, IRF9	LOC26010, APOBEC3A, CHI3L1, EMP3, ECHDC2, OAS1, AP3S2, ENTPD1, MARCKSL1, SGK1, PKIG, ANXA6, BACH2, SP100
5vs7 (n=9)	**6vs7 (n=11)**
RSAD2, GSTK1, BLK, APOBEC3A, RHOB, PNRC1, UPP1, IL15RA, UTS2	RHOB, RSAD2, HMGCL, GSTK1, GRM7, EHF, IRF9, LRRFIP2, DDAH2, CACNA1G, GPX5

Figure 3.8: **Rhinovirus: Performance of boosting classifier consisting of 58 distinct genes. A)** Each individual plot shows the performance of a pairwise boosted classifier. The ROC curve represents the average performance on the test set (hold-out set) in 1,000 bootstrapped copies of data. The error bar represents 95% confidence interval coverage estimation ($\pm$ two standard deviation under normality assumption) for each thresholding point of both detection and false alarm rate. **B)** Predictor genes selected during the bootstrapped boosting phase of each classifier with genes listed according to their relative importance.



B

1vs8 (n=23)	1vs9 (n=12)
SLC25A20, CD177, GCM2, CTRL, KCNMB1, TRAF1, ORMDL2, FFAR2, FCER1G, BMX, DHRS7B, TIMM10, PDK4, RNASE3, CHRNB1, MICB, SP100, SNAP29, BCL9, IFI27, LRRFIP2, EXT1, GADD45B	PCOLCE2, SLC25A20, GCM2, TNF, SOCS3, DUSP3, PARP11, CTRL, POMP, CARS2, TAS2R10, PCNT
1vs10 (n=10)	**8vs9 (n=21)**
GCM2, PCOLCE2, GADD45A, SOCS3, RAB20, MX1, ETV7, FER1L3, IFIT3, TCTN3	CD177, MICB, DSC2, KANK1, C17orf60, NOV, CD72, TRAF1, NOD2, SNAP29, STYXL1, FCER1G, PARP11, KIAA0460, SUSD4, TIMM10, TAS2R10, SELL, CDKN1C, TYMS, TRAF3
8vs10 (n=16)	**9vs10 (n=5)**
TMEM149, CD200, SERPING1, CUL1, SFT2D2, CD177, IGHM, FOLR3, VCL, C14orf94, HMOX1, TMEM62, HLA-DOB, YKT6, NOD2	IFI30, SFT2D2, RSAD2, CTSL1, DNASE1L1

Figure 3.9: **RSV: Performance of boosting classifier consisting of 66 distinct genes. A)** Each individual plot shows the performance of a pairwise boosted classifier. The ROC curve represents the average performance on the test set (hold-out set) in 1,000 bootstrapped copies of data. The error bar represents 95% confidence interval coverage estimation ($\pm$ two standard deviation under normality assumption) for each thresholding point of both detection and false alarm rate. **B)** Predictor genes selected during the bootstrapped boosting phase of each classifier with genes listed according to their relative importance.

A

FLU (TPR, FPR)%	Asx post-challenge	Sx pre-onset	Sx post-onset
Pre-challenge	*(99±3, 10±15)%*	*(91±15, 10±17)%*	*(100±0, 10±4)%*
Asx post-challenge		*(100±0, 10±0)%*	*(100±0, 10±0)%*
Sx pre-onset			*(100±2, 10±12)%*

HRV (TPR, FPR)%	Asx post-challenge	Sx pre-symptom	Sx post-symptom
Pre-challenge	*(80±19, 10±14)%*	*(91±14, 10±16)%*	*(100±0, 10±7)%*
Asx post-challenge		*(100±0, 10±7)%*	*(100±3, 10±8)%*
Sx pre-symptom			*(97±11, 10±13)%*

RSV (TPR, FPR)%	Asx post-challenge	Sx pre-onset	Sx post-onset
Pre-challenge	*(89±15, 10±18)%*	*(65±15, 10±20)%*	*(98±8, 10±18)%*
Asx post-challenge		*(95±9, 10±13)%*	*(100±1, 10±9)%*
Sx pre-onset			*(100±0, 10±8)%*

B

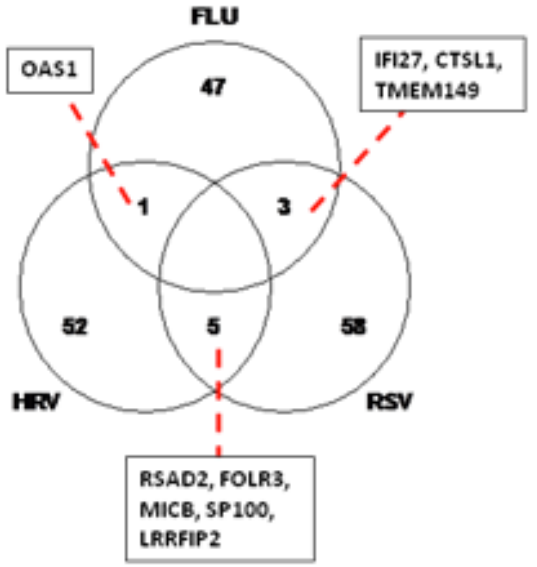


Figure 3.10: **Prediction accuracy table of boosting classifier. A)** The performance is reported at false positive rate level of 10% across all viral challenges. Although it might be conservative in some cases (e.g., FLU Asx post-challenge versus Sx cases), fixing the false positive rate is useful for cross viral comparisons. **B)** Intersection of most discriminating predictor genes for three viral challenges.

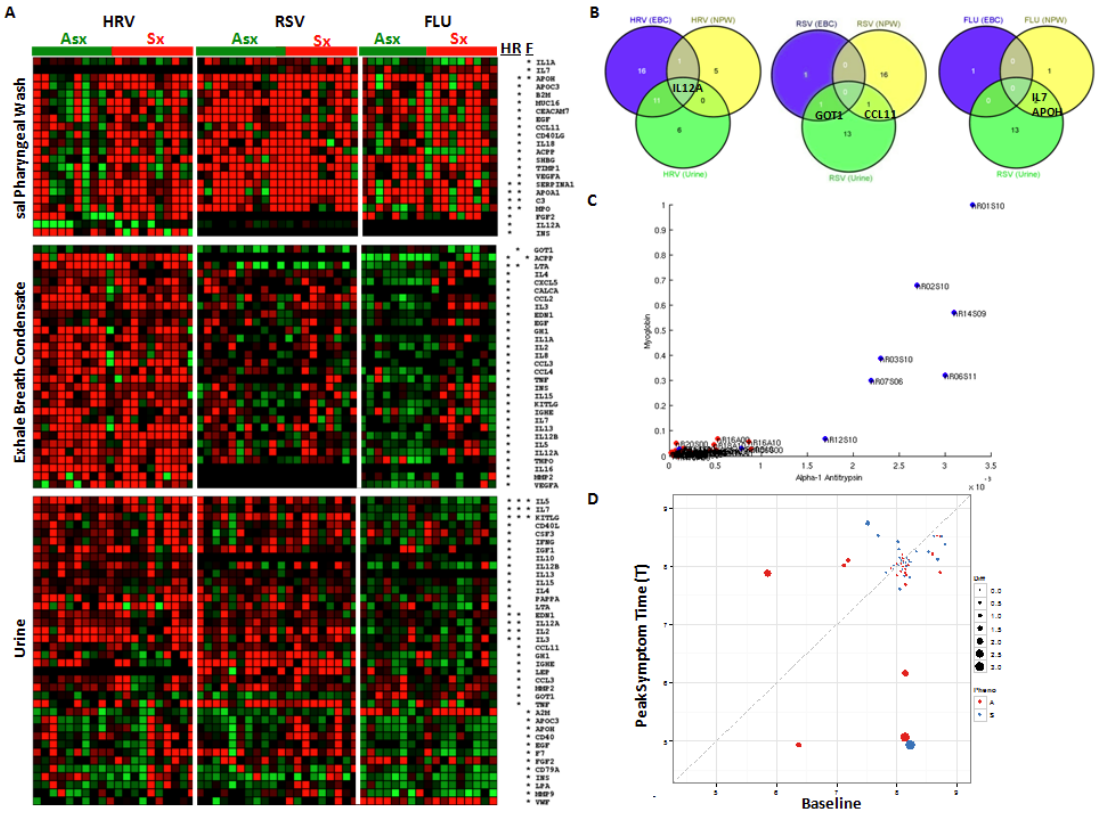


Figure S3.1: **Protein expression and pH values in NPW, EBC, and Urine post inoculation. A)**Protein that show significant alteration in expression levels post inoculation. Note that no phenotypic contrast is shown in figure A except samples are grouped according to phenotype. **B)**Intersection of proteins with differential expression before and post inoculation. Again, no differential expression analysis contrasting Asx vs Sx was shown here in figure B. **C)**Myoblobin (MB) and alpha-1-antitrypsin (SERPINA1) are shown to distinguish Asx from Sx phenotypes at peak symptom time in RSV nasopharyngeal washing. Points in color blue depict Sx phenotype. **D)**Lower pH values in exhaled breath condensate (EBC) at peak symptom time. Total of 57 subjects in all three challenge studies are shown. Size of circle denotes the magnitude of changes with blue color representing symptomatic phenotype. Diagonal dashed line denotes no change. Comparing to pH values measured at prechallenge time, points located above dashed lines are observations with increased pH values at peak time. Similarly, points below dashed lines are subjects show decreased pH values at peak symptom time.

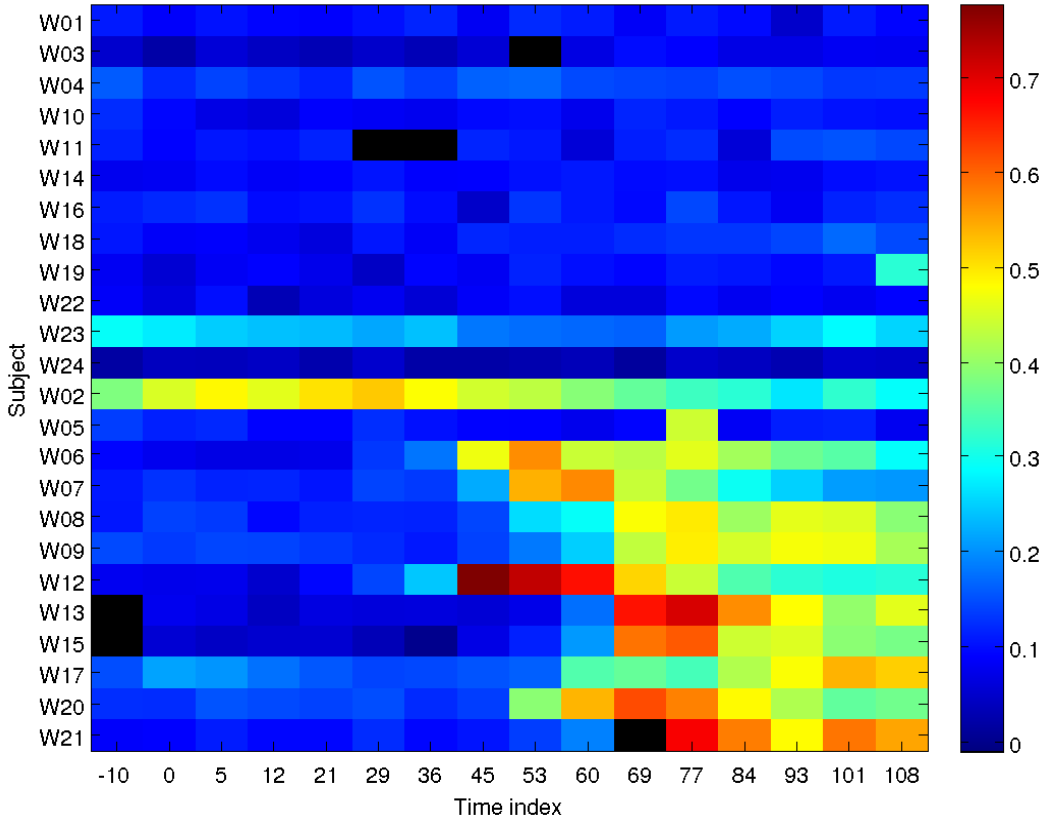


Figure S3.2: Detecting H1N1-mediated host molecular disease signature with unsupervised Bayesian linear unmixing factor analysis

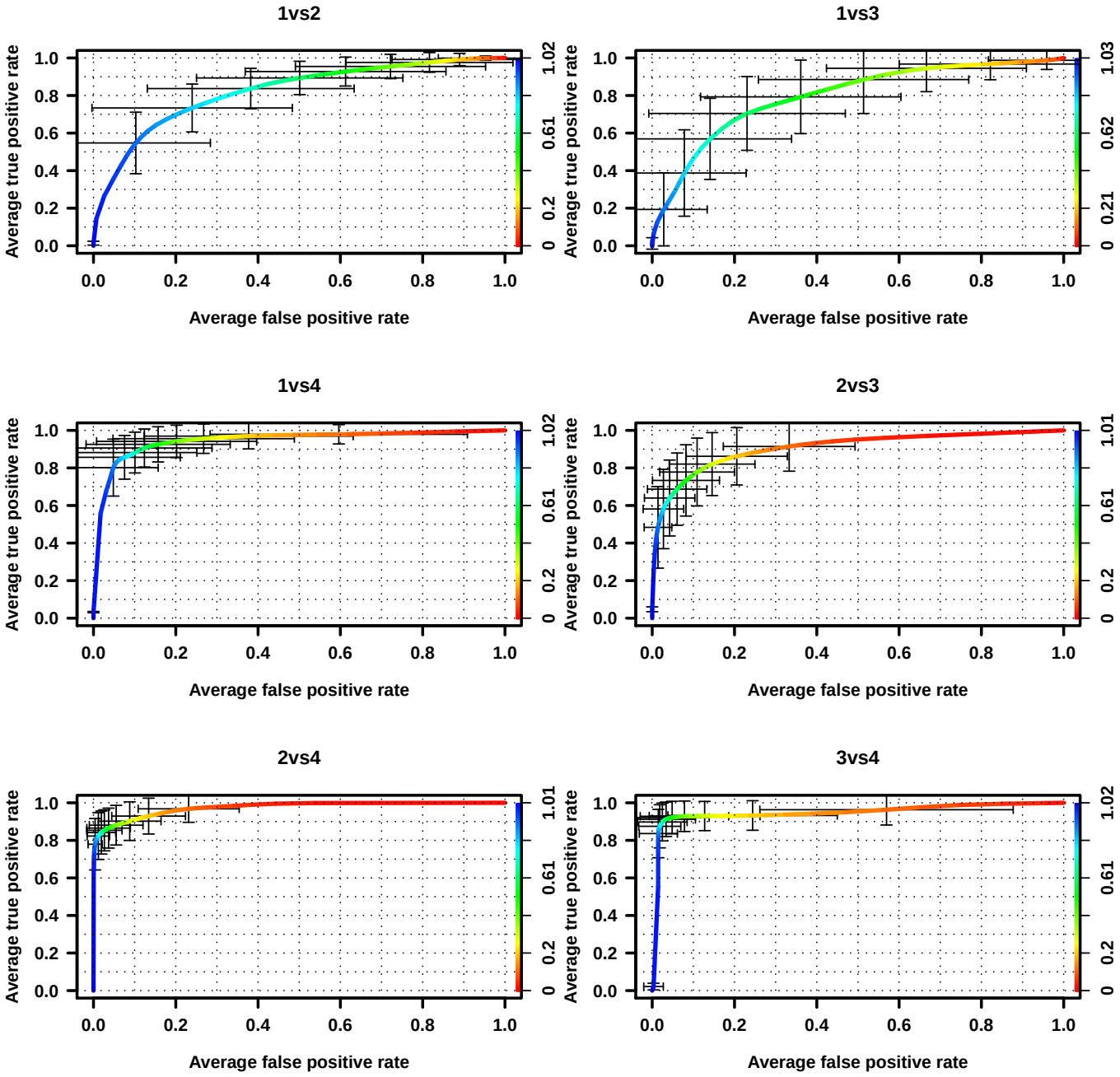


Figure S3.3: Risk stratification in Influenza H1N1 viral infections using H3N2 discriminatory genes (n=52)

Cohort	Number Challenged	Number Symptomatic	Median Time "T": Time to Peak Symptoms (hrs)	Corresponding Time Used for Asymptomatic Subjects (hrs)
Influenza	17	9	80	86
Rhinovirus	20	10	72	72
RSV	20	8	141.5	141.5

Table S3.1: Experimental cohorts for three viral challenge studies

Gene	FLU.classSOM	HRV.classSOM	RSV.classSOM
ASRGL1	1	3	2
CTSH	1	6	2
FOLR3	1	5	1
GSTK1	1	5	2
HLA-DMA	1	3	2
HLA-DPA1	1	3	2
LY86	1	3	2
ORMDL2	1	6	2
PEA15	1	5	2
PSMB3	1	5	2
RETN	1	6	2
SLC29A1	1	3	3
ADAR	2	5	3
ADM	2	5	3
AFF1	2	5	3
ANXA3	2	5	2
AOAH	2	3	3
APOL1	2	5	3
APOL2	2	5	3
ASGR2	2	5	3
ATF3	2	5	2
BMX	2	6	3
C14orf94	2	5	3
C16orf7	2	5	2
CASP5	2	5	3
CCL2	2	5	3
CCR1	2	5	3
CD177	2	5	3
CD59	2	5	2
CEACAM1	2	5	3
CEACAM3	2	6	3
CXCL10	2	5	2
DDAH2	2	5	2
DISC1	2	5	3
DOCK4	2	5	2
DOK1	2	3	3
ETV7	2	5	3
EXT1	2	5	3
FFAR2	2	5	3
GADD45B	2	5	2
GK	2	5	2
GLA	2	5	3
HPSE	2	6	3
IFIT2	2	5	2

Gene	FLU.classSOM	HRV.classSOM	RSV.classSOM
IGSF6	2	6	2
IL1B	2	5	3
IL1RN	2	5	3
IRF1	2	5	3
IRF2	2	5	2
JUNB	2	5	3
KCNJ15	2	5	3
KCNJ2	2	5	2
KIAA0040	2	5	3
LHFPL2	2	5	3
LILRA1	2	6	3
LILRA5	2	5	2
LIMK2	2	5	3
LRRFIP2	2	6	3
LTBR	2	5	3
MAP2K6	2	5	2
MX2	2	5	3
NAPA	2	5	2
NFIL3	2	5	3
NOD2	2	5	2
NRN1	2	5	3
PLAUR	2	5	3
PLEK	2	5	3
PML	2	5	3
RAB20	2	6	3
RHBDF2	2	5	2
SAMHD1	2	5	2
SECTM1	2	5	3
SERPINB1	2	5	2
SH3GLB1	2	5	2
SLC22A4	2	5	3
SLC31A2	2	5	2
SOCS1	2	5	2
SOCS3	2	5	3
SORT1	2	5	3
SPTLC2	2	5	2
STAT2	2	5	3
STK3	2	5	3
STX11	2	5	3
STX3	2	6	3
TAP2	2	5	3
TIMP2	2	5	3
TLR1	2	6	3
TLR5	2	5	3

Gene	FLU.classSOM	HRV.classSOM	RSV.classSOM
TMEM110	2	5	3
TMEM140	2	5	3
TRIM21	2	5	2
TTC26	2	5	3
WARS	2	5	3
YIPF1	2	5	3
ACOT9	3	5	2
AIM2	3	5	2
AKR1A1	3	3	2
ANXA2	3	3	2
ANXA5	3	5	3
APH1B	3	3	3
APOBEC3A	3	5	3
APOL6	3	5	2
ARRB1	3	3	3
ARSB	3	5	3
ASGR1	3	5	2
ASH2L	3	5	2
ATOX1	3	5	2
BLOC1S1	3	6	2
BLVRA	3	5	2
BST1	3	6	2
BST2	3	5	2
BTK	3	6	2
C11orf75	3	5	2
C17orf60	3	5	2
C1GALT1C1	3	5	2
C1QA	3	5	3
C1QB	3	5	3
C2	3	5	3
C20orf24	3	6	2
C3AR1	3	5	3
CALML4	3	3	2
CAMK1	3	3	2
CAPG	3	5	2
CARS2	3	6	3
CASP1	3	5	2
CBARA1	3	5	3
CBR1	3	5	2
CCRL2	3	5	2
CD36	3	5	2
CD63	3	6	3
CD86	3	5	2
CDK5	3	3	2

Gene	FLU.classSOM	HRV.classSOM	RSV.classSOM
CEBPA	3	3	3
CENTA2	3	5	3
CLEC4A	3	6	2
CNP	3	5	2
CST3	3	3	3
CTSL1	3	5	2
CUL1	3	5	2
CYBB	3	5	2
CYFIP1	3	5	3
DDX58	3	5	2
DDX60	3	5	2
DHRS7B	3	6	2
DHRS9	3	6	2
DHX58	3	5	2
DRAM	3	5	3
DRAP1	3	5	2
DUSP3	3	5	2
DYNLT1	3	5	2
EDEM2	3	5	3
EIF2AK2	3	5	2
EIF4E2	3	3	2
EMR1	3	6	2
EPB41L3	3	5	3
FAM46A	3	5	2
FANCA	3	5	2
FCER1G	3	6	3
FCN1	3	5	3
FER1L3	3	5	3
FGF13	3	5	3
FIG4	3	6	3
FKBP15	3	5	3
FLJ11286	3	5	2
FLJ78302	3	5	2
FLVCR2	3	5	2
FUT4	3	5	2
GALE	3	3	3
GALNT3	3	5	2
GBP1	3	5	2
GBP2	3	5	2
GCH1	3	5	2
GORASP1	3	5	2
GSTO1	3	5	2
GTPBP2	3	5	2
GYG1	3	5	2

Gene	FLU.classSOM	HRV.classSOM	RSV.classSOM
HERC5	3	5	2
HERC6	3	5	2
HHEX	3	5	2
HK3	3	5	3
HMGCL	3	3	2
IFI16	3	5	2
IFI27	3	5	2
IFI30	3	5	3
IFI35	3	5	2
IFI44	3	5	2
IFI44L	3	5	2
IFI6	3	5	2
IFIH1	3	5	2
IFIT1	3	5	2
IFIT3	3	5	2
IFIT5	3	5	2
IFITM1	3	5	2
IFITM3	3	5	2
IL15	3	5	2
IL15RA	3	5	3
INDO	3	6	2
IRF5	3	6	3
IRF7	3	5	2
IRF9	3	5	2
ISG15	3	5	2
ISG20	3	5	2
JAK2	3	5	2
KCNMB1	3	5	3
KIAA0082	3	5	2
KIAA0319L	3	6	3
KLF4	3	3	3
KPNB1	3	5	2
KPTN	3	5	2
KYNU	3	5	2
LAMP3	3	5	2
LAP3	3	5	2
LGALS1	3	3	2
LGALS2	3	5	2
LGALS9	3	5	2
LILRB4	3	5	3
LMNB1	3	5	3
LMO2	3	5	2
LOC26010	3	5	2
LOC391020	3	5	2

Gene	FLU.classSOM	HRV.classSOM	RSV.classSOM
LY6E	3	5	2
LYRM1	3	5	2
MAD2L1BP	3	5	2
MAFB	3	5	3
MAPBPIP	3	3	2
MAPKAPK3	3	5	3
MARCO	3	5	2
MICB	3	5	2
MS4A4A	3	5	2
MS4A6A	3	6	2
MSRB2	3	6	3
MT2A	3	5	2
MX1	3	5	2
MYD88	3	5	3
NAGA	3	5	3
NAGK	3	6	2
NEU1	3	3	3
NFKBIE	3	5	2
NIT1	3	3	3
NMI	3	5	2
NPC2	3	5	2
OAS1	3	5	2
OAS2	3	5	2
OAS3	3	5	2
OASL	3	5	2
P2RY14	3	5	2
P2RY2	3	6	3
PANK2	3	5	2
PARP12	3	5	2
PDXK	3	3	3
PHF11	3	5	2
PLA2G4A	3	5	2
PLAC8	3	5	2
PLSCR1	3	5	2
PMVK	3	3	2
PSMB10	3	5	2
PSMB8	3	5	2
PSMB9	3	5	2
PSME1	3	5	2
PSME2	3	5	2
PSTPIP2	3	5	2
RAB32	3	6	3
RGL1	3	5	2
RIN2	3	5	3

Gene	FLU.classSOM	HRV.classSOM	RSV.classSOM
RIPK2	3	5	2
RNASE2	3	6	2
RPP25	3	3	2
RSAD2	3	5	2
RTP4	3	5	2
S100A9	3	6	3
SAMD4A	3	5	2
SAMD9	3	5	2
SAT1	3	5	2
SCO2	3	5	2
SCPEP1	3	6	2
SELL	3	6	2
SERPINB8	3	6	2
SERPING1	3	5	2
SIGLEC1	3	5	2
SLC27A3	3	5	3
SLC39A1	3	5	2
SLC43A3	3	5	3
SLC7A7	3	3	3
SLFN12	3	5	2
SMARCD3	3	5	3
SNX11	3	6	3
SP100	3	5	2
SP110	3	5	2
SQRDL	3	5	2
STAT1	3	5	2
STYXL1	3	3	2
TAP1	3	5	3
TBC1D8	3	5	3
TCN2	3	5	3
TDRD7	3	5	2
TFEC	3	5	2
TIMM10	3	5	2
TLR7	3	5	2
TMEM180	3	5	3
TMEM62	3	5	2
TNF	3	5	3
TNFAIP6	3	5	2
TNFRSF8	3	3	3
TNFSF10	3	5	2
TOR1A	3	5	2
TOR1B	3	5	2
TP53I3	3	6	3
TRAFD1	3	5	2

Gene	FLU.classSOM	HRV.classSOM	RSV.classSOM
TREX1	3	5	2
TRIM22	3	5	2
TRIM38	3	5	2
TRIM5	3	5	2
TXNDC4	3	5	3
TYMP	3	5	2
UBE2L6	3	5	2
UNC93B1	3	5	3
UPP1	3	5	2
USP25	3	5	2
VAMP5	3	5	2
XAF1	3	5	2
YKT6	3	6	2
ZBP1	3	5	2
ZCCHC2	3	5	2
ABLIM1	4	1	1
ALDH9A1	4	1	1
APBA2	4	1	1
ATP6V0E2	4	2	1
B3GNT1	4	2	1
BCL11B	4	1	1
BIN1	4	2	1
BTG1	4	1	1
C11orf2	4	1	1
CCL5	4	2	1
CD247	4	1	1
CD27	4	1	1
CD96	4	2	1
CHI3L1	4	1	1
CYFIP2	4	1	1
DHRS3	4	2	1
DOCK9	4	1	1
DPP4	4	2	1
EDG1	4	2	1
EEF2	4	1	1
EIF3EIP	4	1	1
FAIM3	4	1	1
FOXO1	4	1	1
GATA3	4	2	1
KLHL3	4	1	1
LAT	4	2	1
LDOC1	4	1	1
LEF1	4	1	1
LOC130074	4	1	1

Gene	FLU.classSOM	HRV.classSOM	RSV.classSOM
M-RIP	4	1	1
NELL2	4	1	1
P2RX5	4	1	1
PABPC4	4	2	1
PARP16	4	1	1
PIK3IP1	4	1	1
PLCG1	4	2	1
PLEKHG3	4	1	1
POMT1	4	1	1
PPARD	4	1	1
PRKCH	4	2	1
PVRIG	4	2	1
SATB1	4	1	1
SPOCK2	4	1	1
TAF4	4	1	1
TCF7	4	1	1
TRAC	4	2	1
TRAF5	4	1	1
TRBC1	4	2	1
UNC84B	4	2	1
ZHX2	4	1	1
DCHS1	5	1	1
ENTPD1	5	6	2
FLOT1	5	6	3
PGS1	5	5	3
AKR1B1	6	1	1
CD2	6	2	1
CD72	6	1	1
CD79A	6	1	1
CD8A	6	2	1
CDR2	6	1	1
EHBP1	6	1	1
GLTSCR2	6	1	1
GNB2L1	6	1	1
KLRK1	6	1	1
LCK	6	1	1
LY9	6	1	1
NCR3	6	1	1
PASK	6	1	1
RPL18	6	2	1
SPIB	6	1	1
FBP1	7	3	3
TSPAN4	7	3	2

Table S3.2: Pan-viral differential genes (n=395) and SOM cluster designation

CHAPTER IV

Identification of Hoxa9 and Meis1 Regulatory Functions

4.1 INTRODUCTION

Transcription factor HOXA9 is a homeodomain DNA binding protein that plays a critical role in regulating the differentiation, self-renewal, and proliferation of hematopoietic stem cells (HSC) and their committed progenitors. Among a family of 39 Hox genes, Hoxa9 is the most highly expressed in the HSC compartment. It is directly targeted by many genetic abnormalities that lead to hematological malignancies such as leukemia (Sitwala et al., 2008, Moskow et al., 1995, Nakamura et al., 1996, Armstrong et al., 2002, Casas et al., 2003, Ferrando et al., 2003, Look, 1997, Rozovskaia et al., 2001). The resultant over-expression of Hoxa9 has been associated with a variety of human acute leukemias. It is widely accepted that HOXA9 bind to its target sequences by recognizing a four-letter consensus motif (TAAT) (Mann et al., 2009, Shen et al., 1997, 1999). However, such level of transcriptional specificity cannot sufficiently explain the fact that the whole genome contains many more this consensus motif than what are actually occupied by HOXA9. In addition, Hox proteins are highly evolutionarily conserved. Many Hox family members, including HOXA9, recognize the same TAAT consensus sequence. The question of how Hoxa9 achieves its functional specificity remains unknown. Recent studies have shown cooperative DNA binding and interaction between Hox and Meis1 and Pbx1.

These two factors belong to the three amino acid extension (TALE) family which contain non-Hox homeodomain (Shen et al., 1997, 1999, Sitwala et al., 2008). They are believed to increase the transcriptional specificity and binding affinity of Hox. There is a strong correlation between Hoxa9 and Meis1 expression in human acute myeloid leukemia (AML). In addition, both Hoxa9 or Meis1 are required for leukemic transformation and together they confer adverse prognosis. However, it appears that Hox proteins can have either activating or repressing effects on their targets (Sitwala et al., 2008). Such functional dilemma cannot be attributed solely to the binding of Hoxa9 and its known co-factors. Other transcription factors are likely to play a role in providing additional functional specificity to the transcriptional regulation of Hoxa9. Currently, there are only a few such "collaborators" that have been identified (Mann et al., 2009). These "collaborators" may physically interact with the Hox/Meis1 proteins or bind to adjacent cis-acting sequences. They may function synergistically or antagonistically with Hox/Meis1 proteins in specifying the transcription of their common target sequences.

In order to gain better understanding of the regulatory mechanisms of Hoxa9 and its "collaborators", it is important to identify and characterize their direct *in vivo* binding targets. Towards this end, we performed bioinformatics analyses on a dataset including ChIP-sequencing, ChIP-Chip, genetic profiles of Hoxa9 and Meis1 functional experiments. We identified the patterns of regulatory controls on transcription by Hoxa9 and Meis1 through both sequence-binding and epigenetic modifications. In this chapter, I focus on the methodological components of the analysis and discuss the statistical aspects and modeling challenges involved in analyzing such multi-modal datasets. We refer readers to (Huang et al., 2010) for more direct and detailed biological translation of the findings derived from analysis reported here.

4.2 RESULTS

4.2.1 High confidence Hoxa9/Meis1 (H/M) binding sites were determined with ChIP-seq analysis

Recent advances in parallel sequencing allows the regulatory binding patterns of a given DNA-binding transcription factor to be determined on the entire genome with high efficiency. Briefly, the so-called ChIP-seq refers to *chromatin immunoprecipitation coupled with high-throughput sequencing* technique. For a given transcription factor, the process begins by associating this factor and DNA with crosslinking agents such as formaldehyde. This is followed by a selective precipitation of the crosslinked protein-DNA interactions using an antibody that is highly specific to the protein of interest. The associated DNA are reversely crosslinked and fragmented before they are used for constructing the oligonucleotide adaptor-ligated library. These DNA fragments are then selected for the appropriate size, depending on the specific sequencing technology (e.g., 100 to 300 base pairs for Illumina Sequence Analyzer). These sequences in the library are subsequently analyzed *en mass* and in parallel by a sequencer over a period of 3 – 4 days. The resulted shorter sequence reads, normally ranging from 32 to 40 base pairs for the Illumina sequencer, are then mapped and aligned onto a reference genome of choice. See (Mardis, 2008, Schmidt et al., 2009) for more detailed discussion of the ChIP-sequencing technology.

As ChIP-seq provides extremely high resolution of protein-DNA interaction, it also presents some major challenges for quantitative analysis (Park, 2009). Particularly, the determination of peaks requires careful consideration as it affects every single step of downstream analysis. Although many different analysis techniques have been developed for this purpose (Robertson et al., 2007, Fejes et al., 2008, Ji et al., 2008, Rozowsky et al., 2009, Tuteja et al., 2009, Zhang et al., 2010), the high throughput nature of ChIP-seq can

potentially magnify the number of statistical false discoveries by hundreds and thousands. In our analysis, highly stringent selection criteria were imposed to ensure the fidelity of the final set of peaks. Two biological replicates were sequenced for each one of the Hoxa9 and Meis1 factors (Figure 4.1). Peak detection of enriched binding regions were performed using FindPeaks (Robertson et al., 2007) with false discovery rate < 0.05. After mapping peaks onto mouse genome (UCSC *mus musculus* reference genome version 8; build 2006), any peaks that overlap ($\geq$ 1 base pair) with repetitive genomic regions or regions in controls were discarded. These peaks are further required to be identified in both replicates and not in the control ChIP-seq. In the end, a total of 825 high-confidence peaks were identified as being bound by either Hoxa9 and/or Meis1. It is noteworthy that our peak selection criterion turned out to be consistent with a standard that was later adopted by the ENCODE consortium (Rozowsky et al., 2009). Biological validation of these Hoxa9/Meis1 enriched regions were carried out by members of Hess laboratory and showed significant binding of Hoxa9 and/or Meis1 compared to absent binding in controls (Figure 4.2, Huang et al., 2010).

4.2.2 Genome-wide analysis showed dominant distal binding of Hoxa9 and Meis1

When analyzed for their distribution on the genome, the majority of H/M peaks were found to be located in distal intergic (47.9%) and intronic regions (44.8%). Most of these regions tend to be located more than 10kb away from nearest transcription start sites (TSS). Only 5.1% of H/M bindings are located within promoter region that is important for initiating and regulating transcription (Figure 4.1). This is quite surprising given the functional importance of H/M factors in transcription regulation. Nonetheless, H/M bindings are still significantly closer to TSS than the 229 peaks seen in control ChIP-seq regions (Figure 4.1). The distribution pattern of H/M bindings was also analyzed on individual

chromosome basis (Ji et al., 2006, Shin et al., 2009). Overall, the H/M bindings are evenly distributed across different chromosomes (Figure S4.9). However, in the case of chromosomes 11 and 16, H/M tend to bind more frequently than expected given the relative size of the two chromosomes. This is somewhat interesting because these two chromosomes are among the ones where genetic abnormalities such as translocation often occur.

4.2.3 Hoxa9 and Mesi1 selectively bind to DNA sequences that are highly evolutionarily conserved

A visual examination showed that many of H/M peaks align well with evolutionarily conserved regions on genome. We thus evaluated the level of conservation within H/M peaks. The evolutionary conservation scores were obtained from UCSC phastCons17way database (Siepel et al., 2005) for enriched H/M regions and their adjacent regions extending up to X6 in width. For each H/M peak, a corresponding *virtual* peak is constructed by combining its left- and right-side genomic region subject to the same width of this H/M peak. In a sliding widow fashion, six *virtual* peaks were constructed contiguously outwards from the center of an H/M binding region with no overlap between each other. A two-sample t-test is performed to compare the average conservation score within H/M peaks against that of their corresponding *virtual* peaks. The results clearly showed that the H/M peaks are highly conserved evolutionarily (Figure S4.1). The average conservation score within the H/M peaks is significantly higher than that of their first neighbor *virtual* peaks (1.58 fold change; p-value=5.1E-5). The sharp elbow-shape drop of conservation in the regions sounding the H/M peaks shows that Hoxa9 and Meis1 selectively bind to genomic sequences that are evolutionarily important.

4.2.4 H/M peaks show high potential of regulatory functions

The high level conservation of H/M peaks suggest that H/M bound sequences are functionally important as a result of evolutionary selection pressure. It is still not clear whether they are directly involved in regulating target gene transcription. We next evaluated these H/M peaks for their regulatory potential (RP) (Taylor et al., 2006). The RP is defined on the genomic sequence level based on a comparative genetics study of seven mammalian species. It was shown that high RP scores predict regulatory elements with $\sim 94\%$ accuracy . For each base pair position within a $\pm$8kb window, we computed the average RP scores and the average ChIP-seq sequencing tags across all peaks. The RP scores associated with H/M binding regions show a remarkable coincidence with the sequence tags measured by ChIP-seq ($\rho = 0.81$, p-value≤ 0.0001). The average RP score of H/M peaks is greater than 0.065 whereas score values above 0 correspond to strong regulatory potential of the sequences (Figure 4.3A; Taylor et al., 2006). These findings indicate that H/M peaks, relatively distant from TSS, are associated with high regulatory potential and may function as enhancer sequences.

4.2.5 H/M peaks show epigenetic signatures that are characteristic of enhancers

To validate the hypothesis that H/M peaks function as enhancers, we designed a ChIP-Chip custom tiling array (Nimblegen) to assess whether H/M peaks have the epigenetic characteristics of a typical enhancer, including high-level histone H3 and H4 acetylation, H3K4 monomethylation (H3K4me1), low level histone H3K4 trimethylation (H3K4me3) and high level binding of the histone acetyltransferases p300 and CBP (Figure 4.3) (Heintzman et al., 2007, Visel et al., 2009). All H/M sequences and a selected set ($n = 360$) of the nearest TSSs were tiled, along with 60 negative control sequences that were selected randomly from the mouse genome. The H/M peaks were extended to $\pm$4kb surrounding

regions and TSS were extended to $-2/+1$kb at the TSS.

These experiments revealed that the majority of H/M peaks indeed carry strong enhancer signatures. The H3 and H4 acetylation and p300/CBP binding are clearly centered on the H/M peaks and is flanked by regions of histone H3K4me1 in a bimodal distribution (Figure 4.3B). It was also evident that H/M peaks are unlikely to be promoters, as they differ significantly from the signature of the promoters at adjacent tiled genes. In particular they lack the signature of nucleosomal eviction at the TSS (Figure 4.3C). Interestingly, a subset of the H/M peaks (15%) also showed elevated level of H3K27 trimethylation. There is also another small set of H/M peaks do not show any level of enrichment of epigenetic signals that were tested in this study (Figure 4.3D). Taken together, these results suggest that H/M peaks represent a functionally heterogeneous set of regulatory elements.

4.2.6 Temporal gene expression revealed Hoxa9 regulation on genes mediating proliferation, inflammation and differentiation

The above statistical and experimental analysis results showed that H/M peaks possess regulatory potentials of long-range enhancer sequences that control target gene transcription. We therefore examined the affect of H/M binding sites on the transcription of neighboring genes. Gene expression were profiled on a myeloblastic cell line that is stably retrovirus-transduced with a conditional form of Hoxa9. Specifically, a modified estrogen receptor ligand binding domain (ER) is fused in-frame to the C-terminus of Hoxa9. In the presence of 4-hydroxytamoxifen (4-OHT), an estrogen hormone analog, the protein coded by Hoxa9-ER fusion gene is stably localized in the nucleus and the cells grow normally. Upon 4-OHT withdrawal, the Hoxa9-ER fusion protein rapidly degradate in cytoplasma and the cells undergo growth arrest and differentiate into macrophages (Figure S4.2). For gene expression profiling, RNA from cells harvested at 48, 72, 96 and 120 hours (hrs) following 4-OHT withdrawal was analyzed in triplicate by Affymetrix

microarray hybridization.

No statistically significant gene expression changes (FDR p-value < 0.05) were observed until 72 hours after 4-OHT withdrawal. When expression profiles are analyzed over the complete time course, 6,991 genes show significant changes in expression post 4-OHT withdrawal (composite significance criterion: FDR p-value < 0.05 and median fold change > 1.5). Based on their temporal expression patterns, these genes were clustered into four subgroups (Figure S4.2E). The first group (cluster 1) represents a large number of Hoxa9 upregulated genes ($4,253$) whose expression started decreasing beginning 72 hours after 4-OHT withdrawal and remained decreased at 120 hours. Gene ontology (GO) analysis of this group showed extremely high association with RNA processing (Fisher exact p-value=6.63E-68; Table 4.1), DNA metabolic processes (1.73E-51) and cell cycle regulatory genes (1.79E-44). The group includes Camk2d, Cdk6, Erg, Etv6, Flt3, Foxp1, Gfi1, Kit, Lck, Lmo2, Myb and Sox4, which are strongly implicated in either murine or human leukemias (Figure 4.4) (Li et al., 1999, Mikkers and Berns, 2003). Of the 52 targets showing greater than five-fold altered expression between Hoxa9- and Hoxa9+Meis1-immortalized mouse hematopoietic progenitors published by Wang et al. (Wang et al., 2005), 14% showed association with the H/M ChIP-Seq peaks identified here.

A second major group (cluster 4) represents Hoxa9 downregulated genes (n=2,502) whose expression levels decreased over time (Figure 4.4; Figure S4.2E). GO analysis shows their strong associations with immune response (Fisher exact p-value=1.82E-26; Table 4.1), inflammatory response (1.82E-22) and cell activation (5.57E-14). Many of these genes are located near the H/M binding sites and are related to inflammation and myeloid differentiation, including Ifit1, Tlr4, Ccl3, Ccl4, Csf2rb, Ifngr1, Runx1, Cd28, and Cd33.

Two other clusters showed more transient expression dynamics. One group (cluster

2) consists of 100 genes whose expression was increased at 72 and 96 hrs but decreased by 120 hrs. Many of these genes are involved in cholesterol biosynthesis (6.13E-05; Table 4.1) or sterol biosynthesis (1.38E-4). Another cluster includes 136 genes with decreased expression at 72 and 120 hrs that showed increased expression at 120 hrs. These genes were associated with pattern specification (1.99E-03; Table 4.1) or regulation of nervous system development (2.34E-03).

It is worth noting that we investigated whether there is any direct relationship between H/M peaks and the expression of their nearest genes. We were not able to identify any particular patterns that would attribute the up-regulation or down-regulation of gene expression to the presence of H/M peaks. We also did not observe any linear or nonlinear relations between gene expression and spatial distance between H/M peaks and genes. This suggests that H/M sequences exert functional regulation on their target genes via a mechanism that cannot be simply explained by their genomic arrangement. A more sophisticated regulatory mechanism is likely to be in play, such as specific chromatin structure configuration or DNA looping etc. Such complicated regulatory mechanism will need to be evaluated with more recent biochemical techniques such as Chromosome Conformation Capture (3C) (Dekker et al., 2002).

4.2.7 De novo motif discovery suggested binding of H/M collaborators

A key feature of ChIP-seq technology is that it detects not only where on genome a transcription factor binds, but also exactly what sequences it binds. Under the premise that all DNA-binding transcription factors bind to their targets with high specificity by recognizing their corresponding *consensus sequences*, or *motifs*, H/M peaks made it possible to identify other factors that co-bind with Hoxa9 and Meis1.

We performed *de novo* motif discovery to computationally search for sequence pat-

terns that are shared by the H/M peaks. Three independent runs of de novo motif discovery analysis were performed using Gadem (Li, 2009) and identified motifs of 15 transcription factors or complexes. Several interesting findings were revealed from this analysis. Firstly, it validates the high quality of the H/M peaks as Hoxa9 and Meis1 canonical binding motifs are the most significantly enriched. It occurred in 53% and 34% of H/M sequences, respectively (Figure 4.5A). Secondly, it revealed the enrichment of motifs for a group of transcription factors with some of them known to be functionally related to Hoxa9 or Meis1 or leukemia in general. For instance, the second most frequently occurring motif, which was found in 51% of peaks, is for the ETS family of transcription factors. In addition, a total of 204 sequences (36%) contain the CCAAT-enhancer binding protein C/EBP motif, with more than half of them occurring in combination with the Hoxa9-Meis1-Pbx1 motif. This was followed in frequency by motifs for RUNX, which was present in 14% of peaks and the STAT motif, present in 10% of peaks. There are also motifs for a number of other TF deregulated in hematologic malignancies including MAF and E2A. Thridly, an additional five consensus novel motifs were identified that do not match any known transcription factors (Figure S4.3). They may correspond to motifs of known factors whose consensus motifs are unknow or motifs of unidentified factors.

As *de novo* motif discovery solves a computationally difficult problem, the results can vary a great deal among different techniques (Tompa et al., 2005). It is therefore of great importance to validate the results generated from one method with analysis results of others. In our analysis, we applied three alternative *de novo* discovery methods, including Weeder, CisGenome, and MEME (Pavesi et al., 2004, Ji et al., 2008, Bailey et al., 2006, 2009). These analysis methods represent the entire spectrum of existing methodologies - from exhaustive enumeration (Weeder) to deterministic expectation maximization (MEME) and stochastic expectation maximization (CisGenome). They all yielded com-

parable results to that of Gadem. We therefore conclude that the *de novo* identified motifs are valid and indicative of involvement of multiple transcription factors at the H/M binding regions. Furthermore, the clustering of motifs for many transcription factors in such close proximity suggest that these factors may form a highly integrated regulatory circuit that collaboratively controls the target transcription.

4.2.8 Motif enrichment analysis (MEA) revealed tiered organization of transcriptional control

The *de novo* motif discovery clearly showed that the H/M peaks are enriched with motifs for multiple transcription factors. This raised the possibility that there exist more motifs of other factors than what were identified by *de novo* method. We next sought to directly incorporate the *a priori* motif models that have been experimentally characterized. We searched in H/M peaks for all 727 motif models (170 families) included in Genomatix proprietary Mat Base Matrix Family Library (Version 8.2, January 2010). A motif is considered to be statistically significantly enriched in the H/M peaks if the number of sequences in which the motif is found to be present is significantly higher than its expected whole-genome occurrences according to standard z-test (z-score > 2.81; p-value < 0.005).

Compared to *de novo* discovery, the MEA showed consistent findings with higher level of enrichment of motifs for HOX, MEIS, C/EBP, CREB, STAT RUNX1 and ETS (especially PU.1) (Table 4.2). The Hoxa9 and Meis1 motifs are strikingly concentrated at the center of the binding regions while RUNX1, STAT, CREB and C/EBP motifs are more broadly distributed across the locus of H/M peaks (Figure 4.5B, Figure S4.4, Figure S4.5A). The MEA also identified a number of other significantly enriched motifs for bZIP, MEIS, caudal MYB and MYC TF family members. The complete list of enrichment scores is provided in Table 4.2.

ChIP-chip experiments in myeloblastic cells were carried out to confirm the binding

of a subset of the TF identified in motif analysis. These experiments showed that the H/M peaks are extensively co-bound by C/EBP , Pu.1, and Stat5a/b (Figure 4.5C). In addition, these peaks show higher levels of CBP and p300 binding than H/M peaks lacking C/EBP and Pu.1 motifs. This coincidence of TF binding in H/M peaks is noteworthy given recent studies showing that C/EBP , Pu.1 and Runx1 physically interact (Petrovick et al., 1998) and collectively define myeloid enhancer sequences in murine macrophages (Heinz et al., 2010). Remarkably, over 30% of the 825 H/M peaks overlap with previously identified myeloid enhancers (Figure 4.5D). This creates the opportunity to identify determinants of Hoxa9 binding specificity by comparing motifs that are enriched in H/M bound enhancers versus those that are not. In this analysis major differences were seen in the frequency of HOX consensus motifs. Enrichment was also seen in motifs for a heterodimer of the HOX cofactors MEIS1 and PBX (Figuree 4.5E, Figure S4.5B). Other motifs enriched in H/M peaks but under-represented in enhancers bound by C/EBPA alone included STAT (Figuree 4.5E), MYB, estrogen response elements (ERE), BRN5, PDX1, CDX and peroxisome proliferator activated protein (PPAR) (Figure S4.5B). Collectively, these findings suggest that Hoxa9 is organized into specialized enhancersome or Hoxasomes composed of lineage-specific TFs (Mann et al., 2009).

Our experiments show that Hox consensus sequences are present in over half of Hoxasomes (Figure 4.5A; *de novo* motif analysis) and are a major determinant of Hoxa9 targeting to these enhancers. However, this sequence is present at numerous sites in the genome. Therefore, the genomic binding specificity by itself cannot account for entirely the Hox functional specificity (Mann et al., 2009). Interaction with cooperatively binding cofactors, such as the Meis and Pbx families further increases DNA binding affinity and specificity (Mann et al., 2009). However this cannot completely account for the specificity of Hox binding as these Hox-Meis-Pbx sequences also occur far more frequently than the

number of observed sites. Our experiments suggest that a third tier of Hoxa9 specificity is achieved through combinatorial interactions with TF such as C/ebp, Stat5 and Creb1, each expressed at variable levels in a specific cell type that collectively account for lineage specific Hoxa9 recruitment (Heinz et al., 2010) (Figure 4.6). A variety of biochemical experiments using immunoaffinity purification and mass spectroscopy were performed and confirmed that Hoxa9 is targeted to Hoxasomes through homeodomain and protein-protein interactions, resulting in their stabilization and increased coactivator activity (Huang et al., 2010).

4.2.9 Epigenetic state at H/M peaks are correlated with specific motif configuration

As motifs indicate binding potential of transcription factors, the configuration of motifs may provide important information regarding the effects of combinatorial binding by multiple factors on regulating transcription. Towards this end, we investigated the relationships between motifs and epigenetic signatures of H/M peaks. Using sparse canonical correlation analysis (sCCA) (Hotelling, 1936, Witten et al., 2009), we identified a series of motif configurations (eigen-motif) that are most highly correlated with different epigenetic signatures (eigen-epigenetics) (Table 4.3).

In particular, the first pair of canonical variables showed that ETS motif correlates highly with H4 acetylation. The second, third, and fourth pairs of canonical correlations suggest that enrichment of C/EBPA, Hoxa9, and Stat5 motifs in H/M binding accounts for most *in vivo* binding signals in C/ebpα, Hoxa9, and Stat5 epigenetic profiles. Interestingly, the fifth pair of canonical variates suggest that the presence of GFI1 motif corresponds to the impressive H3K27 trimethylation marker in H/M peaks. There are evidence showing that Gfi1 competes with Hoxa9 for binding sites. This raises the possibility of Gfi1 interfering with the function of Hoxa9 by increasing the amount of suppressive

H3K27m epigenetic modification at regions where Hoxa9 normally bind. Taken together, these results indicate the specific eigen-motif combinations correlate with different epigenetic signatures in H/M peaks. The co-bindings of multiple transcription factors are likely to have important effects on the epigenetic state at H/M binding sites. It is noteworthy that such correlation may be cellular context-dependent because the co-bindings of these factors are lineage specific (Huang et al., 2010).

4.3 DISCUSSION

Hoxa9 transcription factor and its cofactors such as Meis1 and Pbx1 play critical roles in normal development, hematopoiesis, and leukemia. The mechanism through which they regulate transcription and mediate leukemic transformation were not well understood. This research study identified direct *in vivo* binding sites of Hoxa9 and Meis1 and showed that these binding sites serve as the genomic basis for integrating lineage-specific transcription factors to form specialized Hox-regulated enhancersomes or "Hoxasomes" (Huang et al., 2010). The components of hoxasomes include important transcription factors such as C/ebpα, Pu.1, Stat5, and Creb1. Our findings showed that Hoxa9 binding at these sites stabilizes hoxasomes and promotes the recruitment of histone acetyltransferase CBP and P300. This model suggests that the overexpression of Hoxa9 is likely to destabilize the hoxasomes and cause the deregulation of their components, a process that leads to the disruption of cellular function and eventually leukemias (Huang et al., 2010).

From methodological point of view, this study offered several recommendations concerning the multi-modal data analysis, particularly that involving ChIP-seq data. Because of the high technical variability in sequencing, determining the binding sites is a critical issue that cannot be emphasized enough. Regardless of the particular peak calling algorithms used, selection of final set of peaks should be carefully treated. Our results

suggested that enrichment analysis of the target TF motif in the selected sequences is a reasonable way to assess quality of final set of peaks. Our H/M peaks showed more than 98% of enrichment of the canonical Hoxa9-Meis1-Pbx1 motif. This is much higher than what have been reported in literature using ChIP-seq to study other TFs (50% $\sim$ 70%). The selection procedure employed in our analysis ensures the lowest false discovery rate possible. An argument aginst such high standard is that it may be unncecessarily strict and prevents some valid binding sites to be included. Considering however the importance of providing quality binding sites for future biological investigation, it is our belief that such critierion is in the best interest of long-term high quality research. This is echoed in a newly adopted ENCODE consortium standard (Rozowsky et al., 2009).

The motif analysis have become the *de facto* standard in ChIP-seq analysis. However, less attention was given to motif enrichment analysis (MEA). Our findings demonstrated the merit of applying MEA in a general ChIP-seq analysis while *de novo* analysis is more appropriate for finding short length novel binding motifs. MEA takes advantage of a large body of biological knowledge about binding patterns of transcription factors on the order of hundreds. Compared to *de novo* methods, MEA can be carried out rapidly with little or no difficulty. It also eliminates the need for post-analysis query to known TF database because all motif models are known *a priori*. More important, MEA is a deterministic process where enrichment statistic of a motif model is well defined using classic significance measure. On the other hand, the results of *de novo* methods can be affected by a variety of parameters — optimization criteria, sequence background model, similarity measure used for TF model matching etc (Prakash and Tompa, 2005, Tompa et al., 2005, D'Haeseleer, 2006). The *de novo* method is also handicapped by its limitation on the length of motifs that it can identify. A search for a length of $>$ 20bp is almost computationally infeasible. However, motifs of longer length provide important information about

collaborative binding of multiple transcription factors such as Hoxa9-Meis1-Pbx1 protein complex in our case. Our experiences showed that MEA is an extremely valuable tool in finding such multi-factor collaborative binding. Notably, we developed an information geometric estimation and inference framework to study putative protein interactions (VI).

In a complete functional study of a TF, it is critical to jointly analyze the binding data offered by ChIP-seq with other data types such as gene expression and motif enrichment measures. The reason is obvious — the exact locations of bindings on genome alone will not answer all the questions related to the functional role of a TF. That information needs to be integrated with other biological measures to collectively provide useful insights. When jointly modeling ChIP-seq data with other data types, there are two important issues need special consideration. Firstly, the data derived from ChIP-seq sequencing are often discrete. For instance, the basepair sequencing reads and motif occurrences are both discrete. Other types of data such as gene expression or epigenetic profiling data can be either discrete or continuous. Secondly, all datasets are high-dimensional, consisting of hundreds or thousands of variables. Biologically, it is unlikely that all variables in one dataset are related to all variables in another. Rather, they are loosely coupled through associations between subsets of variables or biological components. This presents a very interesting problem for statistical analysis to identify functionally meaningful patterns or associations. Our results suggested that canonical correlation analysis (CCA; Hotelling, 1936, Witten et al., 2009), is a natural fit for such situation and can readily handle all the aforementioned difficulties. It is worth noting that if all variables are categorical, then CCA simply reduces to a classic correspondence analysis (Benzkri, 1982).

In summary, this study revealed functional mechanisms by which Hoxa9 regulates transcription and provides insights into the pathogenesis of acute leukemias. It offers new practical solution to integrate and jointly analyze the next-generation ChIP-seq data with

other modality experimental data.

4.4 MATERIALS AND METHODS

4.4.1 Statistical Analysis

Clustering of Hoxa9-ER significant genes All significant genes were clustered into four groups based on their temporal patterns using Self-Organizing Maps (Kohonen, 1995) in a similar way as Huang, et al (Huang et al., 2001). Prior to clustering, the expression of each gene was normalized to have zero mean and unit variance. A hexagonal topology of 2 by 2 grid of prototypes was initiated and each gene was subject to the network 400 iterations. The L_2 distance measure was used as distance metric. The centroid and one standard deviation of each cluster was computed and plotted in Figure S4.2E. The significant genes were mapped to closest H/M peaks using CisGenome (Ji et al., 2008)

Motif enrichment analysis Comprehensive search of known transcription factor binding motifs was performed for 748 mouse transcription factors included in Genomatix proprietary Mat Base Matrix Family Library (Version 8.2, January 2010) that includes a total of 727 motifs (170 motif families). The DNA sequences of length 300bp from the center of each H/M peak were scanned for presence of any known transcription factor binding motif. A transcription factor binding motif is considered to be statistically significantly enriched in the H/M peaks if the number of sequences in which the motif is found to be present is significantly higher than its expected whole-genome occurrences according to standard z-test (z-score > 2.81; p-value < 0.005).

Sparse canonical correlation analysis of motif configuration and epigenetic profiles We seek two coefficients vectors a_i and b_i that maximize the cross correlation

$$\arg\max_{a_i,b_i} \operatorname{corr}(a_i^T X_i, b_i^T Y_i) \tag{4.1}$$

where $X_{\{n,p\}}$ represents the epigenetic profiles and $Y_{\{n,m\}}$ is the motif enrichment scores (Figure S4.8). Regularizing of coefficient vectors (making some coefficients zero) yields a sparse solution identifying a subset of m motifs that correlate to a subset of p epigenetic signatures. Intuitively, the new canonical variates $U_i = a_i^T X_i$ and $V_i = b_i^T Y_i$ represent a pair of specific motif configuration and specific epigenetic state that are both expressed as a linear combination of the measured X and Y. A total of $min(p,m)$ such pairs can be found in a decreasing order of canonical correlation.

Method implementation The proposed method and visualization and utility functions for motif analysis were implemented in a R package *cMotif*. The code is publicly available at the **Hero Group Reproducible Research** archive under **cMotif** and the official R repository (soon). Several primary functions in the library *libmotif* are listed in

Data Availability The ChIP-sequencing data, gene expression profiles, and Nimblegen tiling array data were deposited to NCBI's Gene Expression Omnibus (Edgar et al. 2002) with GEO series accession number GSE21299. Visualization tracks (UCSC) containing all ChIP-sequencing data, Nimblegen epigenetic modification data for Hoxa9 and Meis1 binding sites, Hoxa9-ER gene expression data, and motif enrichment analysis results are available at http://www.pathology.med.umich.edu/faculty/Hess/index.html.

4.4.2 Experimental Procedure

Chromatin immunoprecipitation (ChIP) A total of 150 million cells were crosslinked sequentially with disuccinimidylglutarate (45 min RT) and 1% formaldehyde (15 min RT). Hoxa9 and Meis1 immunoprecipitation was performed with anti-HA antibody (Abcam) pre-conjugated to Protein G magnetic beads (Dynal/Invitrogen). For C/ebpα ChIP, rabbit anti-C/ebpα (Santa Cruz) was compared with pre-immune rabbit IgG. 4-hour incubation (4 degree with gentle rotation) was followed by washes using Low Salt, High Salt, LiCl,

and Tris-EDTA buffers (Upstate/Millipore). Immunoprecipitates were eluted with 0.1% SDS/0.1M NaHCO3 and DNA-protein crosslinks were reversed overnight at 65 degree in 0.2M NaCl. DNA was RNAse treated and column purified (Qiaquick, Qiagen).

For ChIP-seq, size selection and sequencing were performed at the BC Cancer Agency Genome Sciences Centre (Vancouver, BC) as described previously (Robertson et al., 2007). Peak detection of enriched binding regions was performed using FindPeaks (Robertson et al., 2007) with an estimated false discovery rate < 0.05 as the selection criterion for enriched regions. For ChIP-Chip, DNA was amplified prior to dual hybridization (performed at Nimblegen Systems) of input and immunoprecipitate on a custom tiled mouse genomic array containing putative Hoxa9 and Meis1 target genes (50-mer probes with an average spacing of 35 bp; 15 megabases of total sequence).

ChIP-chip tiling array design Enriched regions were retiled onto Nimblegen 385K/2.1M custom tiling arrays based on the version 8 (Feb 2006) *Mus musculus* reference genome. Each ChIP-seq binding site was extended to 4kb in both directions and the resulting region was tiled with probes at 35bp spacing. In addition, the design included a set of 360 transcription start sites that were closest to ChIP-seq enriched regions, and a set of 60 control regions that were randomly selected from individual chromosomes. The tiled TSS regions were directionally extended 2kb upstream and 1kb downstream. ChIP was performed with antibodies to H3K27me3, H3K9ac, H3K4me1, H3K4me3 and normal IgG control. The chipped DNAs were hybridized to two-color arrays with input sample labeled with Cy3 dye (wavelength=532nm) while experimental ChIP DNA was labeled with Cy5 (wavelength=635nm). Each array consisted of 383,370 reporter probes with 100 bp spacing.

Cell line generation Bone marrow cells were harvested from 5-Fluorouracil treated female 6-8 week old C57BL/6 mice and transduced with an MSCV-based retrovirus expressing HA-tagged or untagged Hoxa9 or Meis1, or Hoxa9 fused to a modified estrogen

receptor ligand binding domain (Hoxa9-ER). Cells were cultured in Iscove's Modified Dulbecco's Medium with 15% Fetal bovine serum (Stem Cell Technologies) and penicillin/streptomycin. IL-3 (R&D) was added to media; alternatively cells were transduced with an IL-3-expressing retroviral vector (pMFGmIL3, obtained from RIKEN DNA Bank with consent of Dr. Hirofumi Hamada); Hoxa9-ER cells were also supplemented with 4-OHT (Sigma). Hoxa9 is required for continued MHP survival, so positive selection of transduced clones was not necessary; double Hoxa9/Meis1 transductants were selected by fluorescence-activated cell sorting (bicistronic Meis1+GFP expression using MigR1 vector; a gift from Dr. Warren Pear).

Plasmids, electroporations, and luciferase reporter assays Twenty-two Hoxa9/Meis1 binding sites were selected based on their proximity to the nearest TSS as well as two control regions randomly selected in the genome where there is no Hoxa9/ Meis1 enrichment. Each $\sim$ 1000bp binding region was amplified from mouse genomic DNA using Advantage HD Polymerase(Clontech) and following restriction enzyme digest, the fragments were cloned into the multiple cloning site of the pTAL-Luc vector(Clontech) using In-Fusion Cloning PCR System(Clontech). For luciferase assays, 2μg of pTAL-luc construct and 500ng of renilla vector were used per electroporation in K562 cells suspended in Gene Pulser Electroporation Buffer (Bio-Rad). Electroporations were performed in 96-well plates using the Gene PulserMXcell Electroporation System (Bio-Rad). After 48 hours upon electroporation, luciferase activity was measured using Dual Luciferase Reporter Assay (Promega), normalizing firefly luciferase to Renilla luciferase.

ChIP-qPCR ChIP was quantified relative to inputs using Taqman probes and an ABI 7500 Real Time PCR System (Applied Biosystems). Taqman primer and probe sequences were designed using Primer Express Software 3.0 (Applied Biosystems)and are available upon request.

Quantitative RT-PCR Primers were designed using DNASTAR software; sequences will be provided upon request. Relative quantitation of real time PCR product was performed using the comparative DDCt method with SYBR green fluorescent labeling on ABI 7500 PCR Detection System. All experiments were performed on at least two different days and yielded similar results.

Gene Expression profiling analysis Hoxa9-ER cells were washed 3x and resuspended in IL-3+ media with/without 100 nM 4-OHT (Sigma). At selected intervals, cells were removed for flow cytometric analysis using anti-Gr1 and anti-Mac1 antibodies (BD Biosciences), morphologic assessment by cytocentrifugation followed by staining with Diff-Quick reagents (Intl. Med. Equip.), and RNA collection. For RNA, pellets were lysed in Trizol reagent (Invitrogen) and RNA was extracted following manufacturer's instructions until phase separation, after which RNeasy columns (Qiagen) were employed for further purification. cRNA probes were synthesized at the University of Michigan microarray core. Probes were hybridized to Affymetrix Mouse 430 2.0 array.

Identification and verification of interacting proteins Proteins were extracted from 1 x 109 HM2 cells using M-PER (Pierce Biotechnology). Thenuclear pellet was solubilized in 250 U/mL benzonase nuclease (EMD Biosciences). After pre-clearing with IgG, immunoprecipitation was performed with anti-HA Affinity Matrix (Roche Applied Science) or anti-FLAG M2 Affinity Gel (Sigma). Bound proteins were washed with M-PER+300 mMNaCl and eluted in SDS-PAGE sample loading buffer. Western blot detection was performed with rabbit polyclonal anti-Hoxa9 (Millipore), rabbit polyclonal anti-Meis1 (Abcam), and rabbit polyclonal anti-Cebpa (Cell Signaling Technology). For mass spectrometry, SDS-PAGE gels were stained with Colloidal Blue (Invitrogen), and lanes were cut into 16 slices for destaining and cysteine reduction/carbamidomethylation (10 mM DTT+50 mMiodoacetamide). Macerated and dried gel slices were re-swollen and di-

gested in ammonium bicarbonate buffer with trypsin (Promega). Peptides were extracted sequentially in using acetonitrile/TFA gradient; extracts were pooled and concentrated prior to reverse phase chromatography (Aquasil C18, Picofrit column, New Objectives). Eluted peptides were directly introduced into an ion-trap mass spectrometer (LTQ-XL, ThermoFisher) with a nano-spray (in MS/MS mode). Data were converted to mzXML format and searched against mouse IPI (v 3.50) + reverse database using X!Tandem with k-score plug-in (Global Proteome Machine). Outputs were subjected to PeptideProphet29 and ProteinProphet30 analysis; proteins with a ProteinProphet probability of ¿ 0.9 were considered for further analysis. MS/MS spectra corresponding to proteins that were unique to the experimental sample were manually verified.

Nuclease-treated extracts were pre-cleared with rabbit IgG conjugated agarose beads (Santa Cruz Biotechnology), and the target proteins immunoprecipitated with either anti-HA Affinity Matrix (Roche Applied Science) or anti-FLAG M2 Affinity Gel (Sigma). Bound proteins were washed with M-PER+300 mMNaCl and eluted in SDS-PAGE sample loading buffer. Western blot detection was performed with rabbit polyclonal anti-HoxA9 (Millipore), rabbit polyclonal anti-MEIS1 (Abcam), rabbit polyclonal anti-Stat5 (C-17; Santa Cruz Biotechnology), rabbit polyclonal anti-C/EBP (Cell Signaling Technology), and rabbit monoclonal anti-CREB (48H2; Cell Signaling Technology).

4.5 ACKNOWLEDGEMENTS

I thank Jay Hess, Alfred Hero, Kajal Sitwala, Monisha Dandekar, Joel Bronstein, Daniel Sanders at the University of Michigan and Gordon Robertson, Timothee Cezard, Misha Bilenky, Nina Thiessen at the British Columbia Cancer Agency for their valuable discussions that lead to the overall design of bioinformatic analysis. All the biological experiments including cell culture, ChIP sample preparation, rtPCR, flow cytometry, mass

spectrometry were designed and/or performed by Kajal Sitwa, Monisha Dandekar, Joel Bronstein, Daniel Sanders, Venkatesha Basrur, and Mark Deming. The initial *de novo* motif analysis using Gadem was performed by Gordon Robertson. The data normalization and initial analysis of gene expression profiles were performed by James MacDonald. This work was supported by a US National Institutes of Health grant R01 CA116570-01A1 to J.L.H. and by a Pilot Grant from the University of Michigan Center for Computational Medicine and Bioinformatics (CCMB) to J.L.H and A.O.H.

Hoxa9 Regulated Genes

Class	GOTerm	P-Value	Class	GOTerm	P-Value
1 Up-regulation (Sustained)	RNA processing	6.63E-68	**2 Up-regulation (Transient)**	cholesterol biosynthetic process	6.13E-05
	DNA metabolic process	1.73E-51		sterol biosynthetic process	1.38E-04
	cell cycle	1.79E-44		isoprenoid metabolic process	6.00E-04
	ncRNA metabolic process	1.86E-40		cholesterol metabolic process	1.70E-03
	cell cycle phase	2.17E-39		steroid biosynthetic process	1.77E-03
	cell cycle process	1.84E-35		sterol metabolic process	2.23E-03
	mRNA metabolic process	1.95E-32		isoprenoid biosynthetic process	2.48E-03
	RNA splicing	5.84E-32		lipid biosynthetic process	1.55E-02
	DNA repair	5.29E-29		steroid metabolic process	1.71E-02
	chromosome organization	2.90E-22		response to oxidative stress	3.50E-02
				response to inorganic substance	3.87E-02
3 Down-regulation (Transient)	pattern specification process	1.99E-03	**4 Down-regulation (Sustained)**	immune response	1.82E-26
	regulation of nervous system development	2.34E-03		inflammatory response	1.82E-22
	anterior/posterior pattern formation	2.70E-03		cell activation	5.57E-14
	regulation of homeostatic process	6.25E-03		intracellular signaling cascade	1.68E-13
	regulation of neurogenesis	9.40E-03		positive regulation of immune system process	4.02E-12
	gut morphogenesis	1.14E-02		regulation of phosphorylation	4.27E-12
	positive regulation of hydrolase activity	1.42E-02		regulation of leukocyte activation	5.66E-11
	response to protein stimulus	1.72E-02		leukocyte activation	6.32E-11
	regulation of cell development	1.76E-02		regulation of cell proliferation	1.35E-10
	neuromuscular process controlling balance	1.91E-02		hemopoietic or lymphoid organ development	2.64E-09

Table 4.1: Gene ontology (GO) analysis of Hoxa9 regulated genes.

TF Families	Number of Sequences	Number of Matches	Expected (genome)	Std.dev.	Over representation (genome)	Z-Score (genome)
V$CEBP	*437*	*624*	*279.41*	*16.7*	*2.23*	*20.6*
V$HOXC	*581*	*836*	*436.39*	*20.86*	*1.92*	*19.13*
V$TALE	*337*	*421*	*222.69*	*14.91*	*1.89*	*13.26*
V$ABDB	*642*	*1198*	*830.79*	*28.75*	*1.44*	*12.75*
V$ETSF	*581*	*976*	*660.43*	*25.65*	*1.48*	*12.28*
V$PBXC	*284*	*362*	*196.11*	*14*	*1.85*	*11.82*
V$AP1F	*190*	*317*	*166.32*	*12.89*	*1.91*	*11.65*
V$MYBL	*341*	*460*	*276.95*	*16.63*	*1.66*	*10.98*
V$PARF	*447*	*902*	*627.62*	*25*	*1.44*	*10.95*
V$HAML	*214*	*253*	*134.55*	*11.59*	*1.88*	*10.17*
V$AP1R	*345*	*554*	*368.57*	*19.18*	*1.5*	*9.64*
V$HOXH	*276*	*348*	*212.27*	*14.56*	*1.64*	*9.29*
V$CDXF	*394*	*557*	*386.35*	*19.63*	*1.44*	*8.67*
V$EBOX	*187*	*288*	*183.74*	*13.55*	*1.57*	*7.66*
V$AARF	*68*	*70*	*29.93*	*5.47*	*2.34*	*7.23*
V$GFI1	*227*	*256*	*164.93*	*12.84*	*1.55*	*7.06*
V$HIFF	*80*	*131*	*73.27*	*8.56*	*1.79*	*6.69*
V$CREB	*385*	*646*	*513.95*	*22.64*	*1.26*	*5.81*
V$CSEN	*76*	*83*	*46.03*	*6.78*	*1.8*	*5.38*
V$HESF	*117*	*174*	*121.72*	*11.03*	*1.43*	*4.69*
V$AP4R	*76*	*95*	*61.45*	*7.84*	*1.55*	*4.22*
V$TCFF	*54*	*56*	*33.05*	*5.75*	*1.69*	*3.91*
V$CHOP	*44*	*45*	*25.73*	*5.07*	*1.75*	*3.7*
V$RBP2	*49*	*53*	*32.73*	*5.72*	*1.62*	*3.45*
V$CDEF	*14*	*15*	*6.18*	*2.49*	*2.43*	*3.35*
V$PAX3	*88*	*91*	*64.68*	*8.04*	*1.41*	*3.21*
V$NF1F	*140*	*171*	*134.36*	*11.59*	*1.27*	*3.12*
V$AP2F	*51*	*59*	*39.05*	*6.25*	*1.51*	*3.11*
V$STAT	*301*	*577*	*510.96*	*22.57*	*1.13*	*2.9*
V$EREF	*143*	*195*	*158.51*	*12.58*	*1.23*	*2.86*
V$E4FF	*92*	*99*	*73.95*	*8.6*	*1.34*	*2.86*

Table 4.2: Scores of motifs enriched in H/M peaks using motif enrichment analysis (MEA). Motifs with enrichment score z-score > 2.81 (p-value < 0.005) compared to random genomic background are selected.

A

EigenEpigenetics	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12
H3K27m					0.999							
H3K4m											-0.04	
Hoxa9			-0.999							-0.999		0.04
Pu.1						-0.999						
Stat5				0.999				0.999				
Cebpa		-0.999										0.999
P300			-0.04				0.04		0.999			
CBP		-0.04						0.04				
RNApII												
H3K4me1				0.04			0.999			-0.04		
H3acetyl	-0.04					-0.04			0.04		-0.999	
H4acetyl	-0.999				-0.04							

B

EigenMotifs	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
CEBP		-0.991		0.135								
HOXA.MEIS.PBX			-0.969			-0.078	0.961	0.276				
TALE			-0.022							-0.961		
HOXA										-0.276		
ETS	-0.961		-0.246			-0.982	0.276		0.981		-0.961	
bZIP		-0.079										
MYBL	-0.276			0.119	-0.276	-0.007			0.176		-0.276	
AML									0.079			
AP1R												0.964
CAUDAL												
MYC		0.08										
GFI1		0.017			0.961							-0.266
CREB		-0.07										
AP4R								-0.961				
PAX3												
AP2F												-0.007
STAT				0.984		-0.169						

Table 4.3: Canonical correlation analysis of epigenetics profiles and motifs enrichment. Sparse constrain was applied on the coefficients to obtain canonical variables (eigen-epigenetics or eigen-motif) by finding a linear combination of at most 30% of original number of variables. (A) Twelve canonical variables determined on the 12 epigenetic profiles of H/M peaks. (B) Twelve canonical variables determined on the 17 motifs similarity measures on the same set of H/M peaks.

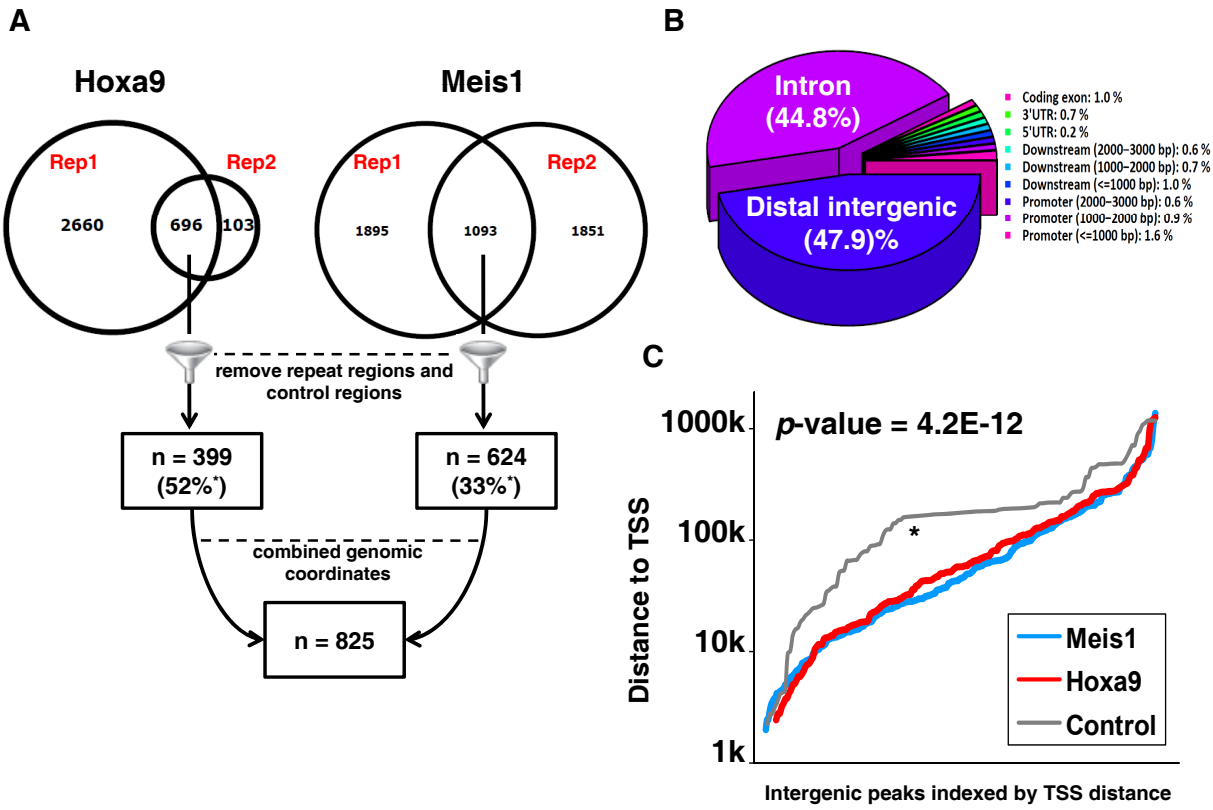


Figure 4.1: Genome-wide identification of Hoxa9 and Meis1 binding sites in leukemia cells (A) Schematic diagram of Hoxa9 and Meis1 binding site identification. Two replicate sequencing runs were performed for each factor and the enriched regions (or peaks) were selected only if they were detected in both biological replicates, consistent with ENCODE consortium standard. The peaks from both factors were subsequently merged into one set of peaks (n=825). Notably, a total of 52% of Hoxa9 peaks overlap with Meis1 peaks and 33% of Meis1 peaks overlap with Hoxa9 peaks. (B) Characterization of genomic localization of Hoxa9 and Meis1 binding sites. (C) Cumulative distribution of genomic localization indicates that Hoxa9 (red) and Meis1 (blue) binding sites are significantly (Kolmogorov-Smirnov test) closer to transcription start sites, compared with control peaks (gray).

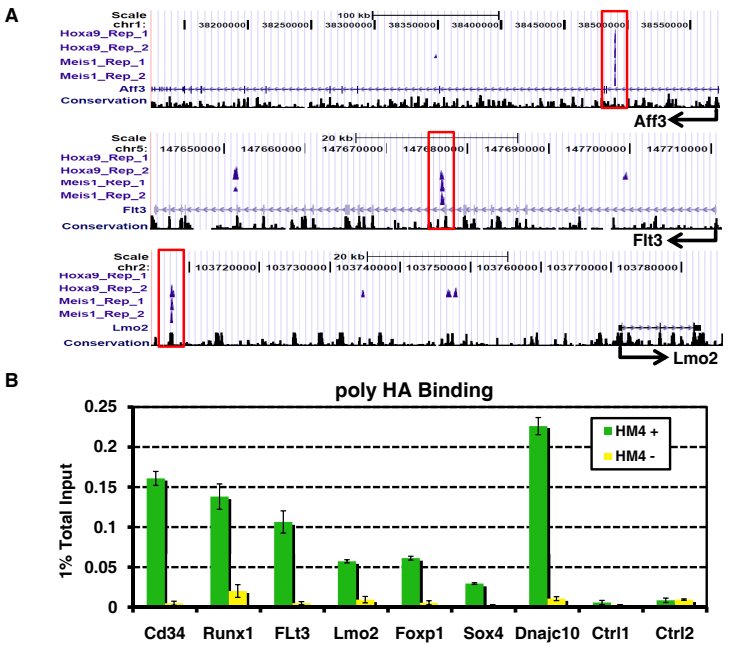


Figure 4.2: Validation of Hoxa9 and Meis1 binding sites identified by ChIP-seq (A) For each binding site enrichment profiles are shown for two replicates of Hoxa9 and Meis1 Chip-seq, with corresponding genomic annotation displayed as UCSC mm8 tracks at the Aff3, Flt3, and Lmo2 loci. A locus is deemed a high-confidence Hoxa9 and Meis1 binding site if it is bound by either Hoxa9 or Meis1 in both of the replicate sequencing runs. The sequence tags of non-significant peak regions (FDR p-value ¡ 0.01) are not displayed. The binding sites are highly conserved as shown by the Phastcon17 conservation track below. No significant binding was detected in the two control lanes at any of the regions shown. (B) Confirmation of selected hoxa9 and Meis1 binding sites by ChIP and Q-PCR. ChIP experiments were performed using polyclonal anti-HA antibodies on HA epitope-tagged Hoxa9-ER/Meis1-transformed myeloblastic cell (HM4) used for ChIP-seq experiments as described in Experimental Procedures. Green bars represent PCR signal as a percent of input for ChIP on cells cultured for 96 hours in the presence of 4-OHT, while yellow bars represent ratios for cells cultured for 96 hours in the absence of 4-OHT. These experiments show that Hoxa9 binds at high levels to ChIP-seq identified binding sites, but not at control peaks and that the Hoxa9 enrichment disappears upon 4-OHT withdrawal.

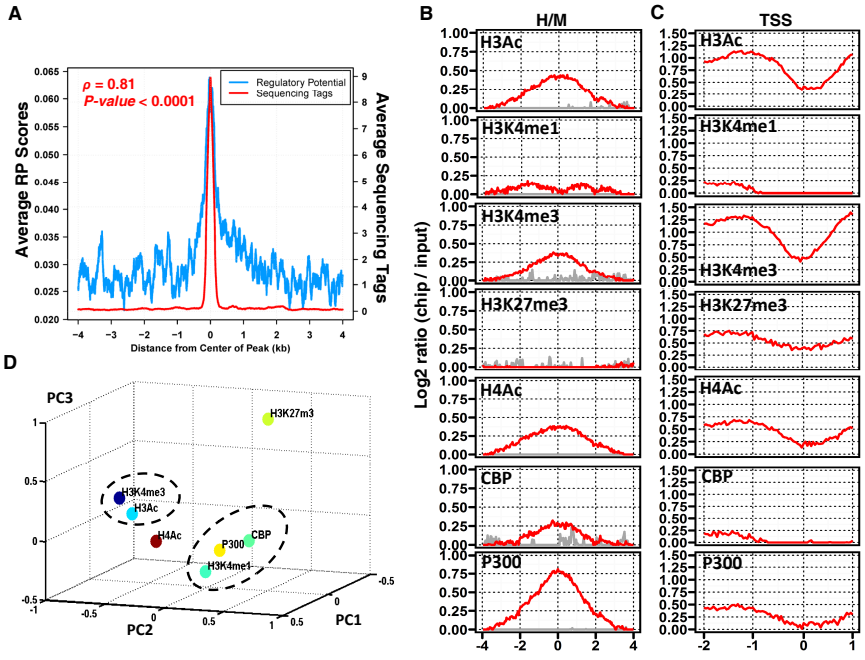


Figure 4.3: H/M binding sites show high regulatory potential and bear the epigenetic signature of enhancer sequences (A) Regulatory potential scores are high at the center of Hoxa9 and Meis1 binding with correlation of 0.81 (p-value ¡ 0.0001). The lines depict average of regulatory potential scores (blue) and sequencing reads in a 1kb region centered at Hoxa9 and Meis1 binding sites. (B) Spatial distribution of epigenetic modifications surrounding high-confidence Hoxa9 and Meis1 binding sites. Epigenetic modification status was examined in +/- 4kb regions centered on Hoxa9 and Meis1 binding loci using a custom Nimblegen tiling array. The normalized log2 ratios of a modification mark over input are shown relative to the center of the binding sites. (C) Spatial distribution of epigenetic modifications at the promoter region (+ 1kb upstream and - 1kb downstream) of a selected set of 360 genes that are closest to Hoxa9 and Meis1 binding sites. The normalized log2 ratio of modification mark over input are shown for each nucleotide with respect to their distance to the transcription start sites. (D) The 3-dimensional projection of seven epigenetic modification markers at the Hoxa9 and Meis1 binding sites using principal component analysis. The first three principal components account for 82.1% of the total variance. The figure shows the loadings of each epigenetic modification on these components. The p300, CBP, H3K4me1 epigenetic signature that is characteristic of enhancer sequences includes more than 65% of all Hoxa9 and Meis1 binding sites.

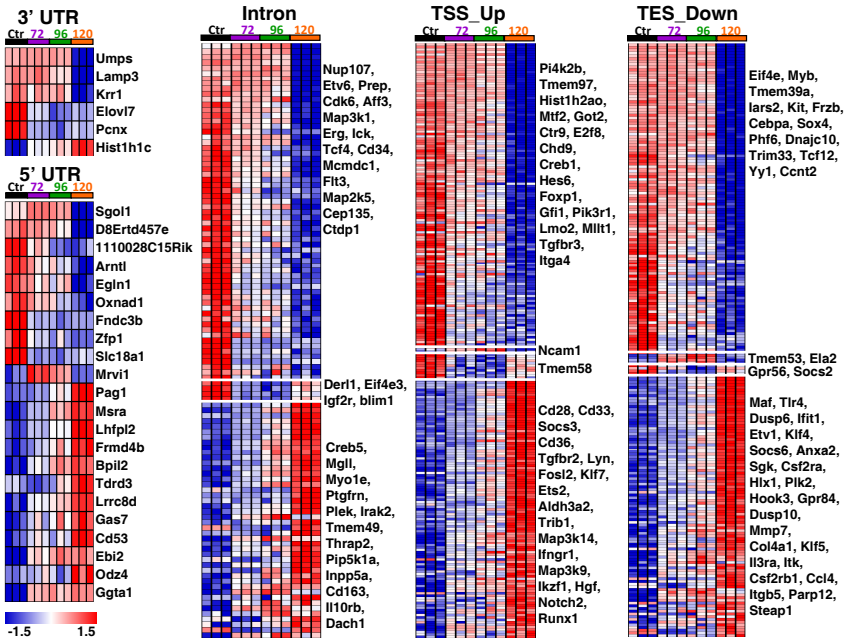


Figure 4.4: Heatmap showing temporal expression of Hoxa9 regulated genes most closely associated with H/M binding sites A subset of genes are shown that are significantly differentially expressed over 120 hrs period post 4-OHT withdrawal ($q < 0.001$ and median fold change > 3.0) compared to controls. Genes are organized according to the relative distance from their genomic features to the nearest H/M peaks based on RefSeq gene annotation. Sequences labeled as 5' UTRs are the regions between the transcription and coding start sites. Similarly, sequences labeled 3' UTRs are defined as the regions between the coding and transcription termination sites. Within each location category, genes are grouped based on their cluster designation (Figure S2) and listed in descending order in their differential statistics. Data are normalized across samples such that the expression value of each individual gene has zero mean and standard deviation of one.

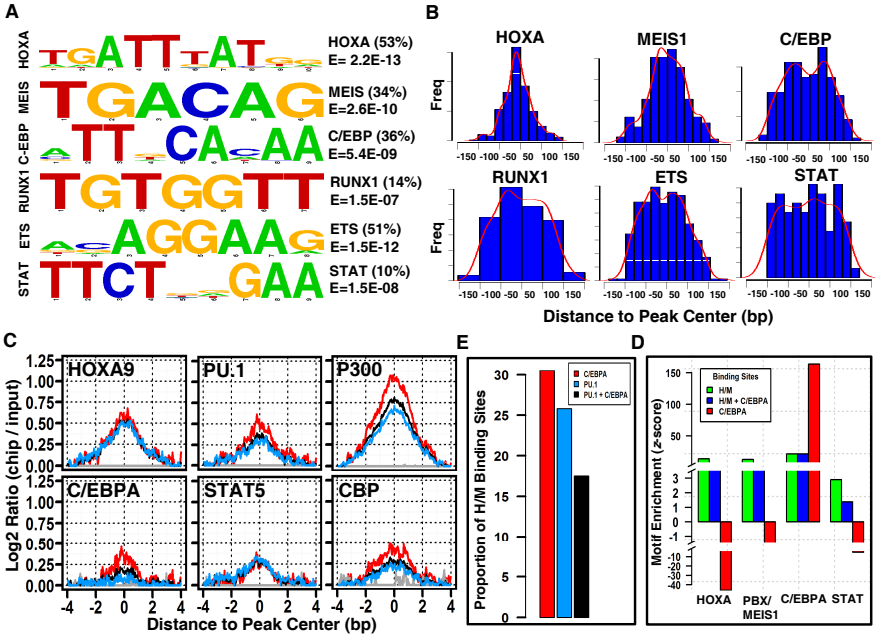


Figure 4.5: De novo motif discovery of transcription factor motifs in H/M binding sites and comparison to previously characterized macrophage enhancer sequences (A) Six de novo DNA sequence motifs and their STAMP logos (Mahony and Benos, 2007) and enrichment statistics, including observed frequencies and similarity measures. A complete list of de novo motifs (n=15) is given in Figure S3. (B) Spatial distributions of motifs listed in (A) with respect to centers of H/M binding sites. (C) Comparison of normalized ChIP-chip signal of H/M peaks that overlap with C/EBPA (red), PU.1 (blue), or both C/EBPA +PU.1 bound sequences. A total six transcription factors are shown in each panel. In most cases (except HOXA9 and STAT5), H/M peaks that are co-bound by C/EBPA showed highest expression intensity, followed by C/EBPA+PU.1 and PU.1. (D) A large proportion of H/M peaks were found to overlap with enhancer sequences bound by C/EBPA and/or PU.1 in LPS-stimulated macrophages (Heinz et al., 2010). E) Motif enrichment analysis showed a universally high level enrichment of HOXA, PBX/MEIS1 and STAT motifs in the set of enhancers described by Heinz et al. that are bound by Hoxa9 and Meis1 (green, blue) compared with those not associated with Hoxa9 and Meis1 (red).

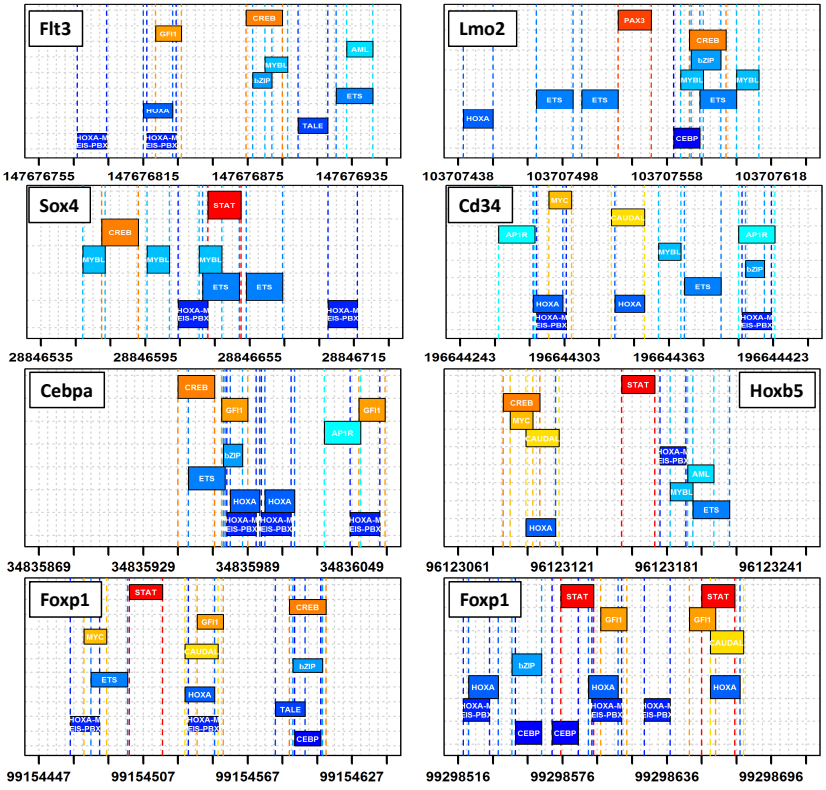


Figure 4.6: Examples of motifs enriched in Hoxa9-regulated Hoxasomes Enriched motifs, each depicted as a colored rectangular box are plotted for the central 200bp of eight representative Hoxasomes. Hoxasomes are highly enriched for HOX, HOX-MEIS-PBX, CREB, MYB, CAUDAL, ETS, MYC, and STAT sites, among others. All motifs shown are significantly enriched in Hoxasomes compared to random genome background. An enrichment statistic is computed with z-test comparing the observed frequencies (in H/M peaks) versus the expected frequencies (in random genomic background) (p-value < 0.001). A complete compendium of Hoxasome motifs is provided in Supplementary Data.

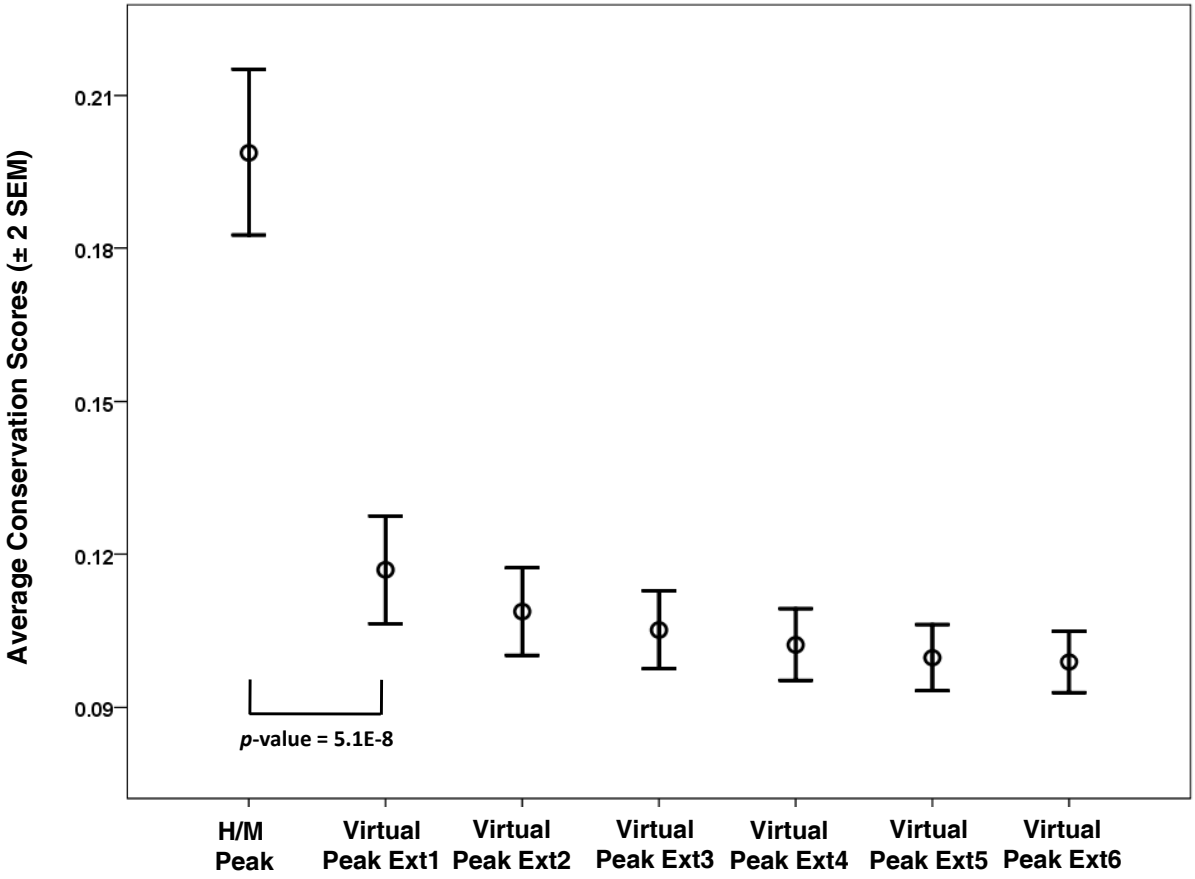


Figure S4.1: Comparison of conservation scores in H/M peaks and the extended regions outside of H/M peaks. For each extended region, an equal length of sequence segment was chosen successively outside of H/M sequences and their associated conservation was computed. Error bars are ± 2 SEM.

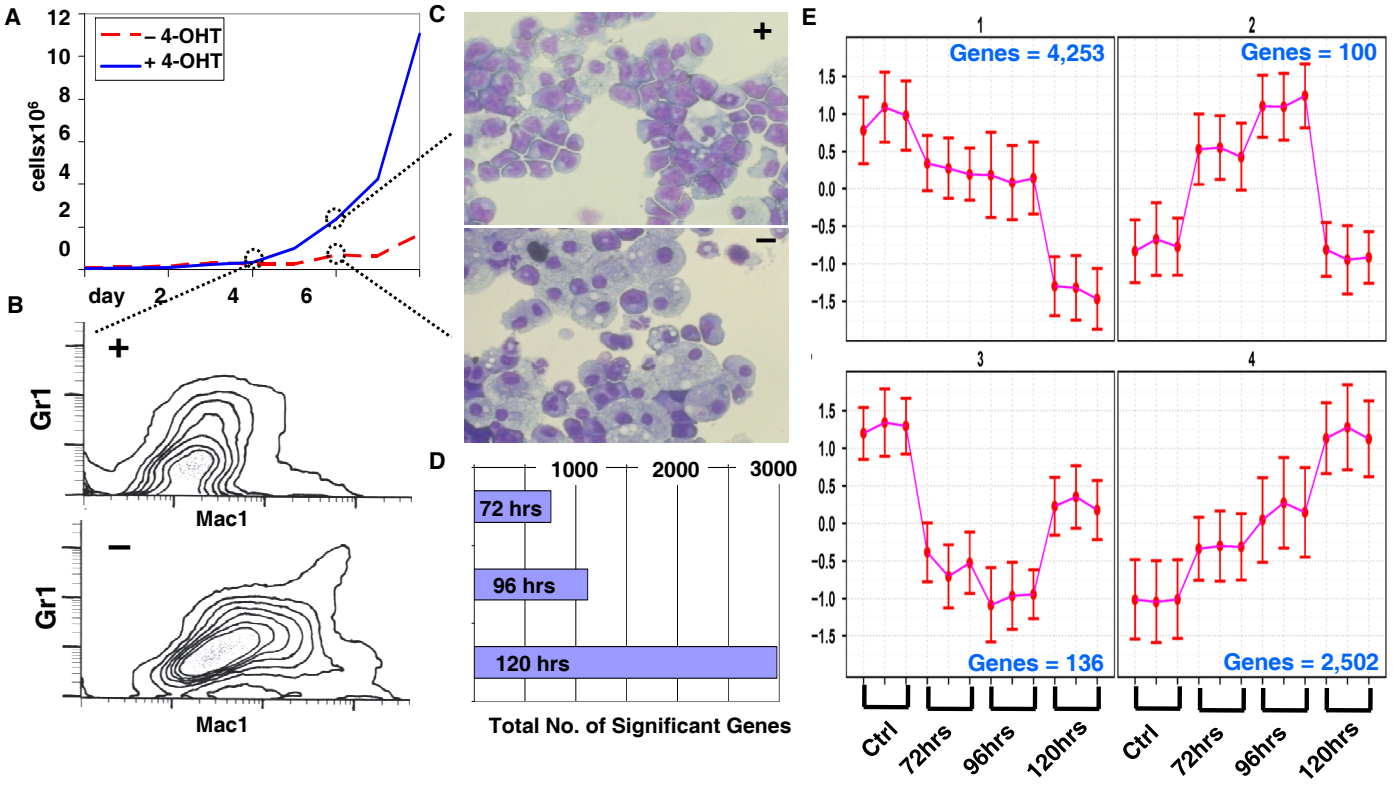
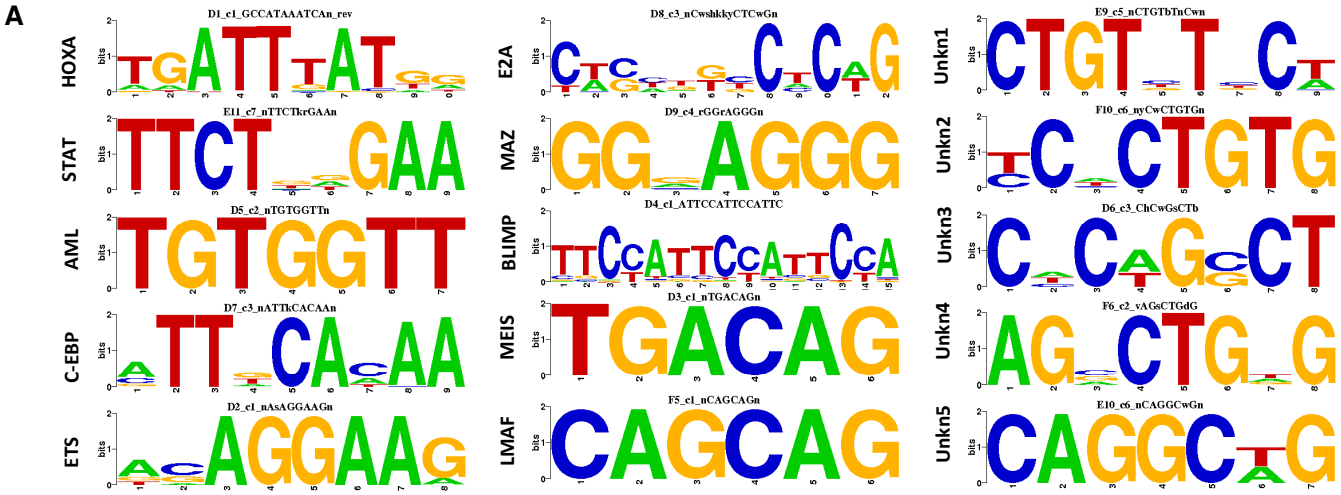


Figure S4.2: Conditional transformation by Hoxa9-ER and identification of Hoxa9 regulated target genes. (A) Hoxa9-ER cells cease dividing within 96 hours of 4-OHT withdrawal and show increased expression of the myeloid/monocytic differentiation markers Gr-1 and Mac-1 by flow cytometry (B). By day 6, the majority of cells showed macrophage morphology, while myeloblast morphology was maintained in cells that were cultured in continuous 4-OHT (C). (D) Cascade of significant changes in gene expression secondary effects following 4-OHT withdrawal. (E) Cluster centroids of significant genes in Hoxa9-ER profiling. Four clusters of genes with significant changes in expression level after 4-OHT withdrawal. Genes are clustered into one of the four clusters based on their temporal dynamics. For each cluster, the centroid and SEM are shown with the total number of genes in that cluster (blue).



	HOXA	STAT	AML	CEBP	ETS	E2A	unkn1	MAZ	BLIMP	MEIS1	unkn2	unkn3	unkn4	unkn5	LMAF
HOXA	425	47	58	110	218	73	76	162	131	124	50	79	71	65	102
STAT		80	9	21	34	16	10	21	24	34	11	8	3	8	13
AML			111	20	64	15	17	41	34	39	27	18	18	20	20
CEBP				204	94	36	31	50	60	70	19	32	24	18	43
ETS					404	70	63	145	157	139	45	72	63	64	97
E2A						148	32	76	50	44	25	32	34	33	38
unkn1							123	44	29	38	19	27	25	17	25
MAZ								274	118	80	40	71	60	67	63
BLIMP									250	77	28	45	37	36	47
MEIS1										264	29	44	36	47	61
unkn2											102	19	18	19	16
unkn3												138	89	50	39
unkn4													118	46	29
unkn5														119	31
LMAF															170

Figure S4.3: De novo motif discovery results. (A) STAMP logos of de novo identified motifs that are enriched in H/M peaks. (B) Co-occurrence matrix of de novo motifs. (C) Bubble plot of co-occurrence of de novo motifs. The size of the bubbles indicates the magnitude of co-localization between two motifs.

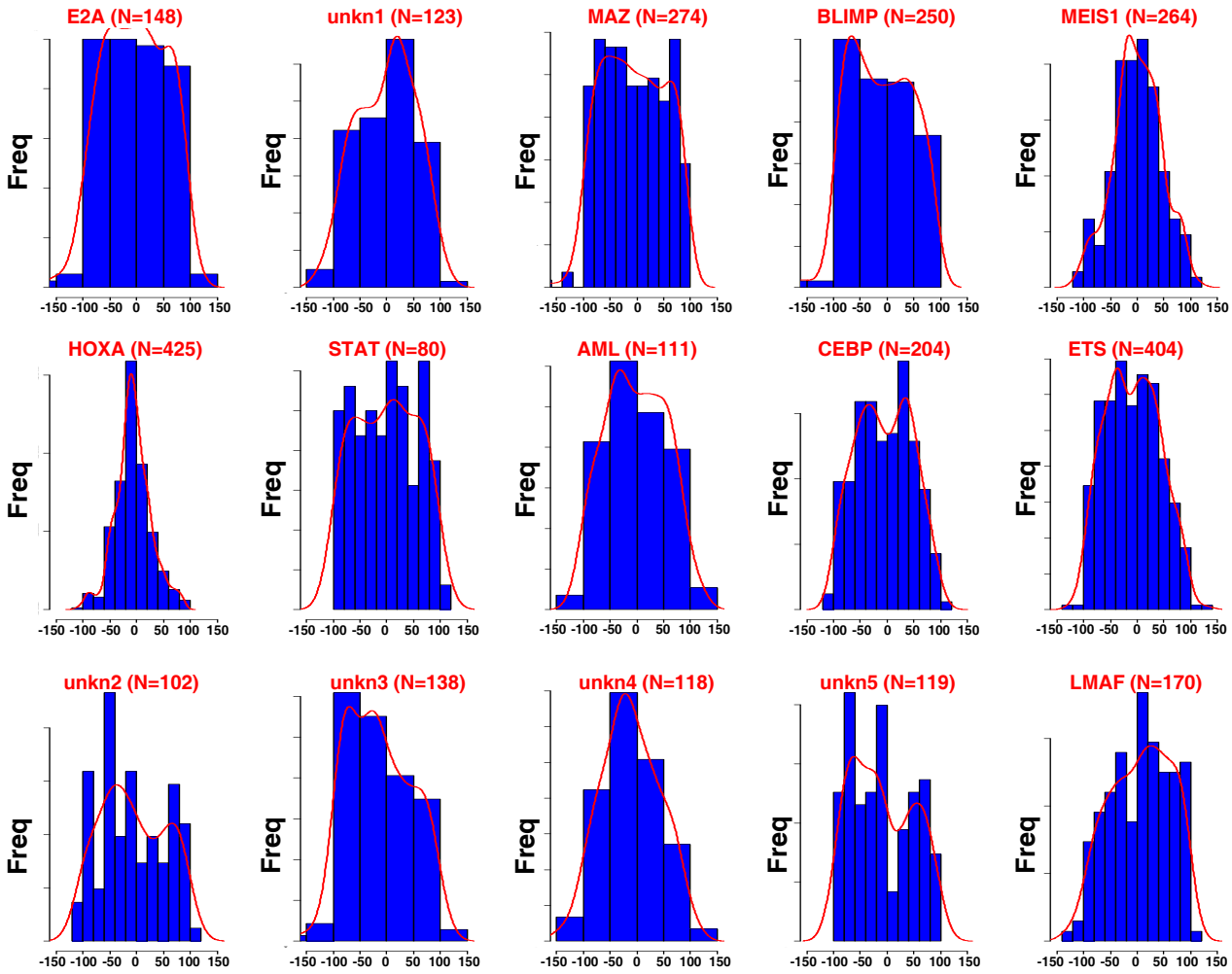


Figure S4.4: Spatial distribution of de novo identified motifs. The histogram and density of de novo motif spatial distribution was computed with respect to the center of the H/M peaks.

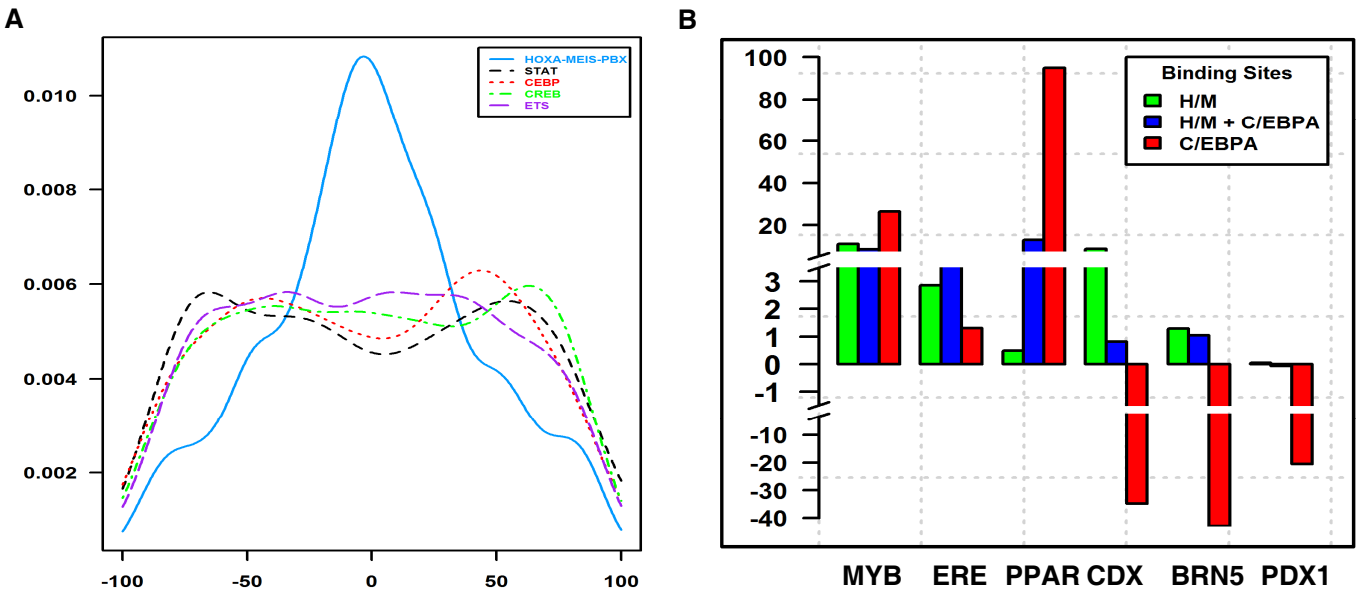


Figure S4.5: Spatial distribution patterns and enrichment statistics of motifs for Hoxa9 and Meis1 cobinding factors. (A) Motifs of Hoxa-Meis1-Pbx, STAT, C/EBPA, CREB, and ETS show complimentary spatial distribution patterns. (B) Comparison of motifs enriched in sequences bound by H/M, H/M + C/EBPA, and C/EBPA alone.

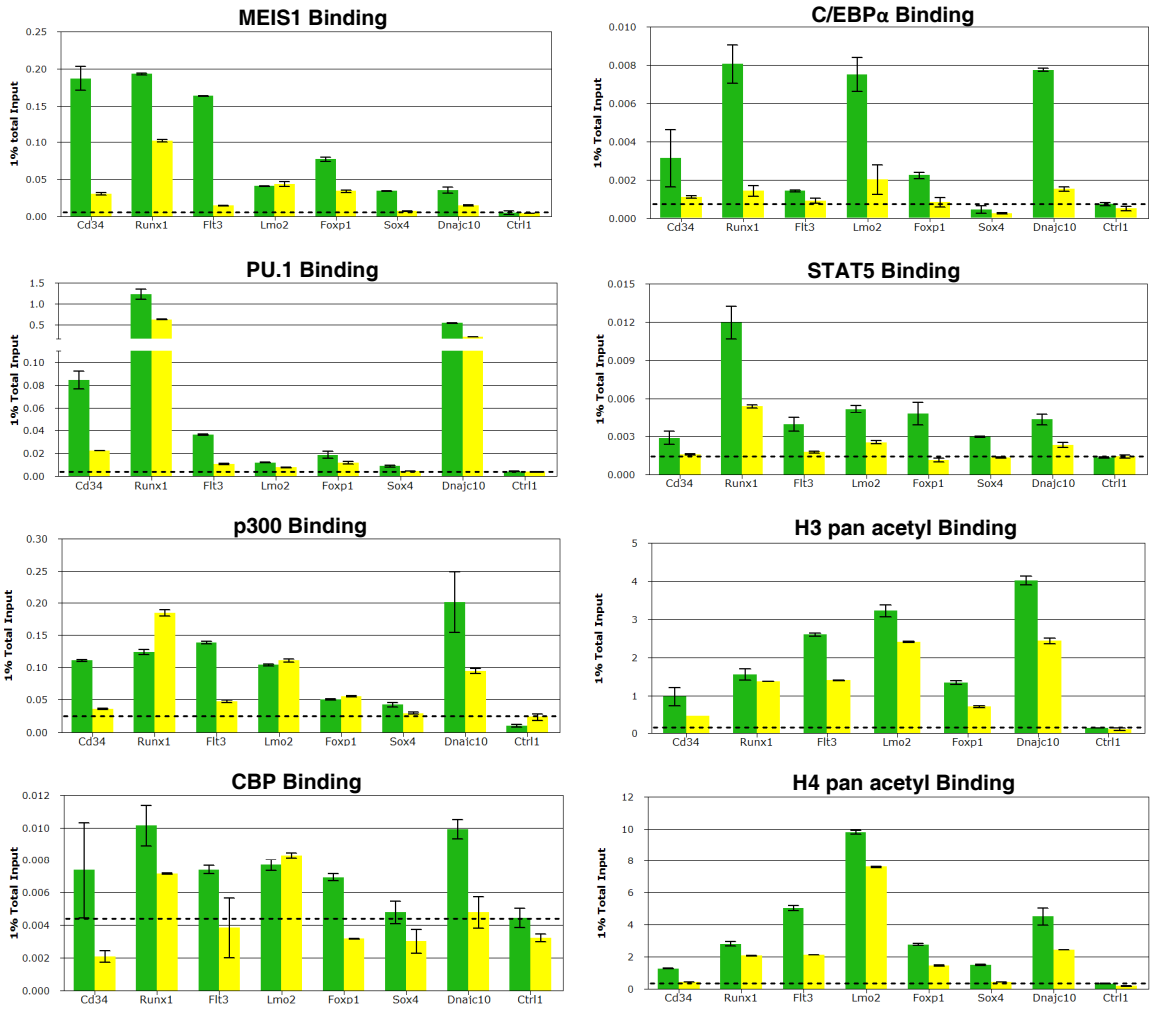


Figure S4.6: Hoxa9-mediated Hoxasome enhancer activity. ChIP experiments showing Meis1, Pu.1, C/ebpα, Stat5a/b, P300, CBP, histone H3 and histone H4 acetylation association with Hoxasomes is dependent on Hoxa9 as evidenced by drop in ChIP signal following 4-OHT withdrawal. See Experimental Procedures and Figure 4.2 legend. Experiments were performed and figures were prepared by Daniel Sanders at the Hess laboratory.

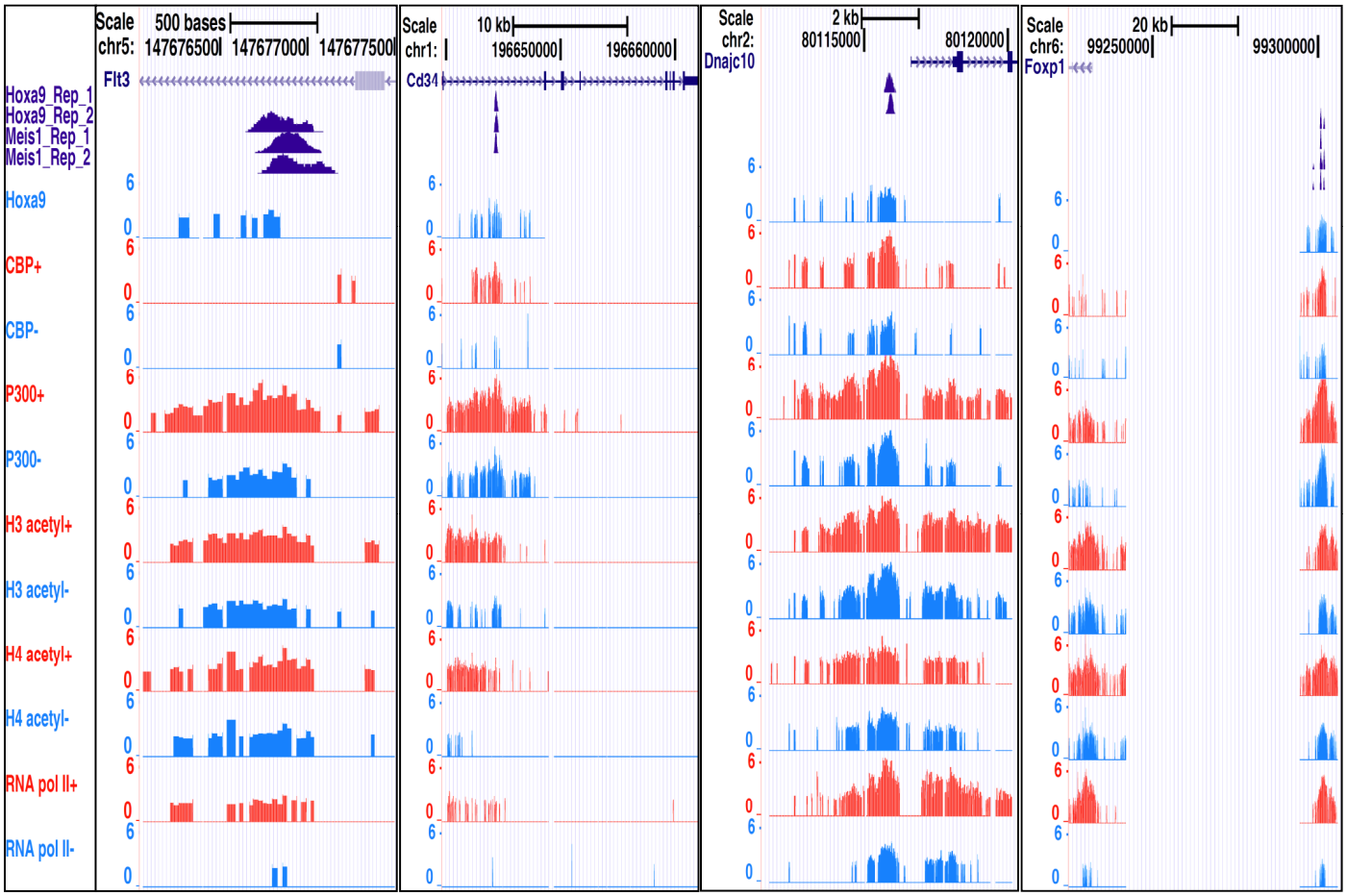


Figure S4.7: Examples of H/M binding sites carrying enhancer signatures. Four H/M bindings sites near or at promoter region of Dnajc10, Cd34, Foxp1, and Flt3 are shown with epigenetic signatures of CBP, H3 acetyl, H4 acetyl, P300, and RNA pol II. For each epigenetic mark, chip-chip results of 4-OHT (+) and 4-OHT withdrawal (–) are shown.

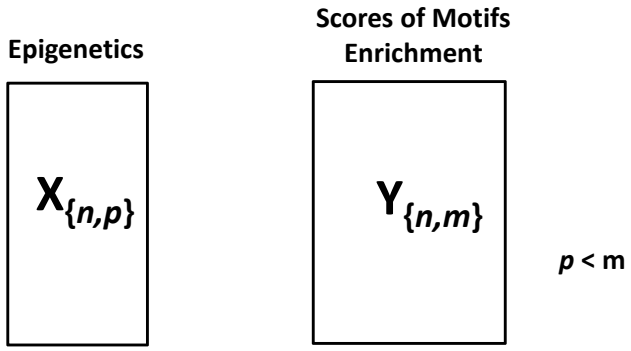


Figure S4.8: Schematic view of sparse canonical correlation analysis on motif enrichment and epigenetics profile at H/M binding sites. n represents the number of H/M peaks. m is the number of motifs found to be significantly enriched (p-value < 0.001) in H/M binding sites. p is the number of epigenetic modifications profiled using Nimblegen custom tiling array.

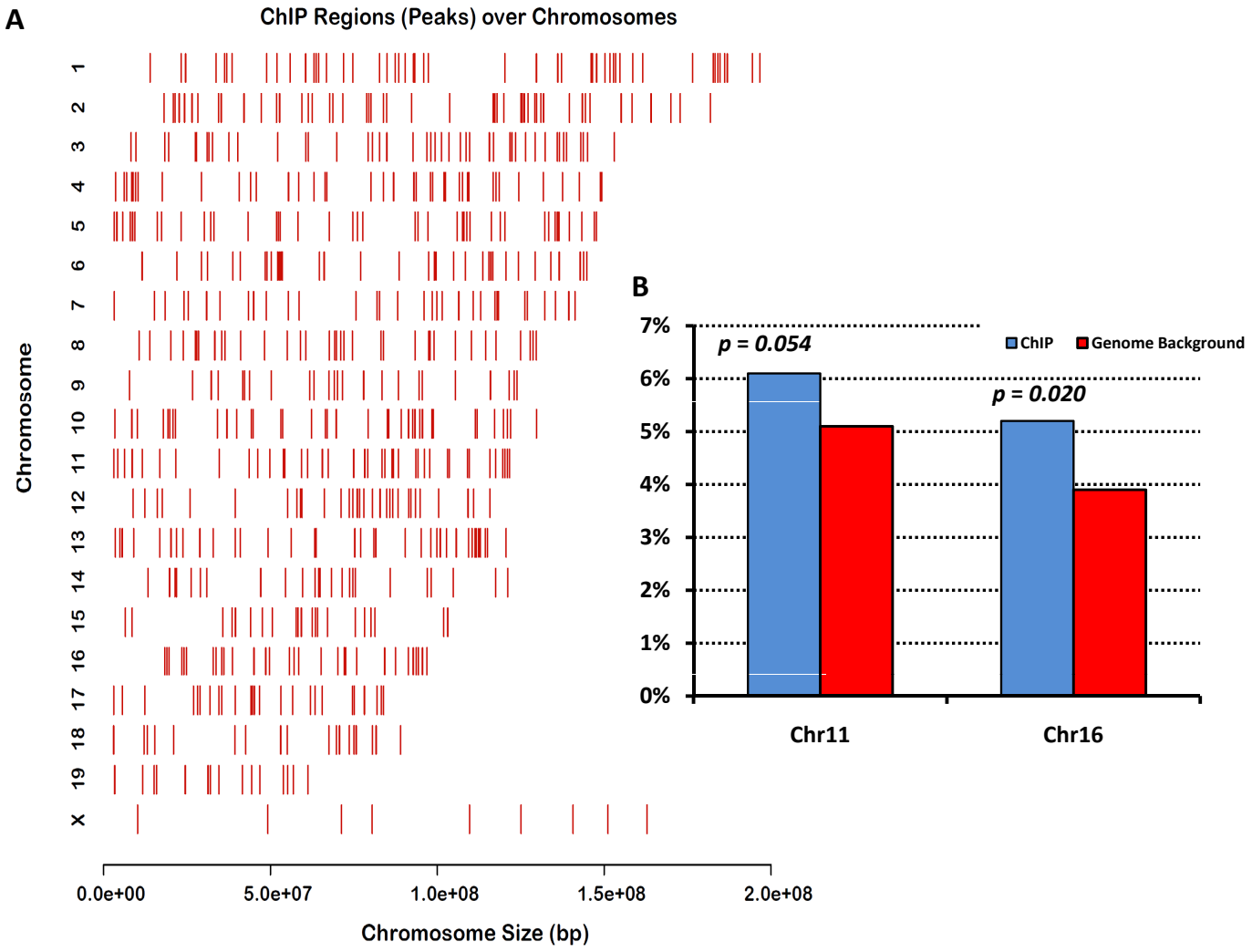


Figure S4.9: Chromosomal distribution pattern shows significant presence of H/M binding on chromosome 11 and 16 compared to random genomic background.

CHAPTER V

Spectral Analysis Of Temporal Gene Pathway Activation During Influenza Virus-induced Symptomatic Disease

5.1 INTRODUCTION

5.1.1 Motivating problem

Influenza viruses are highly infectious and can cause acute respiratory illness in human hosts. Infected hosts present a variety of clinical symptoms including fever, runny nose, sore throat, myalgias, and malaise with potentially more serious complications such as viral pneumonia (Cox and Subbarao, 1999). On the other hand, many hosts also withstand comparable level of viral insult with little or no overt symptoms, exhibiting a higher degree of tolerance (De Jong et al., 2006, Carrat et al., 2008). These subjects constitute a phenotype that is commonly known to be asymptomatic or subclinical infected hosts. Clearly, these asymptomatic infected hosts are able to control and eradicate viral threats more effectively compared to those who become symptomatic. Given the dynamic nature of viral infection, it is now recognized that interactions between hosts and viruses play a crucial role in determining the presence and absence of symptoms (Palese, 2004). This leads to an interesting question - what are the principal factors associated with such divergent disease outcome?

The influx of molecular and genetic evidences has shown that human diseases are a col-

lective consequence of perturbed gene expression through a complex and dynamic process. The infection by pathogens such as influenza viruses results in rapid and dramatic gene expression changes in host system. Hence, identifying and characterizing these changes help understand the mechanism and dynamics of the disease. The host peripheral blood contains key elements of the immune system and the circulating immune cells recruited by the host in response to viral infection and virus-induced tissue damage provides a global view of the host immune response. Thus, we hypothesized that it can be used to monitor the temporal dynamics of host-virus interactions. During a study in which we analyzed whole-genome gene expression profiles from healthy human subjects challenged with influenza H3N2/Wisconsin, we found that the expression levels vary dramatically over time for different genes. At various time points, different sets of genes were involved. A close look at the functions of these genes showed that they often belong to distinct biological pathway networks. As genes exert biological and cellular functions by interacting with others in a concerted manner, this observation suggests the existence of a temporal dynamic association between gene regulatory pathways and the disease development.

In this thesis our objective is to identify such associations by jointly modeling gene associations in a pathway and its pattern of temporal recruitment on the basis of measured gene expression profiles. We refer to such analysis as *temporal pathway analysis*. We propose a method that combines graph Laplacian clustering with flexible pathway significance measures that are derived from non-linearly embedded p-values. We present empirical results that show that our method is capable of identifying and partitioning pathways into coherent groups having temporally coordinated and biologically related activation patterns. The proposed method enables temporal gene pathway programs to be analyzed in a methodologically straightforward and statistically principled manner. To the best of our knowledge, this is the first temporal pathway analysis method that uses manifold embed-

ding of p-values and can be applied to high-dimensional and temporal gene expression data.

5.1.2 The etiology and physiological pathogenesis of the Influenza viruses

The Influenza viruses belong to the class of RNA viruses. They are generally considered as intracellular pathogens, spending most of their life time inside host cells. The viruses spread in their habitat population by entering hosts via aerosol or direct contacts. After successful entry, the viruses rapidly reproduce themselves within the host system by hijacking the host biological machinery to manufacture their own viral components. These components are assembled into new and fully-functioning viral particles, exit from infected host cells, and go into circulation to infect other hosts.

During this cycle, viral activities such as intensified virus reproduction will normally elicit strong immune response in human hosts. Typically, this involves two arms of host immune system, namely innate and adaptive immunity. The former mounts instant first-line attacks towards the virus while the latter functions by producing virus-specific antibody to recognize and bind to the invading pathogens for destruction. Although it was once believed that these two arms of immunity are separate systems, recent studies suggest that they are more intertwined, having no simple and clean separation between each other (Borghesi and Milcarek, 2007). In general, most antiviral strategies employed by host immune system involve shutting-down of protein synthesis, activation of apoptosis (cell death) of infected cells, and inflammation. These responses help reduce viral load within infected hosts, with a cost. The host often experience mild discomfort feelings or even severe systemic symptoms (e.g. fever) largely due to physiological damages as a result of host response to vicious viral activities (Unanue, 2007, Mizgerd, 2008).

Through the evolution, viruses and their hosts developed effective invading and defense

mechanism, respectively, to maximize their own chance of survival. One of the most successful strategy adopted by influenza viruses to evade host protection system is *antigenic variation*. The two types of antigenic variation, *antigenic drifting* and *antigenic shifting*, enable the produce of new strains of viruses bearing distinct structural characteristics that cannot be immunologically detected by host immune system (Cox and Subbarao, 1999, Palese, 2004). This further underscores the importance of gaining studying host-virus interactions through modeling the temporal patterns of gene pathway activities.

5.1.3 Related works on temporal differential gene expression analysis

A typical temporal gene expression experiment consists of a series of measurements of P genes in N samples over a time duration T. At each time slice t ($t \in T$), a random variable $X_t^{(i)}$ ($i = 1, 2, ..., P$) describes the expression profile of a gene i in all N samples or realizations. The temporal differential expression analysis (DEA) concerns with the identification of genes whose expression levels significantly altered at one or more time points when compared to a reference time point (normally the first sampling point or so-called baseline). Additionally, the samples may belong to multiple groups or conditions, e.g., disease versus healthy controls. In such case, the comparison is extended with one more dimension where group contrast needs to be considered.

Unlike analysis for a single time point, temporal DEA often faces the challenge of not having enough samples which in turn limits the available numbers of modeling techniques. Among the common choices, the simplest is the group of classical univariate tests such as F-test, t-test, one-way ANOVA, and SAM. These methods treat the temporal structure as unordered categorical labels and discard the temporal information associated with samples (Storey et al., 2005). By fitting a mixed model, they test the differences in fixed effects over time and/or between groups. Notably, these methods usually impose a strong condition on

the independence structure in the errors. More recent works attempt to jointly model all genes simultaneously and incorporate time structure into the model with spline fitting (Bar-Joseph et al., 2003, Storey et al., 2005). In essence, these methods assume certain stationary properties in the errors which is directly accounted by the smoothing curve fitting. In addition, non-parametric methods have also been proposed (Phang et al., 2003, Tusher et al., 2001). It is worth noting that none of these methods directly incorporates existing gene pathway information and treats each gene as an independent variable.

5.1.4 Related works on incorporating geneset structure

For nearly all gene expression analyses, successfully identifying genes having significant expression changes under different conditions fulfills only a bare minimum task. In most cases except classification problems, a list of significant genes falls short to address the key question of how these genes are related to biological phenomena or medical conditions. On this front, pathway-based analysis has proven to be useful in providing insights to the mechanistic roles played by the candidate genes (Subramanian et al., 2005). Briefly, a gene pathway refers to a group of genes and their translated protein products that are functionally related. Collectively, they influence many biological processes in a synergistic or antagonistic manner. A change in expression of one gene often results in a cascade of changes in the expression of other genes in the same pathway. From a statistical perspective, these genes can be considered as correlates. The realizations (the observed gene expression values in multiple samples) are not completely independent and thus comprise special correlation structure. In principal, such structured information shall be appropriately considered when the whole expression profiles of thousands genes are analyzed. Doing so will also improve the interpretability of discovered candidate genes and facilitate new hypothesis generating.

Generally speaking, there are two types of gene expression analysis based on pathway information. The first one is a two-step procedure in which a univariate type of significance test is performed on a individual gene basis. This is then followed by a hypothesis testing procedure similar to the *Fisher exact test* where a list of all significant genes are compared with known pathway definitions. If a significantly large proportion of them are found to coincide with members of that pathway, this pathway is considered to be enriched or over-represented. Some of the well-known methods in this category includes Gene Set Enrichment Analysis (GSEA) and GO analysis in DAVID (Subramanian et al., 2005, Dennis et al., 2003, Huang da et al., 2009). A second category of pathway analysis methods include the geneset or pathway structure into the model and individual pathways as an independent entirety. These methods aggregate the overall effects by all genes and derive a common statistic for the whole pathway. The most well-known methods in this category include the Gene Set Analysis (GSA), Globaltest, as well as GSEA (Efron, 2007, Goeman et al., 2004, Subramanian et al., 2005).

5.2 METHODS

In this section, we describe the proposed method for assessing temporal activity of gene pathways. We combine the two aforementioned approaches into a single analysis scheme that considers temporal gene expression as well as the incorporation of gene pathway structure.

5.2.1 Measuring significance of functional pathway

Assuming that raw gene expression profiling data have been appropriately preprocessed and checked for quality of both input RNA and image scanning, we first perform a pre-screening of genes to be analyzed in order to reduce both the noise within and the dimension of the dataset. This is analogous to the variance filtering approach commonly

used in static gene expression analysis (Gentleman et al., 2005). Although such screening can be performed on a discrete time point basis, we choose EDGE for its smoothing curve fitting over adjacent time points with increased robustness (Storey et al., 2005). Compared with a typical differential analysis, this pre-screening relaxes selection criteria and ensures the inclusion of genes that would have been excluded in a regular differential analysis.

We next seek to measure the magnitude of significance in differential expression of all genes in a pathway. For each sampling point, we apply the generalized linear model approach suggested by Goeman et al., 2004

$$E(Y_i|\beta) = h^{-1}(\alpha + \sum_{j=1}^{m} X_{ij}\beta_j) \tag{5.1}$$

where the coefficient vector β quantifies the weights of expression values X_{ij} for each subject i on gene j in a pathway with m component genes. The response variable Y is either a discrete class label or continuous measurement. It determines the proper form of h, the *link function* (McCullagh and Nelder, 1989). In our study, $Y \sim Bernoulli$ and takes on -1 for asymptomatic and 1 for symptomatic phenotype, resulting in the canonical *logit* link function. Assuming $\beta \sim \text{Distr}(0, \tau^2 I_m)$, the testing for the null hypothesis $H_0 : \beta_j = 0$ is simplified to a score test for testing $H_0 : \tau^2 = 0$. This score test effectively maximizes the average power of testing all alternatives (Goeman et al., 2004). The significance of the test statistic is calculated by either approximating a scaled χ^2 distribution or using a permutation-based method for small sample size.

5.2.2 Inverse logistic transformation of test statistic of significance measure

With the activity of a pathway properly measured and summarized for all its component genes at individual time points, the temporal activities of that pathway can be represented with a vector of summarized statistics of all time points. The relative magnitude of a test statistic reflects the strength of activation of the pathway at a given time point. We

choose p-values but any other alternative types of measures will likely provide similar results. The challenge is that raw statistics such as p-values have rather poor contrast in their lower range (Figure 5.1 left panel). This raises difficulty when comparing statistics across different pathways as more significant p-values dominate those relatively less significant ones, rendering regular test methods less sensitive to smaller changes. Intuitively, from a probability point of view, we can consider that all p-values are from a common distribution and the most significant statistics are located on very far tail end of that distribution. It then becomes more difficult to discern between such small p-values as the lack of contrast will cause the masking of differences among pathways with relatively small magnitude of changes. Yet these differences are of importance to the fundamental biological processes especially in a temporal setting.

To resolve this issue, we propose a continuous transformation using inverse logistic function to map original p-values to a finer scale with higher resolution:

$$p^* = -\frac{\log(p)}{c - \log(p)} \tag{5.2}$$

Here the constant c is a free tuning parameter and it controls the degree of smoothing by the transformation. It is not clear how to choose the optimal value for c. Based on our simulations, we recommend choosing a value that is close in magnitude to the observed statistics. For instance, in Figure 5.1 (right panel), the median of the 30,000 p-values is on the order of 10^{-1}. We therefore use the value of $-\log_{10}(10^{-1})$ for c.

As shown in Figure 5.1, the inverse logistic transformation effectively recovers p-value contrast at the lower end of the spectrum and achieves satisfactory transformation results (Figure 5.1 right panel). It outperforms other types of transformation such as binary thresholding transformation (Figure 5.1 middle panel). Furthermore, this transformation has a nice property in that it constrains the transformed values to be in the same range as the

original p-values, namely $(0,1)$.

5.2.3 Formulating temporal correlation as a graph partitioning problem

Using the transformation in 5.2.2, we represent the temporal expression activity of a pathway i with $V_i = (V_{i1}, V_{i2}, ..., V_{ij})$ representing the significance measure of pathway i over a series of time points indexed by j. Denoting this collection of $\{V_i\}_{i=1,...,N}$ as V for all N pathways, we seek to group pathways V_i's according to their temporal patterns such that the pathways in the same group share similar temporal expression trajectories. This can be readily converted into a well-known graph partitioning problem for which a nice solution is given by spectral decomposition of the normalized graph Laplacian (Hastie et al., 2001, Xu et al., 2009).

We start by assuming that all V_i are somewhat correlated, i.e., all pathways has some level of similarities in their temporal activities. Let $G = (V, E, W)$ be an undirected graph with a set of vertices V representing the pathways as above, and E being the set of edges between each pair of vertices. The weighted *adjacency matrix* W is the matrix $W = (W_{ij})_{i,j=1,...,N}$ representing the weights of edges. Intuitively, a weight W_{ij} corresponds to the level of temporal similarity in activities between two pathways (i, j). We use the Gaussian kernel function based similarity measure $W_{ij} = \exp(-d_{ij}^2/\alpha)$ where $\alpha > 0$ is the scale parameter for inverse kernel width and d_{ij} is the Euclidean distance between V_i and V_j. The *normalized graph Laplacian* is defined as

$$L = I - D^{-1/2} W D^{-1/2} \tag{5.3}$$

where D is the degree matrix defined by

$$D = \text{diag}(\sum_j w_{i,j}) \tag{5.4}$$

The solution for partitioning these N pathways into K groups is given by applying the K-

means clustering technique on the m eigenvectors corresponding to the m smallest eigenvalues of the normalized graph Laplacian L.

As with most applications based on K-means clustering, the choice of appropriate K is not obvious. The most common practice is to visually examine the decaying rate the eigenvalues as suggested by Xu et al., 2009 In our particular application where temporal recruitment of biological pathways are studied, we reason that the interaction of gene pathways should increase at an exponential rate over time with substantially more pathways getting involved. Accordingly, we suggest to choose K based on

$$K = \log_2(N \cdot T) \tag{5.5}$$

where N is the total number of pathways and T is the total number of assayed time points. Notably, this free parameter selection criterion is in spirit similar to the minimal description length (MDL) principle widely used for data compression and model selection in information theory Rissanen, 2007, Grunwald, 2007. Combined with aforementioned heuristic approach, this delivers satisfactory results in our analysis.

We summarize the algorithm as the following:

Algorithm 1: Temporal Analysis of Gene Pathway

Input: P-values $V_i \in R \times T$, $i = 1,2,...,N$ and $0 \leq V_i \leq 1$
Result: Partitions of pathways
begin
1. Inverse logistic transformation: $\tilde{V}_i = -\frac{\log(V_i)}{c - \log(V_i)}$;
2. **for** $(i, j) \in (1,2,...,N)$ **do**
   - Compute $d_{ij} = \sqrt{\tilde{V}_i^T \tilde{V}_j}$;
   - Compute $W_{ij} = \exp(-d_{ij}^2/\alpha)$;
3. Calculate D, where $D = \text{diag}(\sum_j W_{i,j})$;
4. Calculate L, where $L = I - D^{-1/2} W D^{-1/2}$;
5. Specify or automatically choose $K = \log_2(N \cdot T)$;
6. Cluster K eigenvectors corresponding to the K smallest eigenvalues of L

5.3 RESULTS

5.3.1 Alcohol affected human biological pathways

As a proof of concept, we first analyzed a public dataset which studies the effects of alcohol consumption using four different beverages, including water, juice, alcohol, and red wine (Baty et al., 2006; Material and Methods Section 5.5). We focus on comparing the effects on gene expression by alcohol against water. A total of 54 samples were collected at 5 different time points after subjects drank alcohol-based beverage and after they consumed only water, separately. As described in section 5.2.1, the level of differences in pathway activities by the two groups of subjects are computed and their significance measures (p-values) are transformed using inverse logistic function.

Because of the rather controlled amount of alcohol consumption, only mild pathway activities were observed. In addition, this study has a relatively small sample size with 4 to 5 samples available at one time point. Since it normally would require at least 5 samples in each group to obtain a significance measure at the level of 0.05 based on permutation, the power of detection in this case is rather restricted. Nonetheless, the temporal clustering revealed some interesting findings that were not identified in the original analysis employing a method based on correspondence analysis with instrumental variables (Baty et al., 2006), accounting for individual subject effects with a mixed effects model-like multivariate analysis technique. For example, Figure 5.2 shows that a group of pathways related to the proper function of heart (cardiac) activities are found to be most significantly regulated. Genes from *heart failure ventricle* related pathway are most down-regulated in subjects consumed alcohol based drinks. Furthermore, these pathways tend to cluster together with pathways that are most related to stress response, such as inflammation and cell apoptosis. More important, the results revealed some pathways that are very relevant to the function of alcohol but would have not been identified as significant using other methods.

For instance, most genes on *heart failure ventricle* pathway are slighted down regulated in subject who intook water-based beverage. However, these genes are associated with rather less significant p-values (green color bar in the small subpanel in Figure 5.3). Normally, these genes will be deemed as insignificant in a univariate test as the significance levels of the differential expression changes are above a typical p-value ≤ 0.05 threshold. With our approach, the entire pathway including those weak genes was discovered. Given the known effect on heart and blood circulation by alcohol, the result appears to offer plausible biological interpretation. Other similar examples of dysregulated pathways include the ones that are related to the function of *retina* or *retinal cells*. Again, this does seem to be reasonable given that alcohol's effect on a person's vision. Taken together, these results are encouraging and assure the effectives and power of our method.

5.3.2 Temporal host response networks during influenza infection

We next applied our method on a viral challenge study that investigates the host immune response towards influenza viral infection (Zaas et al., 2009). A total of 17 subjects were challenged with influenza H3N2 viruses. In the end, 9 subjects developed mild to severe symptoms and were clinically labeled as symptomatic subjects. The rest 8 subjects were not clinically infected and thus classified as asymptomatic (Material and Methods Section 5.5). We compare the gene expression profiles of symptomatic against that of asymptomatic subjects. Our objective is to identify gene pathways that are involved in host-virus interactions. Unlike the previous example of beverage study, the acute nature of the viral infection prove to be more dramatic and caused much more drastic changes in host gene expression program. At the peak symptom time, roughly +60 hours post inoculation (hpi) time which is 0hpi, as many as 80 percent of pathways are activated. The same procedure outlined above was performed to measure the significance of associations

between genes/pathways with subject phenotypes.

We show in Figure 5.4 that the inverse logistic function is able to map the original p-values onto a much more refined resolution scale.

A close examination of the eigen-spectrum of the normalized graph Laplacian, we selected a total number of $K = 18$ pathway clusters. This is slightly larger than the $K = 15$ solution according to the $K = \log_2(N \cdot T)$ where $N \cong 1900$ and $T = 15$.

The results are visualized in Figure 5.5 where each symbol representing a pathway and each cluster is in a different color and shape. These 18 clusters form rather distinguishable communities. On one hand, some of these clusters such as cluster 3 (blue) interact with many others and likely to serve as the focal point of the whole host immune response. On the other hand, pathways in cluster 1 and 2 are relatively more separated from the rest. These pathways turn out to be more transiently activated only during a certain period of time. The pathways that belong to the same region or cluster are highly related with each other in terms of their biological functions. For instance, cluster 3 (blue) includes pathways that are most directly related to host immune responses and infectious diseases such as toll-like receptor (TLR) signaling, oxidative stress, and cell apoptosis (Cook et al., 2004, Clarke and Tyler, 2009, Takaoka et al., 2005, Kawai and Akira, 2007). These pathways are known to be the ones that are immediately activated upon the invasion of viral pathogens. In particular, TLR signaling detects the existence of viral components of pathogens and senses internal physiological changes (such as those induced by oxidation). One of the direct consequences of the TLR pathway activation is the triggering of programmed cell death (apoptosis), a process that serves to reduce viral load in the system by eradicating infected cells. This is assuring to observe such cohesive clustering of pathways that are highly relevant in the biological functions that they are involved.

Perhaps most interestingly, these clusters exhibit distinct but subtle temporal dynam-

ics that would have been hard to detect if other methods were used. Figure 5.6 shows the temporal expression trajectory of pathways in clusters 3, 4, and 5. We see that these clusters start being heavily perturbed around roughly same time (+36hpi). However, cluster 3 pathways continued its activation intensity till +69hpi and maintained the that high level of activation till +89hpi. Towards the end three time points, they slightly fall back from their peak level of activation but still remain at a high level. Between +36hpi and +89hpi, cluster 3 pathways are extremely consistent in their expression as shown by the small inter-quantile range (IQR) (Figure 5.6). Likewise, cluster 5 pathways show pattern similar to that of cluster 3. However, there is a relatively different fluctuation pattern in their activation levels at the last three time points. First of all, their fluctuation are more consistent (narrower IQR). Secondly, their activities climb back to higher level at the end, contrasting the lower level in cluster 3. Different from the previous two clusters, cluster 4 showed modest activation magnitude compared to cluster 3 and 5. Its pathways reach their highest level of activation at around +53hpi and stayed at that level till the end of the study. These results demonstrate the effectiveness of the method in identifying the subtle differences in pathway activities.

5.4 CONCLUSION

Graph Laplacian clustering and differential expression analysis with *a priori* pathway structure are not new ideas. Our contribution lies in the fact that we combine the strength of the two approaches and extend it with the embedding and soft-thresholding/smoothing of pathway significance measures. This simple yet effective strategy allows the analysis of temporal dynamics of gene pathway activities in a temporal setting. We also supply analytical solutions for choosing proper values of free parameters without resorting more computationally expensive methods such as resampling techniques. This circumvents the

relatively arbitrary selecting criteria widely used in a K-means analysis. In addition, our choice on directly modeling the significance measure (p-values) goes beyond exploratory analysis. As shown in the Influenza study, the actual p-values is derived from a hypothesis testing on pathway activity and phenotype association via logistic model fitting. The results therefore bear important usefulness if a classifier were to be built.

When applied on public dataset and the dataset from our study of Influenza viral infection, the proposed method demonstrates that it can identify and organize pathways into different groups based on their temporal expression patterns. The partitioning of the pathways relate them to the functional demarcation of the host biological responses. By testing a whole pathway as an entire unity, the present method is capable of identifying pathways that are only weakly modulated yet nonetheless biologically relevant to the disease under study. This provides augmented statistical power in identifying disease biomarkers and aids new biological hypothesis generating.

We also point out that the proposed method is rather flexible. For instance, the mixed effects model adopted here is not the only way to test the conditional differences in pathway activities. Other approaches using different statistical models can also be applied as long as they provide quantitative statistic pertaining to the hypothesis being tested, i.e., whether a pathway relates to the disease outcome. In fact, correlation measures between the gene expression and a disease outcome have been shown to deliver satisfactory results as well (Horvath et al., 2006).

Although this work focuses on the gene expression data, the overall methodology will extend naturally to a similar time-series analysis setting with other types of high-throughput biomedical data collections. For instance, the next-generation sequencing technology (RNA-seq) allows the expression to be measured at a single nucleotide (nt) level, contrasting the 16-25nt probe sequence on a conventional mRNA gene expression

platform. Although the RNA-seq measurement units are discrete rather than continuous as in the case of gene expression, the overall analysis scheme outlined here should largely remain to be the same with only slight adjustment to be made to accommodate the difference in measurements.

5.5 MATERIALS AND METHODS

Beverage study

The wine study was originally conducted by Baty et al., 2006 to investigate the effect of beverage intake on peripheral blood gene expression. Briefly, six healthy volunteers were randomized to consume four different beverages (500 mL each) including grape juice, red wine, 40 gram diluted ethanol, and water. Every volunteer drinks one of the four beverages on four independent days. Blood samples were taken at baseline time, 1, 2, 4, 12 hours after the drink together with standardized nutrition. In total, 108 PBL samples were obtained for RNA extraction and hybridized on standard Affymetrix microarrays (Zaas et al., 2009).

In our analysis, we selected 54 samples that were collected after drinking alcohol and water (highlighted in red in Figure 5.7). The raw affymetrix gene expression dataset was obtained from the Gene Expression Omnibus (GEO) under accession number (GSE3846).

Influenza challenge study

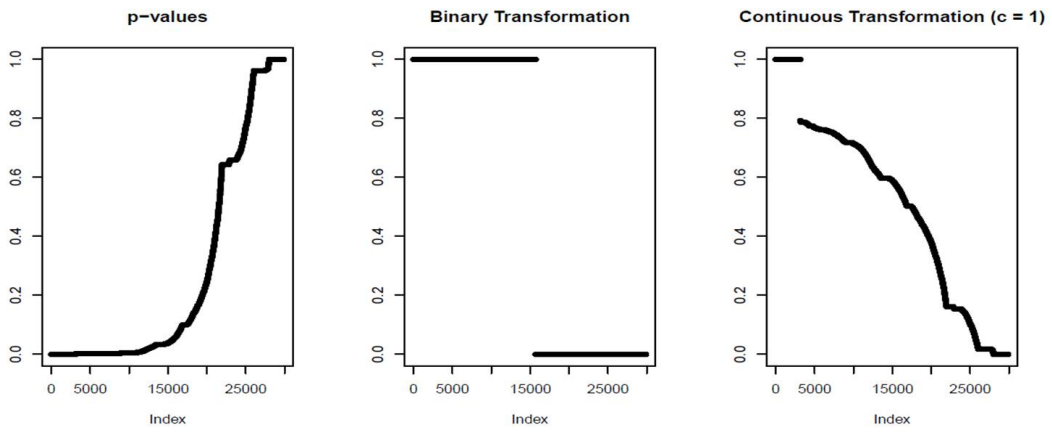
Challenge study and case definition A healthy volunteer intranasal challenge with influenza A A/Wisconsin/67/2005 (H3N2) was performed at Retroscreen Virology, LTD (Brentwood, UK) in 17 pre-screened volunteers who provided informed consent. All volunteers were influenza A antibody negative at pre-inoculation testing. Symptoms were recorded twice daily using standardized symptom scoring (1). The modified Jackson Score requires subjects to rank symptoms of upper respiratory infection (stuffy nose, scratchy

throat, headache, cough, etc) on a scale of 0-3 of "no symptoms", "just noticeable", "bothersome but can still do activities" and "bothersome and cannot do daily activities". For all cohorts, modified Jackson scores were tabulated to determine if subjects became symptomatic from the respiratory viral challenge. A modified Jackson score of ≥ 6 over the quarantine period was the primary indicator of successful viral infection (Zaas et al., 2009) and subjects with this score were denoted as "SYMPTOMATIC". Subjects were classified as "ASYMPTOMATIC" if the Jackson score was less than 6 over the five days of observation and viral shedding was not documented after the first 24 hours subsequent to inoculation. Standardized symptom scores tabulated at the end of each study to determine attack rate and time of maximal symptoms (time "T").

Gene expression sample collections and processing During challenge study, subjects had their peripheral blood samples taken 24 hours prior to inoculation with virus (baseline), then at 8 hour intervals for the initial 120 hours and then 24 hours for the remaining 2 days of the study. Samples were aliquoted and frozen at -80?C immediately. RNA was extracted at Expression Analysis (Durham, NC) from whole blood using the PAXgene™ 96 Blood RNA Kit (PreAnalytiX, Valencia, CA) employing the manufacturer's recommended protocol. Hybridization and microarray data collection was performed at Expression Analysis (Durham, NC) using the GeneChip? Human Genome U133A 2.0 Array (Affymetrix, Santa Clara, CA). Raw gene expression profiles were further preprocessed using robust multi-array analysis (Bolstad et al., 2003) with quantile normalization and probe-level signals were summarized in log base 2 scale. We selected a custom Chip Definition File (CDF) version 10 for more accurate probe mapping to genome (Dai et al., 2005). Data will be deposited into GEO with GSE0101.

Implementation and visualization

Pathway clustering results are visualized using Cytoscape (Shannon et al., 2003) with spring-embedded algorithm. A collection of functions have been implemented in R (Team, 2008). Specifically, the function **tfmPval()** transforms a data frame of p-values using an inverse logistic function. Users can specify the tuning parameter c, or they let the function to determine the amount of smoothing c automatically. The function **tspath()**, based on the **specc()** function in **kernlab** R package, performs the spectral analysis. Given a data frame of pathways p-values and number of clusters k, it finds k partitions of the pathways based on their temporal association trajectory. In addition, a library of temporal pathway analysis functions (**phdpath.r**) are provided to: a) perform pathway-phenotype association study at multiple time points using Globaltest (Goeman et al., 2004); b) plot covariate genes and significant pathways (with a user-supplied significance cutoff); c) map pathways and genes to existing pathway definitions, namely MSigDB and KEGG (Subramanian et al., 2005, Kanehisa, 2002). The function **sim2cyto()** generates attributes files with the between-pathway similarities into the format conformable with Cytoscape for visualization. The code is publicly available at the **Hero Group Reproducible Research** archive under **tspath**.



$$p^* = \begin{cases} 1 & \text{if } p < .05 \\ 0 & \text{otherwise} \end{cases} \qquad p^* = -\frac{\log(p)}{c - \log(p)}$$

Figure 5.1: Comparison of p-value transformation. A total of 30,000 p-values are shown.

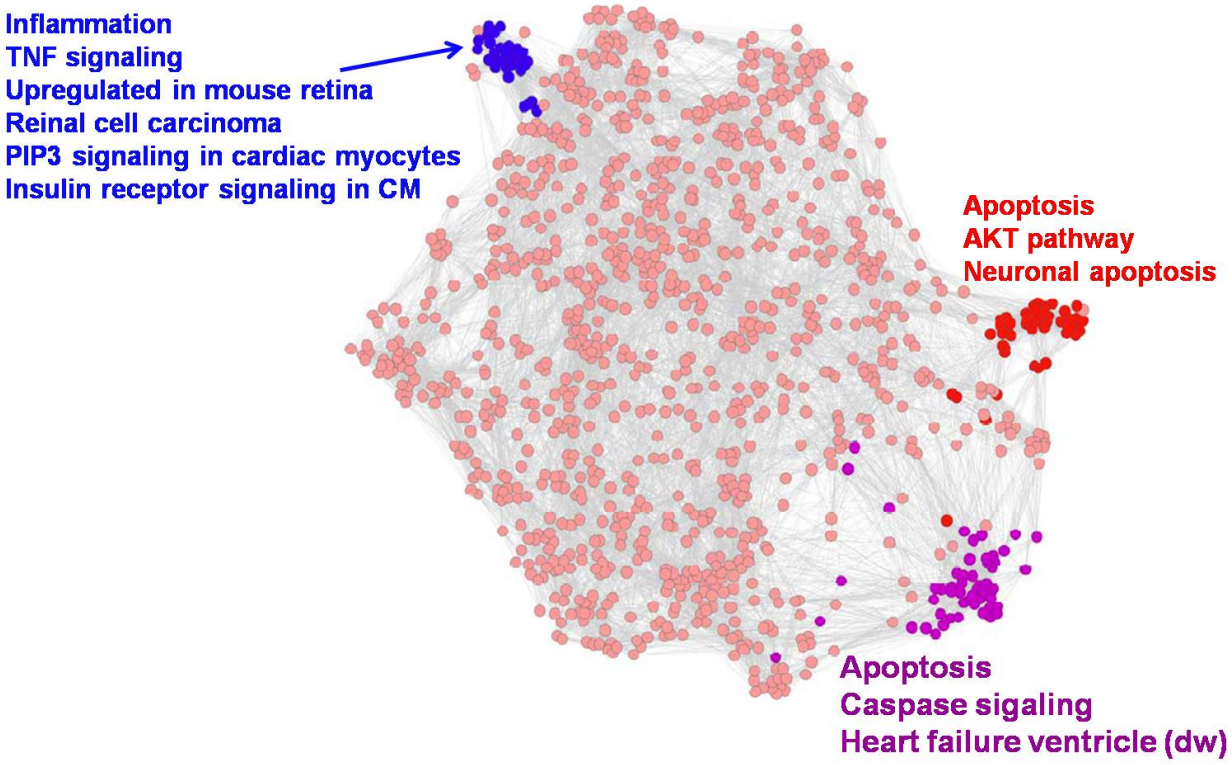


Figure 5.2: Gene pathway networks affected by alcohol consumption

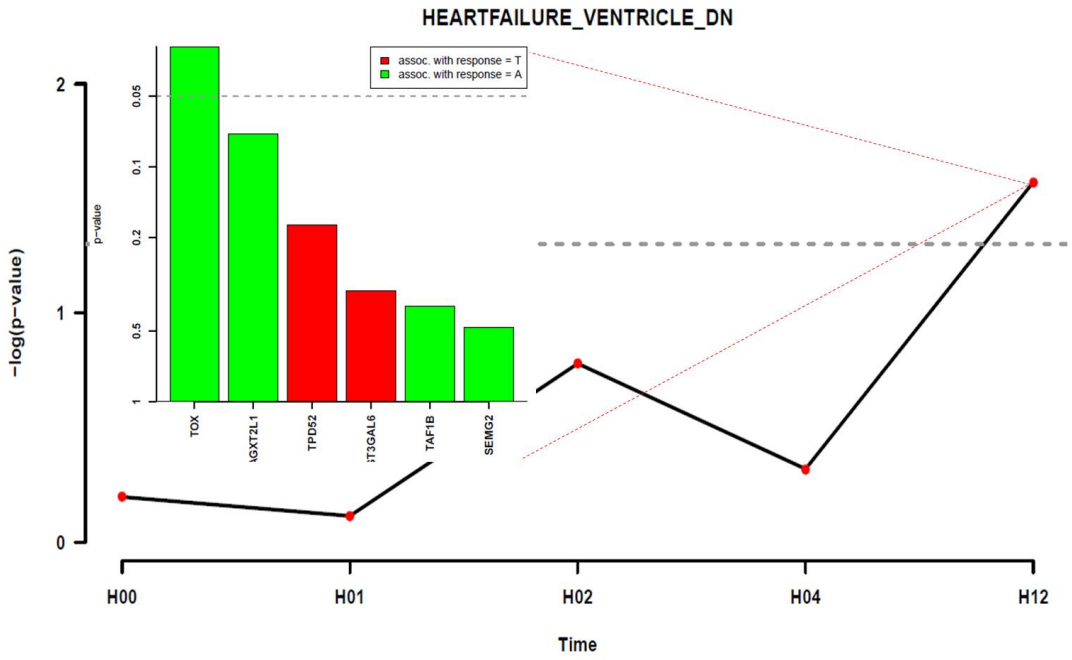


Figure 5.3: An exemplar pathway identified by proposed method

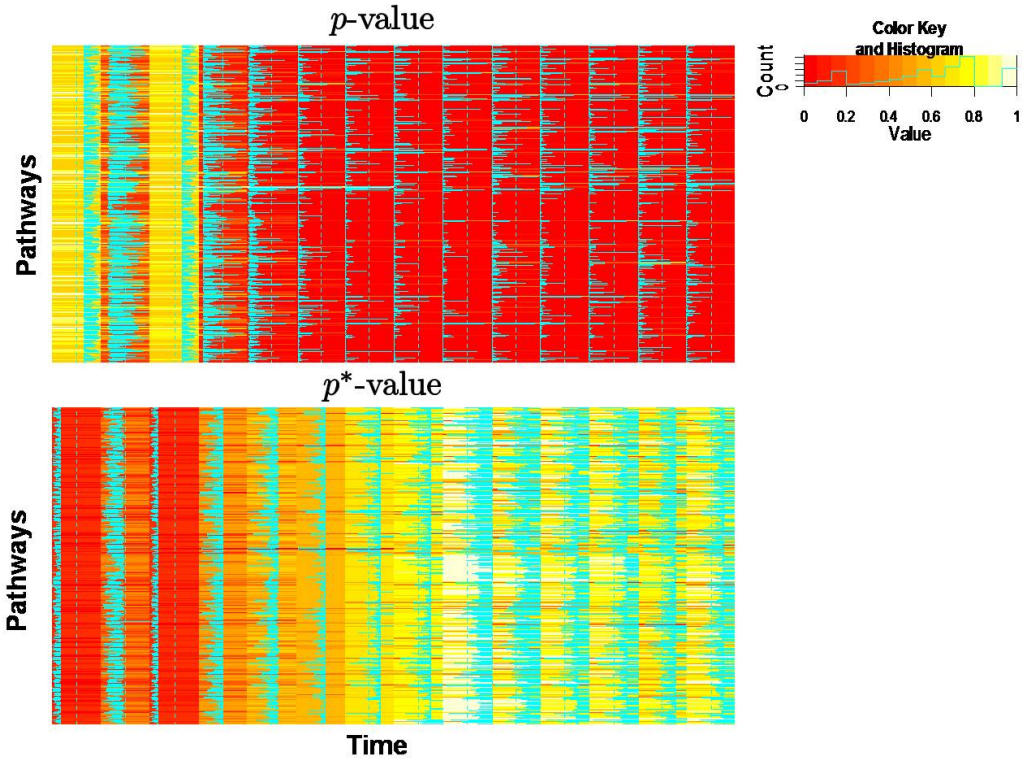


Figure 5.4: Inverse logistic transformed p-values yields refined resolution in significance measurements

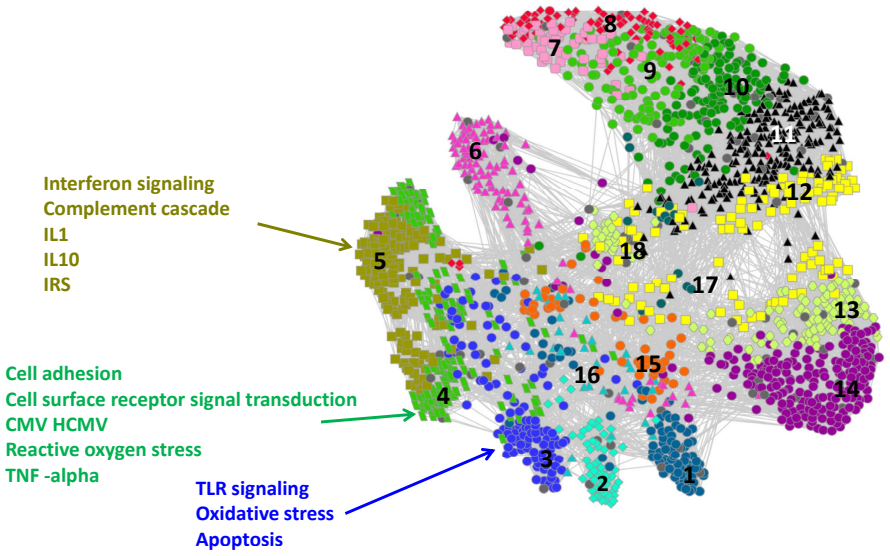


Figure 5.5: Temporal pathway network analysis of Influenza H3N2 viral infection

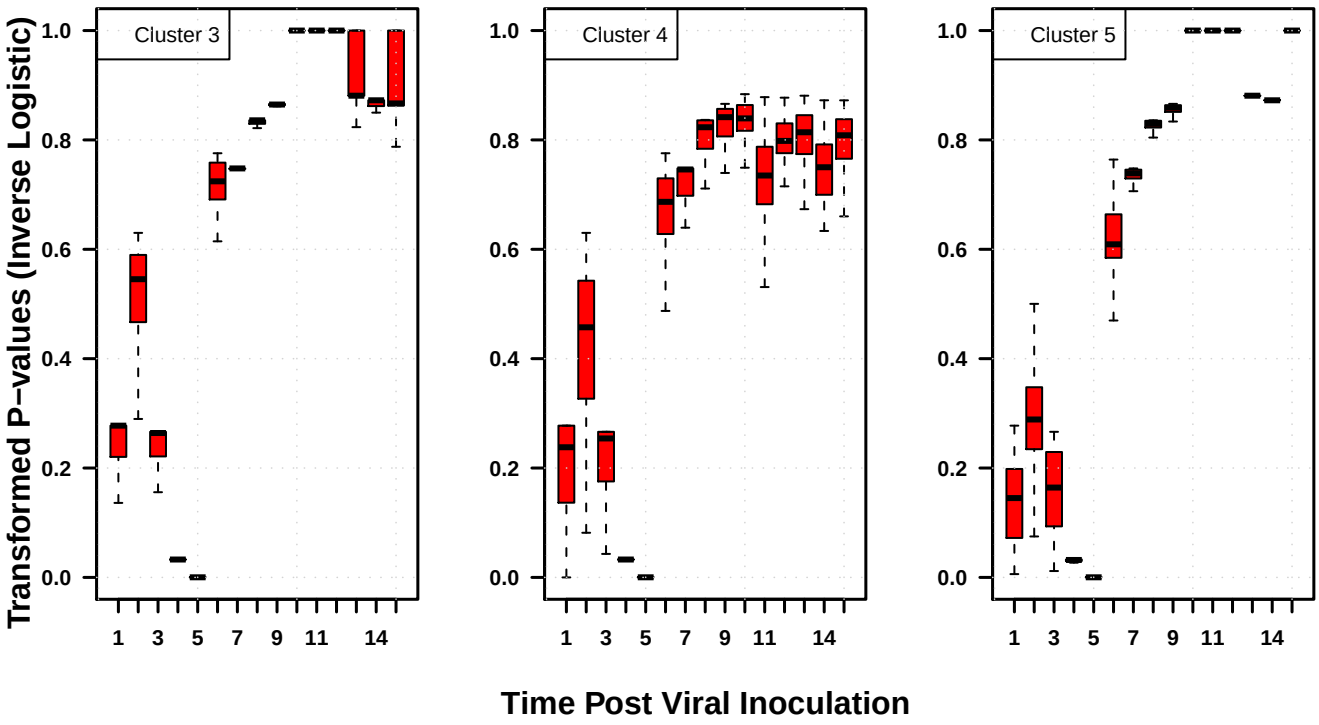


Figure 5.6: Temporal expression pattern of three exemplary clusters. Boxplots are shown on each sampling time for the transformed p-values of all pathways within cluster 3, 4, and 5. The time indices on the x-axis correspond to (baseline, 0, 5, 12, 21, 36, 45, 53, 60, 69, 77, 84, 93, 101, 108 hpi) in the same order.

Subjects	Juice	Wine	Alcohol	Water
CJ	*5*	*5*	*5*	*5*
CM	*4*	*5*	*4*	*5*
IF	*4*	*5*	*5*	*5*
JL	*4*	*3*	*2*	*5*
JR	*5*	*5*	*4*	*5*
MB	*4*	*5*	*5*	*4*
			Tot	108

Figure 5.7: Summary of samples in wine study

CHAPTER VI

Information Geometric Motif Analysis

6.1 INTRODUCTION

Transcription factors are proteins that bind to DNA sequences and drive the transcriptional program. Such binding causes the sequences to undergo important changes in their chromatin structure and also affects the recruitment or dissociation of other transcription factors. This has significant effect on the transcription of DNA and affects many critical biological processes such as cell differentiation and tissue development. Abnormal transcription factor binding is associated with many diseases including leukemia (Orkin and Zon, 2008, Jimenez-Sanchez et al., 2001, Farnham, 2009).

It is now well established that DNA binding proteins bind to their target sequences by recognizing short-length (usually 6 – 20 base pairs) sequence segments with specific patterns. These sequence segments, also known as *consensus sequences* or *motifs*, are unique to their corresponding transcription factors. Transcription factors are capable of recognizing and binding to their corresponding motifs with high specificity. Recently, high-throughput DNA sequencing technology (ChIP-seq) has allowed whole-genome occupancy patterns to be determined for more than 50 different proteins (Table S6.1). One of the most intriguing findings arising from these studies is that the genomic regions bound by one transcription factor often contain a variety of sequence motifs of other transcrip-

tion factors. For instance, our study of leukemia related transcription factors Hoxa9 and Meis1 showed statistically significant enrichment of motifs associated with at least 17 non-Hoxa9-Meis1 proteins (Huang et al., 2010). Some of these motifs appear in as much as 51 percent of the Hoxa9-Meis1 bound sequences. Together with the canonical motif of Hoxa9-Meis1-Pbx1 complex, these 18 motifs are tethered inside a window of approximately 300 base pairs on the genome (Figure 6.1). Binding at such close proximity by multiple proteins and/or protein complexes is likely to have significant regulatory effects on gene expression. These identified motifs demonstrate a variety of spatial distribution patterns along the sequences. In (Huang et al., 2010), we observed four different motif co-localization patterns as shown in Figure 6.1.

Such phenomena are not unique to Hoxa9 and Meis1 binding sites. Indeed, it is a rather common feature that is shared by many other transcription factors as revealed by their ChIP-seq profiles (Table S6.1). As protein binding is highly specific, the co-localization of multiple binding motifs on the same set of ChIP sequences provide strong physical evidence to support the hypothesis that transcription factors form integrated regulatory networks that collaboratively regulate transcription. The interactions between these proteins can be either cooperative or antagonistic and their combined net effect will determine the biological outcome that they set to achieve. As the DNA sequences may contain the important cues on such interactions, we develop statistical tools to fully utilize the power of DNA-sequencing to decode the sequence signals associated with multiple transcription factor interactions.

6.2 METHODS

6.2.1 Problem Formulation

Consider a set of n sequences S ($|S| = n$), each of length l (typically $l \approx$ 200-300 base pairs) that are bound by a transcription factor TF which the ChIP-sequencing experiment was designed to target. All sequences are aligned at the center of each sequence where TF is known to be located according to specified ChIP-sequencing protocol. In addition, a total of m motifs M ($|M| = m$) are found to be statistically significantly enriched in sequences S when compared to some randomly generated genomic background (D'Haeseleer, 2006). From a probabilistic generating model point-of-view, the spatial placement profile of a motif in sequences S can be considered as realizations from a probability distribution of a random variable (associated with this particular motif).

More specifically, the entire motif-sequence profile can be modeled as a statistical manifold (Amari, 1990, Kass and Vos, 1997) consisting of m probability distribution functions whose domain is the sequence set S:

$$\mathbb{M} = \{p(s \mid \theta) \mid \theta \in \Theta, s \in S\} \tag{6.1}$$

where the parameter θ of each distribution is unknown and is to be estimated from the data. We shall see that this formalization allows the binding events of multiple proteins to be examined on the basis of information geometry theory (Amari, 1990). The objective is to determine relationships of these proteins given the one-dimensional spatial distributions of their corresponding motifs. In the following, we outline the methods for: i) estimating motif spatial distribution functions; ii) testing the similarity (or spatial closeness) between these density functions; iii) partitioning transcription factors into functional subgroups based on shared spatial placement patterns of their motifs.

6.2.2 Representing and estimating the density of motif spatial distribution

We start by assuming that a set of motifs have been identified as significantly enriched in a set of DNA sequences. These sequences may be obtained from ChIP pull-down of a target protein followed by microarray analysis (ChIP-Chip) or parallel sequencing (ChIP-seq) etc. The motif analysis can be carried out with a variety of canonical methods. Briefly, these motif analysis methods come in two different flavors. The first flavor is *de novo* motif discovery. It assumes no prior knowledge of the motif to be searched and attempts to discover both its location and probabilistic representation specified in the form of position weight matrix (PWM) (e.g., Bailey et al., 2006, Pavesi et al., 2004, Ji et al., 2008, Zhou and Wong, 2004). The second category of motif analysis methods approaches the problem from the opposite angle. It takes *a priori* information concerning the PWM of a transcription factor and scans it through the entire sequence set. If the PWM is found to be overly represented in the sequences compared to a random genomic background, the corresponding motif is said to be enriched in the sequence set. This type of method is called *motif enrichment analysis (MEA)* (McLeay and Bailey, 2010, Bailey and Elkan, 1994). See also (Tompa et al., 2005) for more detailed information.

Here we do not impose any restrictions on the specific method of motif analysis employed. We only assume that a motif-sequence profile is obtained such that it specifies which motif(s) are enriched in which sequence and gives the exact locus. In our analysis presented here, we use *MEA* because we have found that *de novo* analysis delivers similar results (but fewer motifs). Let $\{X_i\}_{i=1}^{m}$ be a m-dimensional random vector with each element X_i defined as the genomic loci of (n_i) occurrences of motif i in the set of sequences S. The integer n_i accommodates multiple occurrences of one motif on the same sequences and n_i may or may not equal to n_j when $i \neq j$. We further normalize X_i by converting the exact genomic loci into a base pair distance from the center of the motif to the center of

the sequence segment containing it

$$\tilde{X}_i = X_i - \frac{l}{2} + \frac{l_i}{2} \tag{6.2}$$

where l_i is the length of motif i.

The normalized occurrence profile of motif i in sequences S can then be modeled as an independent and identically distributed sample from a distribution having unknown density. This density represents the distribution of the observed motif spatial distribution pattern. The objective will be to cluster these motif patterns, which are described by densities. The clustering procedure will require estimation of the density, for which we use the standard Gaussian kernel density estimation method (Silverman, 1986). In the end, we obtain a collection of probability density functions $\{P(X_i)\}_{i=1}^{m}$ representing the density of spatial distributions of motifs $\{X_i\}_{i=1}^{m}$.

6.2.3 Geodesic distance measure between motif spatial distributions

We will adopt a recently introduced information geometric approach to clustering densities (Carter et al., 2009). The set of motif spatial distribution densities $\{P(X_i)\}_{i=1}^{m}$ can be treated as a statistical manifold defined on the same space of probability distributions. Each density can be viewed as a point on this manifold and the difference between two densities is measured by the geodesic distance defined as the length of the shortest curve connecting two points. This geodesic distance must be estimated as the geometry of the space is assumed to be unknown. There are many non-parametric estimation methods that approximate the geodesic distance including Hellinger distance, Renyi-α entropy, and Kullback-Leibler (KL) divergence. We choose the latter for its close connection to the mutual information theory (Hero et al., 2002).

The KL divergence captures the dissimilarity between two spatial distributions, P and

Q, of a pair of motifs by

$$D_{KL}(P||Q) = \sum_{k=1}^{n} p(k) \log \frac{p(k)}{q(k)} \tag{6.3}$$

where $p(k)$ and $q(k)$ are the occurring frequency of motifs P and Q, respectively, at each basepair position indexed by k. As this divergence measure is non-symmetric and does not satisfy triangular inequality, we transform it into a distance metric by symmetricization,

$$D_{KL} = \frac{1}{2}(D_{KL}(P||Q) + D_{KL}(Q||P)) \tag{6.4}$$

which is both symmetric and non-negative. It has been shown that D_{KL} approximates the Fisher information metric (Carter et al., 2009) which measures the amount of information contained in the motif-sequence profile with regard to the unknown parameter of each motif spatial distribution.

Using the KL dissimilarity measure above, we can infer putative interactions between two proteins by examining how different their spatial distribution profiles are. The more differently they distribute across a genomic sequence, the more likely the two can co-bind at the same time on the same sequence. On the other hand, two proteins may compete for binding if their motifs are found to have almost exactly same distribution pattern. To jointly study multiple motif-sequence profiles, we can partition motifs into subsets based on their spatial distribution profiles. Once the pair-wise differences between motif spatial distributions are measured as described above, such difference measures can be used as the dissimilarity metrics in standard clustering algorithms to group motifs. Here we adopt hierarchical clustering.

6.2.4 Testing statistical significance of the KL distance metric

Given $\binom{m}{2}$ pair-wise distance metrics between spatial distributions of m motifs, it is desirable to quantify the significance of each estimated KL divergence. For a pair of

motifs with spatial distribution density P and Q, we define a null hypothesis H_0 being that their placement on the sequence set S share the same spatial distribution with distance $D_{KL}(P,Q) = 0$.

We employ an empirical procedure based on the bootstrap resampling technique (Efron and Tibshirani, 1993) to assess the statistical significance of the KL distance metrics. For a pair of motif spatial distributions P and Q, we randomly draw B samples with replacement from P and estimate the density of each bootstrap-resampled motif profile, denoting it P^b. We compute the KL distance between each P^b and Q, which form a bootstrap empirical distribution of the D_{KL} estimate. Denoting these distances as $\{D^b_{KL}\}^B_{b=1}$, the $100(1-\alpha)\%$ intervals can then be constructed from the bootstrap samples, e.g., α is typically 0.05. It has been suggested that, compared to the naive percentile method, better performance can be achieved on the confidence interval estimates by correcting the bias and skewness of this bootstrap distribution of D_{KL} (Efron, 1987, Efron and Tibshirani, 1993). Accordingly, we compute the bias-correction constant z_0 as

$$z_0 = \Phi^{-1}(G(D_{KL})) \tag{6.5}$$

where $G(D_{KL})$ is the bootstrap distribution defined as

$$G(D_{KL}) = \Pr(D^b_{KL} < D_{KL}) \tag{6.6}$$

The acceleration constant a is calculated by

$$a = (\sum_g^l U_k^3)/6(\sum_k^l U_k^2)^{3/2} \tag{6.7}$$

where U is the jackknife estimate of standard error of D_{KL} by deleting observations at each nucleotide position, indexed by k (Quenouille, 1956, Tukey, 1958, Efron, 1987). Intuitively, U indicates the skewness of bootstrap empirical distribution of D_{KL}.

With the above bootstrap confidence intervals, the construction of a hypothesis test is straightforward for both one-sided or two-sided tests, e.g., see (Mood et al., 1974) for confidence interval tests of hypothesis. It is worth noting that this empirical evidence-based KL test is distribution-free in that it does not depend on the actual spatial distribution of motifs.

6.2.5 Algorithms

We summarize the algorithms as follows. Algorithm 2, called the Information Geometrical Motif Analysis (IGMA) algorithm, specifies an information geometry based motif analysis procedure to cluster motifs based on their spatial placement patterns. Algorithm 3, called the Information Geometrical Inference of Protein Interaction (IGIPI) algorithm, specifies a hypothesis testing procedure to infer putative protein-protein interaction.

Algorithm 2: Information Geometrical Motif Analysis

Data: Motif-sequence profile $\{X_i\}_{i=1}^{m}$ where $X_i \in R$ indicates the loci of motif i in sequence set S

Result: Partitions of m motifs

begin

1 **for** $i \in (1,2,...,m)$ **do**

2 Calculate motif-sequence center distance $\tilde{X}_i = X_i - \frac{l}{2} + \frac{l_i}{2}$;

3 Estimate spatial distribution density with Gaussian kernel

4 **for** $(i,j) \in (1,2,...,m)$ **do**

5 Calculate $D_{KL}^{(i,j)} = \frac{1}{2}\left[\sum_{k=1}^{n} p(k)\log\frac{p(k)}{q(k)} + \sum_{k=1}^{n} q(k)\log\frac{q(k)}{p(k)}\right]$;

6 Hierarchical clustering motifs using $D_{KL}^{(i,j)}$ as distance metric

Algorithm 3: Information Geometrical Inference of Protein Interaction

Data: Motif-sequence profile $\{X_i\}_{i=1}^{m}$ where $X_i \in R$ indicates the loci of motif i in sequence set S

Result: Significance measures of K-L distance between motifs

begin

Compute D_{KL} for all pairs of motifs using Algorithm 2;

for $(i, j) \in (1, 2, ..., m)$ **do**

Randomly draw B samples with replacement from X_i;

for $b \in (1, 2, ..., B)$ **do**

Compute $D_{KL}^{(i,j)^b}$ using Algorithm 2;

Calculate bootstrap CDF $G(D_{KL}) = \Pr(D_{KL}^{(i,j)^b} < D_{KL}^{(i,j)})$;

Calculate bias constant $z_0 = \Phi^{-1}(G(D_{KL}))$;

for $g \in (1, 2, ..., l)$ **do**

Calculate influence value $U_g = D_{KL}^{(i,j)^{(-g)}}$;

Calculate bootstrap standard error $U = E[U] - U$;

Calculate acceleration constant $a = (\sum_g^l U_g^3)/6(\sum_g^l U_g^2)^{3/2}$;

Calculate $z[\alpha] = z_0 + \frac{z_0 + \Phi^{-1}(\alpha)}{1 - a(z_0 + \Phi^{-1}(\alpha))}$ for $0 < \alpha < 1$;

Calculate bias-corrected CI (lower tail): $G^{-1}(\Phi(z[\alpha]))$;

Calculate significance level of $D_{KL}^{(i,j)} = 0$

6.3 RESULTS

6.3.1 Analysis of motif-sequence profiles of transcription factors Hoxa9 and Meis1

We applied the proposed method to analyze motif-sequence profiles of 17 highly enriched motifs identified in DNA sequences bound by transcription factors Hoxa9 and Meis1 using ChIP-sequencing (Huang et al., 2010). Hoxa9 and Meis1 interact with each other and with another protein Pbx1. Together, they form a tri-meric protein complex (Hoxa9-Meis1-Pbx1) that synergistically drives leukemic transformation. Using Algorithms 2 and 3, we evaluated the KL distances between their motif spatial distribution densities (Figure 6.2).

These KL distances were used to cluster the motifs and revealed biologically sensible results. For instance, when compared to Hoxa9-Meis1-Pbx1 motif spatial profile, it appears that the motif of protein GFI1 is the most similar in its spatial distribution to that of Hoxa9-Meis1-Pbx1 complex. There is a high level of similarity between the two, suggesting that GFI1 may play an antagonistic role by competing with Hoxa9-Meis1-Pbx1 for binding sites *in vivo*. Indeed, there have been corroborative experimental results showing that Gfi1 represses Pbx1 and Meis1 and competes with HoxA9 autoregulation (Horman et al., 2009).

On the other hand, a subset of transcription factors including CREB, C/EBPα, and STAT, whose motif distributions are very complementary to that of Hoxa9-Meis-Pbx1 trimetric protein complexes, suggesting cooperative binding with Hoxa9-Meis1-Pbx1. In support of this, proteomic data shows that STAT, CREB, and C/EBPα physically interact with Hoxa9 and Meis1 (Huang et al., 2010).

6.3.2 Comparison with Kolmogrov-Simirnov (KS) test

In order to gain more insight on the performance of the proposed KL test, we analyzed the Hoxa9 and Meis1 data using the Kolmogorov-Smirnov (KS). Overall, KS and KL tests yield comparable results with KS statistics being larger than KL distance measures. This is reasonable as KS measures the largest difference between points on two cumulative distribution functions. For instance, the distance between GFI1 motif and Hoxa9-Meis1-Pbx1 is as low as 0.038 (KS) and 0.006 (KL) in contrast to the median distance measure of all motifs (0.118 (KS) and 0.101 (KL)). Both KS test and KL test reject the hypothesis that the two motifs have different spatial distributions, suggesting a compete-for-binding relationship between the two proteins.

In some cases however, there are significant differences between the two tests. Specifically, KS test did not reject the hypotheses that motifs of AP2F/AP4R/PAX3 have same spatial distribution profiles as that of Hoxa9-Meis1-Pbx1. On the other hand, the KL test rejects the hypothesis that these motifs with significant p-values of 0.004 (AP2F), 0.003 (AP4R), and 0.005 (PAX3). It suggests that these motifs are likely co-bind with Hoxa9-Meis1-Pbx1. These KL results are supported by experimental evidences showing that Hox expression are functionally related to the expression of PAX3 and AP2 proteins (Pruitt et al., 2004, Maconochie et al., 1999, Doerksen et al., 1996). The co-localized binding by these two proteins with Hoxa9-Meis1-Pbx1 is likely to be the reason for their functional association. Therefore, KL test seems to be more sensitive in detecting the relative weaker signals from the motif-sequence profiles.

6.4 DISCUSSION

It is known that transcription factors rarely function alone. Instead, the majority of cellular phenomena are carried out by protein machines, or aggregates of ten or more proteins

(Alberts, 1998). Recent advances in high-throughput DNA-sequencing allows genome-wide mapping of transcription factor binding sites. Analyzing these sites for enrichment of consensus motifs offers great opportunity to look into novel interactions between factors. In fact, motif analysis has become the *de facto* standard in ChIP-seq/ChIP-Chip and other high-throughput data analysis and proved to be valuable in studying transcription factor-genome interactions.

While a plethora of methods have been developed to identify motifs enriched in DNA sequences, there has been relatively less effort to characterize the spatial distributions of these motifs. Yet the spatial relationship between motifs is as important as the information about where a protein binds on the genome and what other motifs exist near by. A group of proteins that bind together can have important influence on regulating the transcriptional program. For instance, collaborative cobinding may critically increase the regulatory specificity whereas competing binding interferes with the normal function of proteins. The core Hoxa9 binding motif is only 4 letters in length (TAAT) and it is shared by almost all members of Hox protein family. There are many more sequences in a mammalian genome than that actually bound by Hoxa9. The question of how Hoxa9 achieves its high specificity in controlling heamatopoiesis and development not likely to be answered solely by a 4-letter motif. It is therefore of great importance to identify other proteins that bind collaboratively or antagonistically with Hoxa9. Having the ability to quantify the similarity of their motif spatial distributions and test its significance helps an experimenter to infer protein interactions.

In this paper, we have developed an information geometry based inference strategy for analyzing multi-motif spatial relations using high-throughput DNA-sequencing data. Modeling the motif-sequence profile with a simple statistical manifold, this method is robust and completely data-driven. We further extended it with a jackknife-after-bootstrap

hypothesis testing scheme to quantify uncertainty of the dissimilarity estimates. The proposed framework provides a statistically principalled approach to infer putative interaction of transcription factors. The results of applying the method to analyze motif-sequence profiles from real data established its effectiveness in characterizing motif spatial distributions. The results were corroborated by experimental validations and biological interpretations. In summary, our information geometry framework for motif analysis can provide insights of multi-protein interactions that induce or mediate gene transcription program.

It should be noted that the interactions of transcription factors are highly context-dependent and that their net effects are very likely distinct in different cell lineages and at different developmental stages. Nonetheless, the proposed methods provide an analysis strategy for studying the spatial arrangement of multiple transcription factors on genome. We believe that the proposed method and its implementation provide a useful and timely addition to the study of transcriptional regulatory networks using the next-generation high-throughput sequencing technology.

6.5 MATERIALS AND IMPLEMENTAITON

Chromatin immunoprecipitation (ChIP) A total of 150 million cells were crosslinked sequentially with disuccinimidylglutarate (45 min RT) and 1% formaldehyde (15 min RT). Hoxa9 and Meis1 immunoprecipitation was performed with anti-HA antibody (Abcam) pre-conjugated to Protein G magnetic beads (Dynal/Invitrogen). For ChIP-seq, size selection and sequencing were performed at the BC Cancer Agency Genome Sciences Centre (Vancouver, BC) as described previously (Robertson et al., 2007). Peak detection of enriched binding regions was performed using FindPeaks (Robertson et al., 2007) with an estimated false discovery rate < 0.05 as the selection criterion for enriched regions.

Motif enrichment analysis Comprehensive search of known transcription factor bind-

ing motifs was performed for 748 mouse transcription factors included in Genomatix proprietary Mat Base Matrix Family Library (Version 8.2, January 2010) that includes a total of 727 motifs (170 motif families). The DNA sequences of length 300bp from the center of each H/M peak were scanned for presence of any known transcription factor binding motif. A transcription factor binding motif is considered to be statistically significantly enriched in the H/M peaks if the number of sequences in which the motif is found to be present is significantly higher than its expected whole-genome occurrences according to standard z-test (z-score > 2.81; p-value < 0.005).

Method implementation The proposed method and visualization and utility functions for motif analysis were implemented in a R package *cMotif*. The code is publicly available at the **Hero Group Reproducible Research** archive under **cMotif** and the official R repository (soon). Several primary functions in the library *libmotif* are listed in Table S6.2.

6.6 ACKNOWLEDGEMENTS

This work was supported by a Pilot Grant from the University of Michigan Center for Computational Medicine and Bioinformatics (CCMB) to J.L.H and A.O.H.

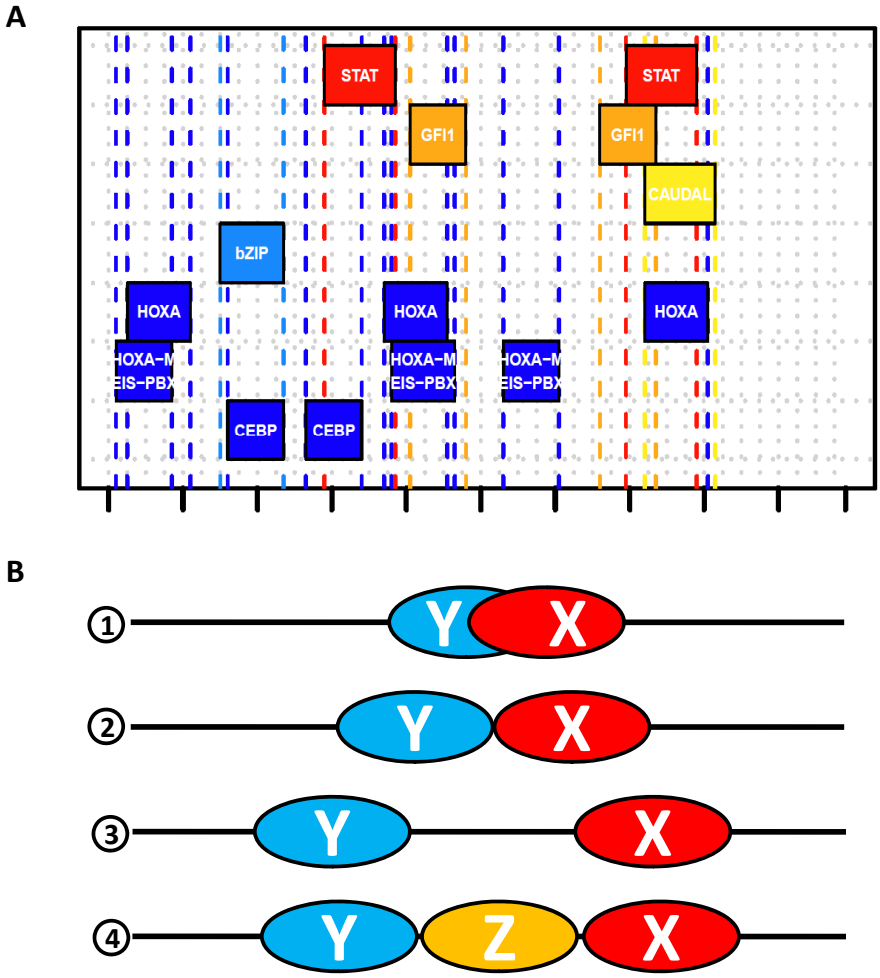


Figure 6.1: (A) An example of co-localization by multiple binding motifs at a Hoxa9 and Meis1 binding region (chr6:99298517-99298715). (B) Schematic plot of four typical motif binding patterns. (1) Two proteins bindings to the same locus; (2) Two proteins bind adjacently; (3) Two proteins bind closely but with gap; (4) Three proteins form a tri-meric protein complex.

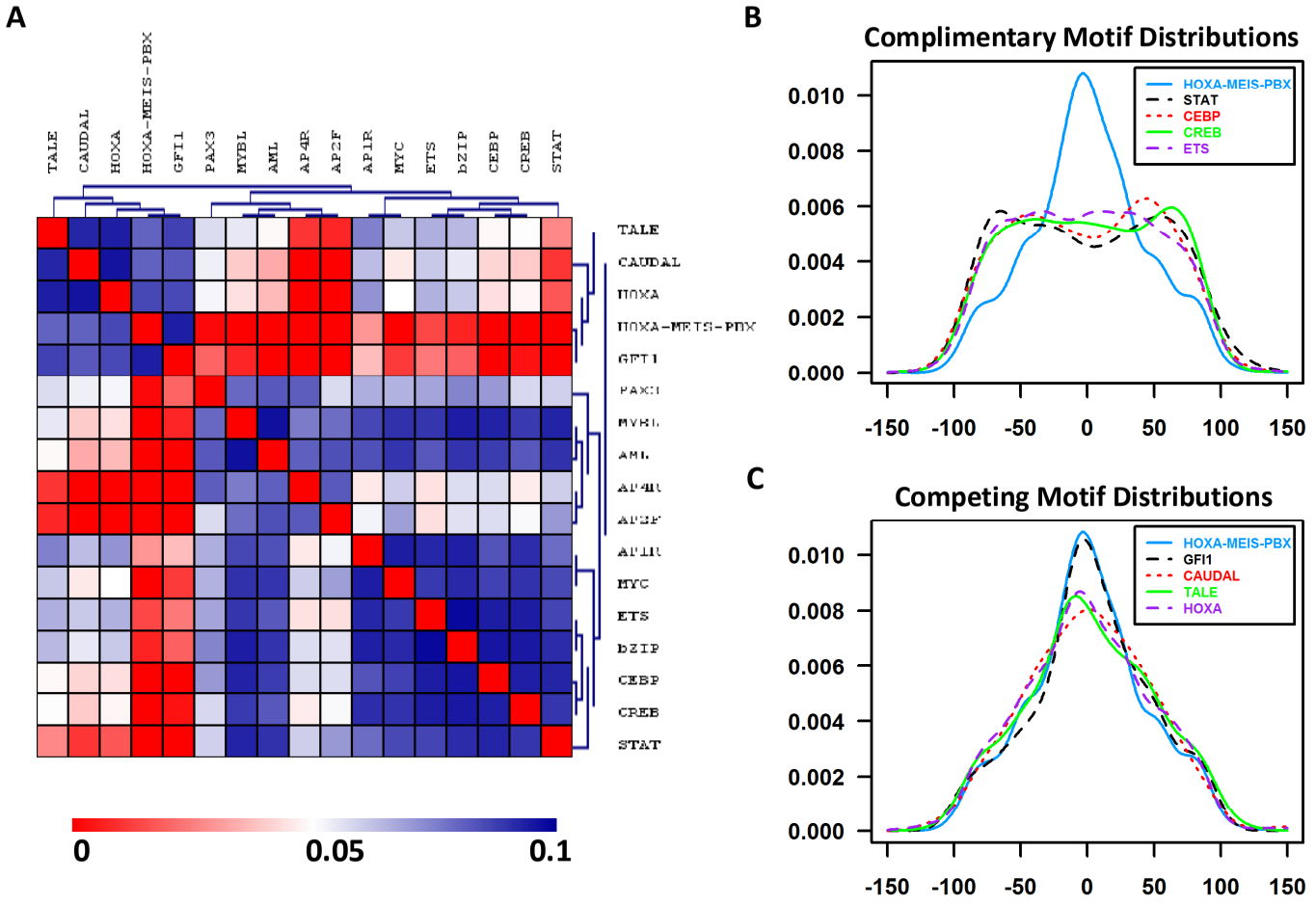


Figure 6.2: Information Geometrical Analysis of Hoxa9 and Meis1 ChIP-sequencing profiles. Spatial distributions 17 motifs enriched in DNA sequences bound by leukemia transcription factors Hoxa9 and Meis1 are analyzed. (A) Heatmap of pair-wise KL distance measures between motifs. (B) Motifs of STAT, CEBP, CREB, and ETS show complimentary spatial distribution with HMP complex. (C) Motifs of GFI1, CAUDAL, TALE, and HOXA show antagonistic spatial distribution with HMP complex.

	Kullback-Leibler		Kolmogorov-Smirnov	
Motif	**dist**	***p*-value**	**dist**	***p*-value**
AML	0.123	**0.004**	0.128	**0.002**
AP1R	0.071	**0.005**	0.100	**0.001**
AP2F	0.184	**0.004**	0.217	0.010
AP4R	0.173	**0.003**	0.139	0.064
bZIP	0.093	**0.004**	0.118	**0.000**
CAUDAL	0.019	0.009	0.056	0.158
CEBP	0.120	**0.004**	0.125	**0.000**
CREB	0.111	**0.003**	0.124	**0.000**
ETS	0.086	**0.003**	0.118	**0.000**
GFI1	0.006	0.108	0.038	0.913
HOXA	0.014	0.013	0.050	0.064
MYBL	0.113	**0.003**	0.120	**0.000**
MYC	0.102	**0.005**	0.123	**0.001**
PAX3	0.099	**0.005**	0.112	0.233
STAT	0.153	**0.003**	0.132	**0.000**
TALE	0.019	0.015	0.068	0.091

Table 6.1: Comparison of KS and KL distance measures of 17 motifs and Hoxa9-Meis1-Pbx1 motif. The *p-value* is the significance measure of how likely two motifs are from the same distributions. The smaller the *p-value* is, the less likely the two motifs share similar spatial distribution patterns. A total 500 bootstrap samples were drawn to estimate the *p-value*. Significance measures (p-values) ≤ 0.005 are shown in red.

Journal	Study	TFs (N)	Species	Tissue / Cell lines	Transcription Factors	No. of Peaks
Cell Stem Cell	Wilson 2010	*10*		hematopoietic progenitor cells	Scl/Tal1	7096
					Lyl1	4350
					Lmo2	9604
					Gata2	9234
					Runx1	5258
					Meis1	8386
					Pu.1	22720
					Erg	36167
					Fli-1	19601
					Gfi1b	8853
Cell Molecular Cell	Heinz 2010	*4*	mouse	Primary Macrophages	PU.1	45631
				Splenic B Cells	PU.1	32575
				peritoneal macrophages	C/ebpa	40,000
				peritoneal macrophages	C/ebpb	40,000
				spenic B cells	Oct2	13840
Nature	Kim 2010	*5*			CBP	28000
					SRF	
					CREB	
					NPAS4	
					RNAPII	
Immunity	Ghisletti 2010	*1*	mouse	macrophages (untreated)	p300	8064
			mouse	macrophages (LPS-treated)	p300	2742
Cell	Chen 2008	*15*	mouse	Embryonic stem cells	Nanog	
					Oct4	
					STAT3	
					Smad1	
					Sox2	
					Zfx	
					c-Myc	
					n-Myc	
					Klf4	
					Esrrb	
					Tcfcp2l1	
					E2f1	
					CTCF	
					p300	
					Suz12	
Nature	Visel 2009	*1*	mouse	embryonic forebrain	p300	
				embryonic midbrain	p300	
				embryonic limb tissue	p300	
Genes & Dev	Soler 2009	*5*		C88 MEL cells	Ldb1	
					Gata1	
					Tal1 (Scl)	
					Mtgr1	
					Eto2	
Genes & Dev	Nelson 2010	*3*		3T3-L1	PPAR gamma	
					RAR	
					RNApol-II	
		3	human	chronic myelogenous leukemia cells (K562, ATCC#CCL-243)	E2F4	
					E2F6	
					YY1	
Cell	Lupien 2008	*1*	human	MCF7 (breast cancer) cells	FOXA1	
Nat Method	Valouev 2008	*1*	human	Jurkat T cells	GABP	6442
					SRF	2429
					NRSF polyclonal	2960
					NRSF monoclon	2596
Nat Methods	Robertson 2007	*1*	human	HeLa S3 cells (stimulated)	STAT1	41,582
			human	HeLa S3 cells (unstimulated)	STAT1	11,004
Science	Johnson 2007	*1*	human		NRSF	1946

Table S6.1: All 51 published ChIP-seq experiments of transcription factors and/or transcription regulators. In this paper, we focused on 14 of them which are pertinent to leukemia differentiation and development.

Table S6.2: Primary functions in R implementation of Information Geometrical Motif Analysis

Function	**Description**
mfseqDist	Main function for computing Kullback-Leibler and Kolmogrov-Smirnov distance and perform hypothesis testing
kl.boot.test	compute confidence intervals from two data vectors of motif spatial distribution (Bias-Correctoin adjusted with jack-knife influence function)
kl.dist	Compute KL distance of two data vectors of motif locations
bi.bcanon	bivariate jackknife-after-bootstrap BCA non-parametric confidence interval for a list of α)
xtabSeqMotif	Cross tabulate sequences and motifs
getDistCenter	Compute (per region) distance from center of motif to center of region given a dataframe of motif occurrences in different chromosomal regions
plotMotif	Plot motif incidences on each genomic sequence (big PDF)
plotMotifTest	Plot motif, allow motif families to be combined
getMotifMatrix	Get motif incidence matrix (sequence by motif) All sequences are aligned at the center

CHAPTER VII

Concluding Remarks

Integrative learning based on statistical theories is a powerful solution for identifying characteristic and predictive disease features from an increasingly large volume of biomedical data. Aiming at translating data to actionable knowledge, this dissertation research addressed several issues related to the integrative statistical learning in bioinformatics analysis.

In a supervised learning setting, this dissertation studied and characterized the host transcriptional response patterns towards respiratory viral pathogens (Chapter II, III). These patterns not only revealed important host factors contributing to diverse clinical signs and symptoms, but also linked a non-passive response state to subclinical outcomes in human hosts who withstand viral insult. These results offered important insights to the molecular defense mechanism of human immune system. In an unsupervised learning setting, the findings were generalized to statistical models that can be used for exposure detection and risk stratification. Collectively, these findings offered a valuable tool for monitoring and managing infectious disease in natural environments. For the second problem studied in this thesis, a large body of genomic, genetic, and epigenetic data were jointly modeled to identify features that attribute to the leukemogenesis function of transcription factor Hoxa9. A direct result of this integrative analysis is a new biological model highlighting

the sophisticated multi-tier organization of Hoxa9-mediated long range enhancersomes. In both studies, a variety of methods such as unsupervised clustering, random mixed linear model, canonical correlation, and ensemble boosting classification techniques were integrated. These methods were chosen for modeling simplicity, practical adaptability, and optimal performance. The third part of this dissertation (Chapter VI, V) illustrated the importance of flexible learning strategy when new method development becomes necessary. The information geometric motif analysis (IGMA) algorithm presented in (Chapter VI) provides a new and simple estimation and inference framework for studying the cobinding effects of multiple transcription factors. By modeling motif spatial profiles in a statistical manifold, our method effectively characterized relationships between transcription factors. To model the temporal relationships between gene pathways, we developed a spectral analysis method based on graph theory (Chapter V). This method partitioned pathways into communities by decomposing the graph Laplacian embedding of significance measures on their activities. Practical solutions were provided for choosing model parameters with ease.

The Hoxa9 study reported in this dissertation proved the value of motif enrichment analysis (MEA). Our results suggested that MEA should be considered as a standard analysis routine as it has the unusual properties that *de novo* discovery methods cannot offer. It is easy to perform, fast to run, and provides deterministic results that do not vary from analysis to analysis. Currently, the genome-wide binding patterns have been characterized for more than 50 transcription factors. In the foreseeable future, more and more ChIP-seq analysis will be performed as the technology becomes more reliable and more cost-effective. The MEA is a valuable tool that cannot be neglected. Equipped with our IGMA algorithm (Chapter VI), the MEA helps the analysis of protein-DNA interactions and uncovers novel protein-protein interactions. These are of important value to functional

studies.

This dissertation opened the door to some interesting problems. The results on the studying host responses to respiratory viruses (Chapter II, III) promotes a difficult and exciting question — how to differentiate the exposure of different viral pathogens using blood based gene expression profiles? A even more challenging question is whether the detection can be carried out using a non-invasive techniques rather than blood sample acquisition. This may require new modeling techniques or new data to be obtained. It has been suggested that human social network topology may be used to improve the detection accuracy. The question of how and to what extent the sparse social network information can help improve the detection rate is worth investigation. Furthermore, to integrate and model the socio-structure with molecular signatures may require an adjustment to the design of analysis strategy.

The respiratory viral challenge study (Chapter II, III) also offered an very interesting problem that is whether we can further improve modeling efficiency by simultaneously analyzing genetic profiles, clinical observations, temporal information, and potentially geographical or social interactions. Our preliminary results showed that the high-order tensor analysis (HOTA) might be a natural fit in this situation. The HOTA is an extension of singular value decomposition (SVD) for decomposing multi-way matrices. It finds low rank approximation to multidimensional dataset that cannot be treated by SVD. Tensor analysis may represent a computationally efficient alternative solution and provide satisfactory level of approximation (Lathauwer et al., 2000, Kolda and Bader, 2009).

Another interesting problem related to the IGMA algorithm (Chapter VI) is to investigate to what extent that the motif relationships can be generalized to ChIP-sequencing profiles generated in different cellular contexts. Generally speaking, the developmental stages and disease progression are hierarchical and involve cascades of multiple signal-

ing pathways. As a result, the binding and interaction among transcription factors are presumed to be highly dynamic and context dependent. The question of whether this dynamic relationship can be modeled using IGMA is enticing. As the binding patterns of more and more transcription factors are being characterized using the ChIP-seq technology, a meta analysis that is not entirely restricted to the ChIP-seq profile of a single TF can be very interesting.

In addition, the IGMA algorithm itself can also be improved. Currently, it employs a *jackknife-after-bootstrap* scheme to assess the statistical significance of estimated distance between motif spatial profiles. This procedure provides excellent approximation to the distribution of similarity estimates comparing motif spatial profiles. However, the method is computational expensive. It is not a problem for our study of Hoxa9 because of the relatively small number of binding sites. However, other ChIP-seq data are on the order of tens and thousands of observations. This might render the algorithm computationally prohibitive. A possible solution to this problem might be to adopt another approximation approach similar to the k-fold cross-validation scheme that is often used in classification regime. The whole dataset may be divided into a few dozens of smaller sets. The estimation can then be carried out on individual sets independently, followed by construction of confidence intervals and test for significance from the k-fold estimates. That will eliminate the use of *jackknife* and *bootstrap* resampling techniques and lower the overall computation overload.

Bibliography

C. Abraham and J. H. Cho. Inflammatory bowel disease. *N Engl J Med*, 361(21):2066–2078, 2009. 25

S. Akira, S. Uematsu, and O. Takeuchi. Pathogen recognition and innate immunity. *Cell*, 124(4):783–801, 2006. 16

B. Alberts. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3):291–4, 1998. 206

I. C. Allen, M. A. Scull, C. B. Moore, E. K. Holl, E. McElvania-TeKippe, D. J. Taxman, E. H. Guthrie, R. J. Pickles, and J. P. Ting. The nlrp3 inflammasome mediates in vivo innate immunity to influenza a virus through recognition of viral rna. *Immunity*, 30(4): 556–65, 2009. 18, 26

R. D. Allison, A. Katsounas, D. E. Koziol, D. E. Kleiner, H. J. Alter, R. A. Lempicki, B. Wood, J. Yang, B. Fullmer, K. J. Cortez, M. A. Polis, and S. Kottilil. Association of interleukin-15-induced peripheral immune activation with hepatic stellate cell activation in persons coinfected with hepatitis c virus and hiv. *J Infect Dis*, 200(4):619–623, 2009. 88

S. Amari. Differential-geometrical methods in statistics. *Springer*, page 294, 1990. 197

J. Andrejeva, K. S. Childs, D. F. Young, T. S. Carlos, N. Stock, S. Goodbourn, and R. E.

Randall. The v proteins of paramyxoviruses bind the ifn-inducible rna helicase, mda-5, and inhibit its activation of the ifn-beta promoter. *Proc Natl Acad Sci U S A*, 101(49): 17264–9, 2004. 16

S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 30(1):41–7, 2002. 129

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994. 198

T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue):W369–73, 2006. 138, 198

T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. Meme suite: tools for motif discovery and searching. *Nucleic Acids Res*, 37(Web Server issue):W202–8, 2009. 138

Z. Bar-Joseph, G. K. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon. Continuous representations of time-series gene expression data. *J Comput Biol*, 10(3-4):341–56, 2003. 174

B. Barrett, R. Brown, R. Voland, R. Maberry, and R. Turner. Relations among questionnaire and laboratory measures of rhinovirus infection. *Eur Respir J*, 28(2):358–63, 2006. 29, 34, 105

F. Baty, M. Facompre, J. Wiegand, J. Schwager, and M. H. Brutsche. Analysis with re-

spect to instrumental variables for the exploration of microarray data structures. *BMC Bioinformatics*, 7:422, 2006. 180, 185

M. J. Belsey, B. de Lima, A. K. Pavlou, and J. W. Savopoulos. Influenza vaccines. *Nat Rev Drug Discov*, 5(3):183–4, 2006. 79

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57: 289300, 1995. 31, 38, 39

J.P. Benzkri. L'analyse des donnes. *Dunod Publishing*, II(L'analyse des correspondances): 632, 1982. 144

P Bhlmann. Boosting for high-dimensional linear models. *Annals of Statistics*, 34(2): 559–583, 2006. 43, 44, 109

P Bhlmann and Torsten Hothorn. Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, 22(4):477505, 2007. 23, 43, 44

P Bhlmann and B Yu. Boosting with the l2 loss: regression and classification. *Journal of the American Statistical Association*, 98:324339, 2003. 32, 43, 44

B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–93, 2003. 35, 107, 186

L. Borghesi and C. Milcarek. Innate versus adaptive immunity: a paradigm past its prime? *Cancer Res*, 67(9):3989–93, 2007. 172

L Breiman and J. H. Friedman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 45

F. Carrat, E. Vergu, N. M. Ferguson, M. Lemaitre, S. Cauchemez, S. Leach, and A. J. Valleron. Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *Am J Epidemiol*, 167(7):775–85, 2008. 10, 27, 49, 170

K. M. Carter, R. Raich, W. G. Finn, and 3rd Hero, A. O. Fine: fisher information nonparametric embedding. *IEEE Trans Pattern Anal Mach Intell*, 31(11):2093–8, 2009. 199, 200

L. Carthagena, A. Bergamaschi, J. M. Luna, A. David, P. D. Uchil, F. Margottin-Goguet, W. Mothes, U. Hazan, C. Transy, G. Pancino, and S. Nisole. Human trim gene expression in response to interferons. *PLoS One*, 4(3):e4894, 2009. 87

S. Casas, B. Nagy, E. Elonen, A. Aventin, M. L. Larramendy, J. Sierra, T. Ruutu, and S. Knuutila. Aberrant expression of hoxa9, dek, cbl and csf1r in acute myeloid leukemia. *Leuk Lymphoma*, 44(11):1935–41, 2003. 129

J. C. Castelli, B. A. Hassel, K. A. Wood, X. L. Li, K. Amemiya, M. C. Dalakas, P. F. Torrence, and R. J. Youle. A study of the interferon antiviral mechanism: apoptosis activation by the 2-5a system. *J Exp Med*, 186(6):967–72, 1997. 26

G. Chen, X. Guo, F. Lv, Y. Xu, and G. Gao. p72 dead box rna helicase is required for optimal function of the zinc-finger antiviral protein. *Proc Natl Acad Sci U S A*, 105(11): 4352–7, 2008. 24

G. Chen, M. H. Shaw, Y. G. Kim, and G. Nunez. Nod-like receptors: role in innate immunity and inflammatory disease. *Annu Rev Pathol*, 4:365–98, 2009. 18

P. Clarke and K. L. Tyler. Apoptosis in animal models of virus-induced disease. *Nat Rev Microbiol*, 7(2):144–55, 2009. 182

M. J. Clemens and A. Elia. The double-stranded rna-dependent protein kinase pkr: structure and function. *J Interferon Cytokine Res*, 17(9):503–24, 1997. 26

D. N. Cook, D. S. Pisetsky, and D. A. Schwartz. Toll-like receptors in the pathogenesis of human disease. *Nat Immunol*, 5(10):975–9, 2004. 182

N. J. Cox and K. Subbarao. Influenza. *Lancet*, 354(9186):1277–82, 1999. 10, 170, 173

M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res*, 33(20): e175, 2005. 35, 186

M. D. de Jong, C. P. Simmons, T. T. Thanh, V. M. Hien, G. J. Smith, T. N. Chau, D. M. Hoang, N. V. Chau, T. H. Khanh, V. C. Dong, P. T. Qui, B. V. Cam, Q. Ha do, Y. Guan, J. S. Peiris, N. T. Chinh, T. T. Hien, and J. Farrar. Fatal outcome of human influenza a (h5n1) is associated with high viral load and hypercytokinemia. *Nat Med*, 12(10): 1203–7, 2006. 10, 170

J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–11, 2002. 137

Jr. Dennis, G., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3, 2003. 175

P. D'Haeseleer. How does dna sequence motif discovery work? *Nat Biotechnol*, 24(8): 959–61, 2006. 143, 197

N. Dobigeon, S. Moussaoui, M. Coulon, J-Y. Tourneret, and A Hero. Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery. *IEEE Trans on Signal Processing*, 57(11):4355–4368, 2009. 22, 31, 41

L. F. Doerksen, A. Bhattacharya, P. Kannan, D. Pratt, and M. A. Tainsky. Functional interaction between a rare and an ap-2 binding site in the regulation of the human hox a4 gene promoter. *Nucleic Acids Res*, 24(14):2849–56, 1996. 205

E. Durand, A. Al Haj Zen, F. Addad, C. Brasselet, G. Caligiuri, F. Vinchon, P. Lemarchand, M. Desnos, P. Bruneval, and A. Lafont. Adenovirus-mediated gene transfer of superoxide dismutase and catalase decreases restenosis after balloon angioplasty. *J Vasc Res*, 42(3):255–65, 2005. 19, 26

B Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. 32, 37, 45, 110

B. Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987. 201

B. Efron and R. Tibshirani. An introduction to the bootstrap. *Chapman and Hall*, 1993. 201

R. Efron, B. Tibshirani. On testing the significance of sets of genes. *Annals of Applied Statistics*, 1(1):107–129, 2007. 175

M. Falco, R. Biassoni, C. Bottino, M. Vitale, S. Sivori, R. Augugliaro, L. Moretta, and A. Moretta. Identification and molecular cloning of p75/airm1, a novel member of the sialoadhesin family that functions as an inhibitory receptor in human natural killer cells. *J Exp Med*, 190(6):793–802, 1999. 86

A. R. Falsey, P. A. Hennessey, M. A. Formica, C. Cox, and E. E. Walsh. Respiratory syncytial virus infection in elderly and high-risk adults. *N Engl J Med*, 352(17):1749–59, 2005. 79

J. Faraway. Linear models with r. *Chapman and Hall*, 2004. 31, 40

P. J. Farnham. Insights from genomic profiling of transcription factors. *Nat Rev Genet*, 10 (9):605–16, 2009. 195

A. P. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. J. Jones. Findpeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–30, 2008. 131

J. E. Fenner, R. Starr, A. L. Cornish, J. G. Zhang, D. Metcalf, R. D. Schreiber, K. Sheehan, D. J. Hilton, W. S. Alexander, and P. J. Hertzog. Suppressor of cytokine signaling 1 regulates the immune response to infection by a unique inhibition of type i interferon activity. *Nat Immunol*, 7(1):33–9, 2006. 11, 80, 84, 98

A. A. Ferrando, S. A. Armstrong, D. S. Neuberg, S. E. Sallan, L. B. Silverman, S. J. Korsmeyer, and A. T. Look. Gene expression signatures in mll-rearranged t-lineage and b-precursor acute leukemias: dominance of hox dysregulation. *Blood*, 102(1):262–8, 2003. 129

R. A. Floyd, K. Hensley, F. Jaffery, L. Maidt, K. Robinson, Q. Pye, and C. Stewart. Increased oxidative stress brought on by pro-inflammatory cytokines in neurodegenerative processes and the protective role of nitrone-based free radical traps. *Life Sci*, 65(18-19): 1893–9, 1999. 26

R Gentleman, Vincent Carey, W Huber, R Irizarry, and S Dudoit. Bioinformatics and

computational biology solutions using r and bioconductor. *Springer*, page 473, 2005. 176

J. E. Gern and W. W. Busse. Association of rhinovirus infections with asthma. *Clin Microbiol Rev*, 12(1):9–18, 1999. 79

J. E. Gern and W. W. Busse. Relationship of viral infections to wheezing illnesses and asthma. *Nat Rev Immunol*, 2(2):132–8, 2002. 80

J. E. Gern, A. G. Mosser, C. A. Swenson, P. J. Rennie, R. J. England, J. Shaffer, and H. Mizoguchi. Inhibition of rhinovirus replication in vitro and in vivo by acid-buffered saline. *J Infect Dis*, 195(8):1137–43, 2007. 111

J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1): 93–9, 2004. 31, 39, 175, 176, 187

P. Grunwald. The minimum description length principle. *The MIT Press*, page 504, 2007. 179

M. Harada, K. Nakashima, T. Hirota, M. Shimizu, S. Doi, K. Fujita, T. Shirakawa, T. Enomoto, M. Yoshikawa, H. Moriyama, K. Matsumoto, H. Saito, Y. Suzuki, Y. Nakamura, and M. Tamari. Functional polymorphism in the suppressor of cytokine signaling 1 gene associated with adult asthma. *Am J Respir Cell Mol Biol*, 36(4):491–6, 2007. 99

T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning. *Springer*, 2001. 178

T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. 1990. 36, 107

N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 39(3):311–8, 2007. 134

S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol Cell*, 38(4):576–89, 2010. 140, 141, 159

H. Hemmi, T. Kaisho, O. Takeuchi, S. Sato, H. Sanjo, K. Hoshino, T. Horiuchi, H. Tomizawa, K. Takeda, and S. Akira. Small anti-viral compounds activate immune cells via the tlr7 myd88-dependent signaling pathway. *Nat Immunol*, 3(2):196–200, 2002. 16

J. O. Hendley. Rhinovirus colds: immunology and pathogenesis. *Eur J Respir Dis Suppl*, 128 (Pt 1):340–4, 1983. 79

A. Hero and G. Fleury. Pareto-optimal methods for gene ranking. *Journal of VLSI Signal Processing*, 38:259–275, 2004. 31, 38

A. Hero, B. Ma, O. Michel, and J. Gorman. Alpha-divergence for classification, indexing and retrieval. *Technical Report*, 2002. 199

K. Honda, H. Yanai, H. Negishi, M. Asagiri, M. Sato, T. Mizutani, N. Shimada, Y. Ohba, A. Takaoka, N. Yoshida, and T. Taniguchi. Irf-7 is the master regulator of type-i interferon-dependent immune responses. *Nature*, 434(7034):772–7, 2005. 11, 80

S. R. Horman, C. S. Velu, A. Chaubey, T. Bourdeau, J. Zhu, W. E. Paul, B. Gebelein, and

H. L. Grimes. Gfi1 integrates progenitor versus granulocytic transcriptional programming. *Blood*, 113(22):5466–75, 2009. 204

S. Horvath, B. Zhang, M. Carlson, K. V. Lu, S. Zhu, R. M. Felciano, M. F. Laurance, W. Zhao, S. Qi, Z. Chen, Y. Lee, A. C. Scheck, L. M. Liau, H. Wu, D. H. Geschwind, P. G. Febbo, H. I. Kornblum, T. F. Cloughesy, S. F. Nelson, and P. S. Mischel. Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. *Proc Natl Acad Sci U S A*, 103(46):17402–7, 2006. 184

H. Hotelling. Relations between two sets of variates. *Biometrika*, (28):321–377, 1936. 141, 144

O Hssjer. On the coefficient of determination for mixed regression models. *Journal of Statistical Planning and Inference*, 138(10):3022–3038, 2008. 31, 40

Q. Huang, D. Liu, P. Majewski, L. C. Schulte, J. M. Korn, R. A. Young, E. S. Lander, and N. Hacohen. The plasticity of dendritic cell responses to pathogens and their components. *Science*, 294(5543):870–5, 2001. 11, 13, 80, 82, 145

Y. Huang, K. Sitwala, J. Bronstein, D. Sanders, M. Dandekar, C. Collins, G. Robertson, J. MacDonald, T. Cezard, M. Bilenky, N. Thiessen, Y. Zhao, T. Zeng, M. Hirst, A. Hero, S. Jones, and J. Hess. Identification of hoxa9-regulated hoxasomes in hematopoietic cells. *Submitted*, 2010. 130, 132, 141, 142, 196, 204

W. Huang da, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009. 175

J. P. Hugot, M. Chamaillard, H. Zouali, S. Lesage, J. P. Cezard, J. Belaiche, S. Almer, C. Tysk, C. A. O'Morain, M. Gassull, V. Binder, Y. Finkel, A. Cortot, R. Modigliani,

P. Laurent-Puig, C. Gower-Rousseau, J. Macry, J. F. Colombel, M. Sahbatou, and G. Thomas. Association of nod2 leucine-rich repeat variants with susceptibility to crohn's disease. *Nature*, 411(6837):599–603, 2001. 25

A. Iannello, S. Samarani, O. Debbeche, R. Ahmad, M. R. Boulassel, C. Tremblay, E. Toma, J. P. Routy, and A. Ahmad. Potential role of interleukin-18 in the immunopathogenesis of aids: involvement in fratricidal killing of nk cells. *J Virol*, 83 (12):5999–6010, 2009. 90

T. Ichinohe, H. K. Lee, Y. Ogura, R. Flavell, and A. Iwasaki. Inflammasome recognition of influenza virus is essential for adaptive immune responses. *J Exp Med*, 206(1):79–87, 2009. 11, 80, 100

G. G. Jackson, H. F. Dowling, I. G. Spiesman, and A. V. Boand. Transmission of the common cold to volunteers under controlled conditions. i. the common cold as a clinical entity. *AMA Arch Intern Med*, 101(2):267–78, 1958. 29, 34, 104, 105

S. M. Janciauskiene, I. M. Nita, and T. Stevens. Alpha1-antitrypsin, old dog, new tricks. alpha1-antitrypsin exerts in vitro anti-inflammatory activity in human monocytes by elevating camp. *J Biol Chem*, 282(12):8573–82, 2007. 90

H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong. An integrated software system for analyzing chip-chip and chip-seq data. *Nat Biotechnol*, 26(11): 1293–300, 2008. 131, 138, 145, 198

X. Ji, W. Li, J. Song, L. Wei, and X. S. Liu. Ceas: cis-regulatory element annotation system. *Nucleic Acids Res*, 34(Web Server issue):W551–4, 2006. 133

G. Jimenez-Sanchez, B. Childs, and D. Valle. Human disease genes. *Nature*, 409(6822): 853–5, 2001. 195

W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–27, 2007. 107

L. A. Joosten, B. Heinhuis, S. Abdollahi-Roodsaz, G. Ferwerda, L. Lebourhis, D. J. Philpott, M. A. Nahori, C. Popa, S. A. Morre, J. W. van der Meer, S. E. Girardin, M. G. Netea, and W. B. van den Berg. Differential function of the nacht-lrr (nlr) members nod1 and nod2 in arthritis. *Proc Natl Acad Sci U S A*, 105(26):9017–22, 2008. 25

M. Kanehisa. The kegg database. *Novartis Found Symp*, 247:91–101; discussion 101–3, 119–28, 244–52, 2002. 187

D. C. Kang, R. V. Gopalkrishnan, Q. Wu, E. Jankowsky, A. M. Pyle, and P. B. Fisher. mda-5: An interferon-inducible putative rna helicase with double-stranded rna-dependent atpase activity and melanoma growth-suppressive properties. *Proc Natl Acad Sci U S A*, 99(2):637–42, 2002. 16

T. D. Kanneganti, N. Ozoren, M. Body-Malapel, A. Amer, J. H. Park, L. Franchi, J. Whitfield, W. Barchet, M. Colonna, P. Vandenabeele, J. Bertin, A. Coyle, E. P. Grant, S. Akira, and G. Nunez. Bacterial rna and small antiviral compounds activate caspase-1 through cryopyrin/nalp3. *Nature*, 440(7081):233–6, 2006. 18

R. Kass and Paul. Vos. Geometrical foundations of asymptotic inference. *Wiley*, page 355, 1997. 197

T. Kawai and S. Akira. Tlr signaling. *Semin Immunol*, 19(1):24–32, 2007. 10, 80, 182

T. Kawai, S. Sato, K. J. Ishii, C. Coban, H. Hemmi, M. Yamamoto, K. Terai, M. Matsuda, J. Inoue, S. Uematsu, O. Takeuchi, and S. Akira. Interferon-alpha induction through toll-like receptors involves a direct interaction of irf7 with myd88 and traf6. *Nat Immunol*, 5(10):1061–8, 2004. 10, 80

T. Kino and G. P. Chrousos. Human immunodeficiency virus type-1 accessory protein vpr: a causative agent of the aids-related insulin resistance/lipodystrophy syndrome? *Ann N Y Acad Sci*, 1024:153–67, 2004. 24

T. Kino, A. Gragerov, A. Valentin, M. Tsopanomihalou, G. Ilyina-Gragerova, R. Erwin-Cohen, G. P. Chrousos, and G. N. Pavlakis. Vpr protein of human immunodeficiency virus type 1 binds to 14-3-3 proteins and facilitates complex formation with cdc25c: implications for cell cycle arrest. *J Virol*, 79(5):2780–7, 2005. 24

S. Kirchberger, O. Majdic, P. Steinberger, S. Bluml, K. Pfistershammer, G. Zlabinger, L. Deszcz, E. Kuechler, W. Knapp, and J. Stockl. Human rhinoviruses inhibit the accessory function of dendritic cells by inducing sialoadhesin and b7-h1 expression. *J Immunol*, 175(2):1145–52, 2005. 100

K. S. Kobayashi, M. Chamaillard, Y. Ogura, O. Henegariu, N. Inohara, G. Nunez, and R. A. Flavell. Nod2-dependent regulation of innate and adaptive immunity in the intestinal tract. *Science*, 307(5710):731–4, 2005. 18

S. Kofler, T. Nickel, and M. Weis. Role of cytokines in cardiovascular diseases: a focus on endothelial responses to inflammation. *Clin Sci (Lond)*, 108(3):205–13, 2005. 26

T Kohonen. Self-organizing maps. *Series in Information Sciences*, 30, 1995. 13, 30, 36, 107, 145

T. Kolda and B. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3): 455–500, 2009. 216

B. D. Korman, D. L. Kastner, P. K. Gregersen, and E. F. Remmers. Stat4: genetics, mechanisms, and implications for autoimmunity. *Curr Allergy Asthma Rep*, 8(5):398–403, 2008. 20

L. Lathauwer, B. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM Journal of Matrix Analysis Application*, 21(4):1253–1278, 2000. 216

L. L. Lau, B. J. Cowling, V. J. Fang, K. H. Chan, E. H. Lau, M. Lipsitch, C. K. Cheng, P. M. Houck, T. M. Uyeki, J. S. Peiris, and G. M. Leung. Viral shedding and clinical illness in naturally acquired influenza virus infections. *J Infect Dis*, 201(10):1509–16, 2010. 48

EH Lehmann. Nonparametrics: Statistical methods based on ranks. *Holden Day*, page 233, 1975. 38

J. Li, H. Shen, K. L. Himmel, A. J. Dupuy, D. A. Largaespada, T. Nakamura, Jr. Shaughnessy, J. D., N. A. Jenkins, and N. G. Copeland. Leukaemia disease genes: large-scale cloning and pathway predictions. *Nat Genet*, 23(3):348–53, 1999. 136

L. Li. Gadem: a genetic algorithm guided formation of spaced dyads coupled with an em algorithm for motif discovery. *J Comput Biol*, 16(2):317–29, 2009. 138

M. Lohoff, G. S. Duncan, D. Ferrick, H. W. Mittrucker, S. Bischof, S. Prechtl, M. Rollinghoff, E. Schmitt, A. Pahl, and T. W. Mak. Deficiency in the transcription factor interferon regulatory factor (irf)-2 leads to severely compromised development of natural killer and t helper type 1 cells. *J Exp Med*, 192(3):325–36, 2000. 98

A. T. Look. Oncogenic transcription factors in the human acute leukemias. *Science*, 278 (5340):1059–64, 1997. 129

Z. Luo, Y. Zhang, F. Li, J. He, H. Ding, L. Yan, and H. Cheng. Resistin induces insulin resistance by both ampk-dependent and ampk-independent mechanisms in hepg2 cells. *Endocrine*, 36(1):60–9, 2009. 24

A. D. Luster. Chemokines–chemotactic cytokines that mediate inflammation. *N Engl J Med*, 338(7):436–45, 1998. 98

A. Ma, R. Koka, and P. Burkett. Diverse functions of il-2, il-15, and il-7 in lymphoid homeostasis. *Annu Rev Immunol*, 24:657–79, 2006. 20

F. S. Machado, J. E. Johndrow, L. Esper, A. Dias, A. Bafica, C. N. Serhan, and J. Aliberti. Anti-inflammatory actions of lipoxin a4 and aspirin-triggered lipoxin are socs-2 dependent. *Nat Med*, 12(3):330–4, 2006. 20

M. Maconochie, R. Krishnamurthy, S. Nonchev, P. Meier, M. Manzanares, P. J. Mitchell, and R. Krumlauf. Regulation of hoxa2 in cranial neural crest cells involves members of the ap-2 family. *Development*, 126(7):1483–94, 1999. 205

S. Mahony and P. V. Benos. Stamp: a web tool for exploring dna-binding motif similarities. *Nucleic Acids Res*, 35(Web Server issue):W253–8, 2007. 159

R. S. Mann, K. M. Lelli, and R. Joshi. Hox specificity unique roles for cofactors and collaborators. *Curr Top Dev Biol*, 88:63–101, 2009. 129, 130, 140

E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends Genet*, 24(3):133–41, 2008. 131

F. Martinon and J. Tschopp. Nlrs join tlrs as innate sensors of pathogens. *Trends Immunol*, 26(8):447–54, 2005. 18, 25, 99, 100

F. Martinon, K. Burns, and J. Tschopp. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proil-beta. *Mol Cell*, 10(2): 417–26, 2002. 18

P. McCullagh and J. Nelder. Generalized linear models. *Chapman and Hall/CRC*, page 532, 1989. 176

R. C. McLeay and T. L. Bailey. Motif enrichment analysis: a unified framework and an evaluation on chip data. *BMC Bioinformatics*, 11:165, 2010. 198

H. Mikkers and A. Berns. Retroviral insertional mutagenesis: tagging cancer pathways. *Adv Cancer Res*, 88:53–99, 2003. 136

J. P. Mizgerd. Acute lower respiratory tract infection. *N Engl J Med*, 358(7):716–27, 2008. 172

A. Mood, F. Graybill, and D. Boes. Introduction to the theory of statistics. *McGraw Hill*, 0070428646:480, 1974. 202

J. J. Moskow, F. Bullrich, K. Huebner, I. O. Daar, and A. M. Buchberg. Meis1, a pbx1-related homeobox gene involved in myeloid leukemia in bxh-2 mice. *Mol Cell Biol*, 15 (10):5434–43, 1995. 129

T. Nakamura, D. A. Largaespada, M. P. Lee, L. A. Johnson, K. Ohyashiki, K. Toyama, S. J. Chen, C. L. Willman, I. M. Chen, A. P. Feinberg, N. A. Jenkins, N. G. Copeland, and Jr. Shaughnessy, J. D. Fusion of the nucleoporin gene nup98 to hoxa9 by the chromosome translocation t(7;11)(p15;p15) in human myeloid leukaemia. *Nat Genet*, 12(2):154–8, 1996. 129

T. Oda, T. Akaike, T. Hamamoto, F. Suzuki, T. Hirano, and H. Maeda. Oxygen radicals in influenza-induced pathogenesis and treatment with pyran polymer-conjugated sod. *Science*, 244(4907):974–6, 1989. 19, 26

K. Ogasawara, S. Hida, N. Azimi, Y. Tagaya, T. Sato, T. Yokochi-Fukuda, T. A. Waldmann, T. Taniguchi, and S. Taki. Requirement for irf-1 in the microenvironment supporting development of natural killer cells. *Nature*, 391(6668):700–3, 1998. 98

S. H. Orkin and L. I. Zon. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132(4):631–44, 2008. 195

K. Ozato, D. M. Shin, T. H. Chang, and 3rd Morse, H. C. Trim family proteins and their emerging roles in innate immunity. *Nat Rev Immunol*, 8(11):849–60, 2008. 87

P. Palese. Influenza: old and new threats. *Nat Med*, 10(12 Suppl):S82–7, 2004. 10, 170, 173

P. J. Park. Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–80, 2009. 131

G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole. Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res*, 32(Web Server issue):W199–203, 2004. 138, 198

M. S. Petrovick, S. W. Hiebert, A. D. Friedman, C. J. Hetherington, D. G. Tenen, and D. E. Zhang. Multiple functional domains of aml1: Pu.1 and c/ebpalpha synergize with different regions of aml1. *Mol Cell Biol*, 18(7):3915–25, 1998. 140

T. L. Phang, M. C. Neville, M. Rudolph, and L. Hunter. Trajectory clustering: a non-parametric method for grouping gene expression time courses, with applications to mammary development. *Pac Symp Biocomput*, pages 351–62, 2003. 174

L. C. Platanias, S. Uddin, A. Yetter, X. J. Sun, and M. F. White. The type i interferon

receptor mediates tyrosine phosphorylation of insulin receptor substrate 2. *J Biol Chem*, 271(1):278–82, 1996. 24

C. M. Pombo, J. V. Bonventre, A. Molnar, J. Kyriakis, and T. Force. Activation of a human ste20-like kinase by oxidant stress defines a novel stress response pathway. *Embo J*, 15 (17):4537–46, 1996. 19

J. Pothlichet, M. Chignard, and M. Si-Tahar. Cutting edge: innate immune response triggered by influenza a virus is negatively regulated by socs1 and socs3 through a rig-i/ifnar1-dependent pathway. *J Immunol*, 180(4):2034–8, 2008. 20

A. Prakash and M. Tompa. Discovery of regulatory elements in vertebrates through comparative genomics. *Nat Biotechnol*, 23(10):1249–56, 2005. 143

D. Proud, R. B. Turner, B. Winther, S. Wiehler, J. P. Tiesman, T. D. Reichling, K. D. Juhlin, A. W. Fulmer, B. Y. Ho, A. A. Walanski, C. L. Poore, H. Mizoguchi, L. Jump, M. L. Moore, C. K. Zukowski, and J. W. Clymer. Gene expression profiles during in vivo human rhinovirus infection: insights into the host response. *Am J Respir Crit Care Med*, 178(9):962–8, 2008. 11, 80

S. C. Pruitt, A. Bussman, A. Y. Maslov, T. A. Natoli, and R. Heinaman. Hox/pbx and brn binding sites mediate pax3 expression in vitro and in vivo. *Gene Expr Patterns*, 4(6): 671–85, 2004. 205

L. Pulliam, B. Sun, and H. Rempel. Invasive chronic inflammatory monocyte phenotype in subjects with high hiv-1 viral load. *J Neuroimmunol*, 157(1-2):93–8, 2004. 17

M. Quenouille. Notes on biase in estimation. *Biometrika*, 43:353–360, 1956. 201

O. Ramilo, W. Allman, W. Chung, A. Mejias, M. Ardura, C. Glaser, K. M. Wittkowski, B. Piqueras, J. Banchereau, A. K. Palucka, and D. Chaussabel. Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood*, 109(5):2066–77, 2007. 46, 106

J. Rissanen. Information and complexity in statistical modeling. *Springer*, page 144, 2007. 179

G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8):651–7, 2007. 131, 132, 147, 207

C. V. Rothlin, S. Ghosh, E. I. Zuniga, M. B. Oldstone, and G. Lemke. Tam receptors are pleiotropic inhibitors of the innate immune response. *Cell*, 131(6):1124–36, 2007. 20

Peter Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987. 36, 108

T. Rozovskaia, E. Feinstein, O. Mor, R. Foa, J. Blechman, T. Nakamura, C. M. Croce, G. Cimino, and E. Canaani. Upregulation of meis1 and hoxa9 in acute lymphocytic leukemias with the t(4 : 11) abnormality. *Oncogene*, 20(7):874–8, 2001. 129

J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nat Biotechnol*, 27(1):66–75, 2009. 131, 132, 143

L. Rui, M. Yuan, D. Frantz, S. Shoelson, and M. F. White. Socs-1 and socs-3 block insulin signaling by ubiquitin-mediated degradation of irs1 and irs2. *J Biol Chem*, 277(44): 42394–8, 2002. 24

A. Ryo, N. Tsurutani, K. Ohba, R. Kimura, J. Komano, M. Nishi, H. Soeda, S. Hattori, K. Perrem, M. Yamamoto, J. Chiba, J. Mimaya, K. Yoshimura, S. Matsushita, M. Honda, A. Yoshimura, T. Sawasaki, I. Aoki, Y. Morikawa, and N. Yamamoto. Socs1 is an inducible host factor during hiv-1 infection and regulates the intracellular trafficking and stability of hiv-1 gag. *Proc Natl Acad Sci U S A*, 105(1):294–9, 2008. 11, 80, 84, 99

A. Sabbah, T. H. Chang, R. Harnack, V. Frohlich, K. Tominaga, P. H. Dube, Y. Xiang, and S. Bose. Activation of innate immune antiviral responses by nod2. *Nat Immunol*, 10 (10):1073–80, 2009. 18, 19

C. E. Samuel, K. L. Kuhen, C. X. George, L. G. Ortega, R. Rende-Fournier, and H. Tanaka. The pkr protein kinase–an interferon-inducible regulator of cell growth and differentiation. *Int J Hematol*, 65(3):227–37, 1997. 26

K. S. Schluns and L. Lefrancois. Cytokine control of memory t-cell development and survival. *Nat Rev Immunol*, 3(4):269–79, 2003. 20

D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, and D. T. Odom. Chip-seq: using high-throughput sequencing to discover protein-dna interactions. *Methods*, 48(3):240–8, 2009. 131

K. B. Schwarz. Oxidative stress during viral infection: a review. *Free Radic Biol Med*, 21 (5):641–9, 1996. 26

Y. Seki, K. Hayashi, A. Matsumoto, N. Seki, J. Tsukada, J. Ransom, T. Naka, T. Kishimoto, A. Yoshimura, and M. Kubo. Expression of the suppressor of cytokine signaling-5 (socs5) negatively regulates il-4-dependent stat6 activation and th2 differentiation. *Proc Natl Acad Sci U S A*, 99(20):13003–8, 2002. 20

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, 2003. 187

W. F. Shen, J. C. Montgomery, S. Rozenfeld, J. J. Moskow, H. J. Lawrence, A. M. Buchberg, and C. Largman. Abdb-like hox proteins stabilize dna binding by the meis1 homeodomain proteins. *Mol Cell Biol*, 17(11):6448–58, 1997. 129, 130

W. F. Shen, S. Rozenfeld, A. Kwong, L. G. Kom ves, H. J. Lawrence, and C. Largman. Hoxa9 forms triple complexes with pbx2 and meis1 in myeloid cells. *Mol Cell Biol*, 19 (4):3051–61, 1999. 129, 130

H. Shin, T. Liu, A. K. Manrai, and X. S. Liu. Ceas: cis-regulatory element annotation system. *Bioinformatics*, 25(19):2605–6, 2009. 133

A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–50, 2005. 133

B. Silverman. Density estimation for statistics and data analysis. *Chapman and Hall*, 1986. 199

K. V. Sitwala, M. N. Dandekar, and J. L. Hess. Hox proteins and leukemia. *Int J Clin Exp Pathol*, 1(6):461–74, 2008. 129, 130

D. B. Stetson and R. Medzhitov. Type i interferons in host defense. *Immunity*, 25(3): 373–81, 2006. 10, 80

J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A*, 102(36):12837–42, 2005. 12, 30, 35, 36, 82, 107, 108, 173, 174, 176

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50, 2005. 31, 39, 174, 175, 187

J. C. Sun, S. M. Lehar, and M. J. Bevan. Augmented il-7 signaling during viral infection drives greater expansion of effector t cells but does not enhance memory. *J Immunol*, 177(7):4458–63, 2006. 20

A. Takaoka, H. Yanai, S. Kondo, G. Duncan, H. Negishi, T. Mizutani, S. Kano, K. Honda, Y. Ohba, T. W. Mak, and T. Taniguchi. Integral role of irf-5 in the gene induction programme activated by toll-like receptors. *Nature*, 434(7030):243–9, 2005. 17, 80, 98, 182

J. Taylor, S. Tyekucheva, D. C. King, R. C. Hardison, W. Miller, and F. Chiaromonte. Esperr: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res*, 16(12):1596–604, 2006. 134

R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, 2008. 187

M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–44, 2005. 138, 143, 198

J. Tukey. Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*, 29(2):614, 1958. 201

R. B. Turner. Ineffectiveness of intranasal zinc gluconate for prevention of experimental rhinovirus colds. *Clin Infect Dis*, 33(11):1865–70, 2001. 29, 34, 103, 105

R. B. Turner, J. O. Hendley, and Jr. Gwaltney, J. M. Shedding of infected ciliated epithelial cells in rhinovirus colds. *J Infect Dis*, 145(6):849–53, 1982. 79, 103

R. B. Turner, R. Bauer, K. Woelkart, T. C. Hulsey, and J. D. Gangemi. An evaluation of echinacea angustifolia in experimental rhinovirus infections. *N Engl J Med*, 353(4): 341–8, 2005. 49

V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–21, 2001. 174

G. Tuteja, P. White, J. Schug, and K. H. Kaestner. Extracting transcription factor targets from chip-seq data. *Nucleic Acids Res*, 37(17):e113, 2009. 131

D. A. Tyrrell, S. Cohen, and J. E. Schlarb. Signs and symptoms in common colds. *Epidemiol Infect*, 111(1):143–56, 1993. 110

S. Uddin, L. Yenush, X. J. Sun, M. E. Sweet, M. F. White, and L. C. Platanias. Interferon-

alpha engages the insulin receptor substrate-1 to associate with the phosphatidylinositol 3'-kinase. *J Biol Chem*, 270(27):15938–41, 1995. 24

S. Uddin, B. Majchrzak, P. C. Wang, S. Modi, M. K. Khan, E. N. Fish, and L. C. Platanias. Interferon-dependent activation of the serine kinase pi 3'-kinase requires engagement of the irs pathway but not the stat pathway. *Biochem Biophys Res Commun*, 270(1): 158–62, 2000. 24

E. R. Unanue. Viral infections and nonspecific protection–good or bad? *N Engl J Med*, 357(13):1345–6, 2007. 172

M Vandermeer, A Thomas, L Kamimoto, A Reingold, K Gershman, J Meek, M Farley, P Ryan, R Lynfield, J Baumbach, W Schaffner, and N Bennett. Role of statins in preventing death among patients hospitalized with lab-confirmed influenza infections. *Infectious Diseases Society of America Annual Meeting*, page 706, 2009. 26

A. Visel, M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin, and L. A. Pennacchio. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–8, 2009. 134

D. Wald, J. Qin, Z. Zhao, Y. Qian, M. Naramura, L. Tian, J. Towne, J. E. Sims, G. R. Stark, and X. Li. Sigirr, a negative regulator of toll-like receptor-interleukin 1 receptor signaling. *Nat Immunol*, 4(9):920–7, 2003. 85, 113

G. G. Wang, M. P. Pasillas, and M. P. Kamps. Meis1 programs transcription of flt3 and cancer stem cell character, using a mechanism that requires interaction with pbx and a novel function of the meis1 c-terminus. *Blood*, 106(1):254–64, 2005. 136

D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–34, 2009. 141, 144

K. Xu, M. Kliger, Y. Chen, P. Woolf, and A. Hero. Revealing social networks of spammers through spectral clustering. *Proc Int Conference on Communications*, 2009. 178, 179

M. Yamamoto, S. Sato, H. Hemmi, K. Hoshino, T. Kaisho, H. Sanjo, O. Takeuchi, M. Sugiyama, M. Okabe, K. Takeda, and S. Akira. Role of adaptor trif in the myd88-independent toll-like receptor signaling pathway. *Science*, 301(5633):640–3, 2003. 11, 80

H. Yasukawa, A. Sasaki, and A. Yoshimura. Negative regulation of cytokine signaling pathways. *Annu Rev Immunol*, 18:143–64, 2000. 19

M. Yoneyama, M. Kikuchi, T. Natsukawa, N. Shinobu, T. Imaizumi, M. Miyagishi, K. Taira, S. Akira, and T. Fujita. The rna helicase rig-i has an essential function in double-stranded rna-induced innate antiviral responses. *Nat Immunol*, 5(7):730–7, 2004. 11, 16, 80, 85

A. K. Zaas, M. Chen, J. Varkey, T. Veldman, 3rd Hero, A. O., J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams, N. C. Oien, B. Nicholson, S. Kingsmore, L. Carin, C. W. Woods, and G. S. Ginsburg. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host Microbe*, 6(3):207–17, 2009. 12, 17, 28, 46, 94, 181, 185, 186

X. Zhang, G. Robertson, M. Krzywinski, K. Ning, A. Droit, S. Jones, and R. Gottardo. Pics: Probabilistic inference for chip-seq. *Biometrics*, 2010. 131

Q. Zhou and W. H. Wong. Cismodule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, 101(33):12114–9, 2004. 198

Q. Zhu, C. Egelston, A. Vivekanandhan, S. Uematsu, S. Akira, D. M. Klinman, I. M. Belyakov, and J. A. Berzofsky. Toll-like receptor ligands synergize through distinct dendritic cell pathways to induce t cell responses: implications for vaccines. *Proc Natl Acad Sci U S A*, 105(42):16260–5, 2008. 11, 80, 84, 98

M. J. Zilliox, G. Parmigiani, and D. E. Griffin. Gene expression patterns in dendritic cells infected with measles virus compared with other pathogens. *Proc Natl Acad Sci U S A*, 103(9):3363–8, 2006. 21