Published in final edited form as:

Nat Immunol. 2014 February ; 15(2): 118–127. doi:10.1038/ni.2787.

Unifying immunology with informatics and multiscale biology

Brian A Kidd^{1,2,3}, Lauren A Peters^{3,4}, Eric E Schadt^{1,2,3}, and Joel T Dudley^{1,2,3}

¹Department of Genetics and Genomic Sciences, New York, New York, USA.

²Icahn Institute for Genomics and Multiscale Biology, New York, New York, USA.

³Icahn School of Medicine at Mount Sinai, New York, New York, USA.

⁴Graduate School of Biomedical Sciences, New York, New York, USA.

Abstract

The immune system is a highly complex and dynamic system. Historically, the most common scientific and clinical practice has been to evaluate its individual components. This kind of approach cannot always expose the interconnecting pathways that control immune-system responses and does not reveal how the immune system works across multiple biological systems and scales. High-throughput technologies can be used to measure thousands of parameters of the immune system at a genome-wide scale. These system-wide surveys yield massive amounts of quantitative data that provide a means to monitor and probe immune-system function. New integrative analyses can help synthesize and transform these data into valuable biological insight. Here we review some of the computational analysis tools for high-dimensional data and how they can be applied to immunology.

Substantial progress has been made in elucidating specific pathway constituents, interactions and mechanisms in the immune system. Understanding how immune cells and molecules interact with each other, the surrounding tissue architecture and more recently the microbiome, suggest many new important questions and research opportunities for immunologists. The potential to examine global cross-talk between pathways and cell populations is only just emerging. Advances in high-throughput profiling technologies—such as high-throughput genomic sequencing and mass cytometry (using the CyTOF mass cytometer)—enable comprehensive measurement of the immune system across multiple cellular components and time points. These technologies provide vast quantities of rich, high-dimensional data that capture system-wide properties at molecular and cellular resolution. Such measurements have greatly expanded the potential parameters to be analyzed and have increased the complexity of the mathematical models required for determining how immune processes operate and relate to various physiological conditions.

© 2014 Nature America, Inc. All rights reserved.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

Correspondence should be addressed to J.T.D. (joel.dudley@mssm.edu).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

The volume and complexity of these data necessitate computational tools and techniques to aid discovery and advance immunological research.

In this Review we focus on computational tools and methodologies for analyzing and integrating high-dimensional biomedical data relevant to understanding the organization, function and dynamics of the immune system, and its relevance to disease. We describe how integrative informatics and network biology techniques applied to large data sets can be used to elucidate complex immune-system states (see Box 1 for key terms). We discuss some of the most important challenges facing systems immunology and how computational tools can be applied to immunology to advance our understanding of how various functional molecular circuits interact in the immune system, and lay the groundwork for translating systems immunology data into clinical applications.

Immunological profiling

Two major tasks in immunology are to identify markers (for example, genes or proteins) or the functional characteristics that define various immune cell states or developmental stages and to determine how these components interact in a variety of circumstances. High-throughput molecular profiling technologies enable diverse strategies for investigating complex immune states.

Genome-wide transcriptional profiling is a systematic, unbiased approach to examine how transcript changes correlate with diverse states of the immune system. Hypothesis-free evaluation of these states by transcriptional profiling can be used to identify relationships that may have been more difficult to identify or even completely missed using more targeted approaches. Transcriptional profiles of immune-system cells have been used to develop molecular signatures for autoimmunity^{1–3}, to explore vaccine efficacy^{4–7}, to distinguish various phases of infection^{8–11} and to suggest new treatment options for patients with rheumatological disease¹² and lymphomas^{12,13}.

Population studies designed to determine the links between genotype and phenotype have uncovered numerous genetic variations that influence function of the immune system^{14,15}. A recent study identified 23 nucleotide variants from 13 genetic loci that regulate frequencies of immune-system cells¹⁶. To date, genome-wide association studies (GWAS) have linked more than 275 genetic loci with one or more autoimmune diseases¹⁷. Many of these loci form clusters of risk variants, as their gene products map to common biological pathways and suggest common molecular underpinnings in autoimmunity^{18–21}. Further high-resolution mapping of immune-system genes with high-throughput sequencing is likely to highlight underlying roles for both common and rare immune system-associated genetic variations across a broad number of diseases^{2,22}.

Human leukocyte antigen (HLA) genes are among the most polymorphic in the genome. Given the thousands of allelic variants and the exquisite specificity of antigen recognition, accurate assignment of HLA genotype is required for certain organ or bone marrow transplants. Switching from serological HLA typing to high-resolution HLA typing using high-throughput sequencing platforms has improved the donor-matching process and increased survival outcomes in recipients^{23,24}. Recently, a new sequencing strategy was

developed for high-throughput yet high-resolution HLA typing by deep sequencing²⁵. This technique is cost-effective, which increases the likelihood of its adoption in a clinical setting, and can be applied to other polymorphic genes. In addition to HLA genotyping for clinical applications, high-throughput sequencing can help (i) identify immunogenetic risk factors for disease²⁶, (ii) examine HLA polymorphism across population and pathogenic diversity²⁷ and (iii) select immunodominant HLA-restricted T cell epitopes (for example, for vaccine or tetramer design²⁸).

DNA sequencing technologies have also been successfully used to monitor the human immune system in multiple contexts including responses of vaccines^{29,30}, evolution of viral variants to escape immune detection³¹, diagnostics for leukemia³² and profiling of T cell antigen receptor (TCR)³³ or antibody repertoires³⁴. A general computational strategy for repertoire analysis involves alignment, clustering and phylogenetic tree construction techniques^{29,35}. This strategy requires robust computational algorithms to handle the massive data sets. One example of such a program is the open-source TCR repertoire analysis software MiTCR, which can efficiently process and analyze TCR sequences from hundreds of millions of raw high-throughput sequencing reads³⁶ (other programs are listed in Table 1). Repertoire analysis relies heavily on having good reference sequences for proper statistical comparisons and diversity estimates. The ImMunoGeneTics (IMGT) information database (<http://www.imgt.org/>) contains the largest collection of reference sequences for immunoglobulins, TCRs and major histocompatibility complexes (MHCs). This database provides a web portal for researchers to compare up to 450,000 of their measured sequences against a reference to identify germ-line rearrangements, track hypersomatic evolution in response to antigenic challenge and identify specific clones that might be associated with disease³⁷. The techniques and technical issues for high-throughput sequencing and their diagnostic applications for medicine are reviewed in ref. 38.

Functional and integrative genomic analysis

The investigation of genome-wide patterns of mRNA expression in cells of the immune system can identify drivers of the immune-system response to environmental factors (such as antigens, cytokines and small molecules) within and across cells and organisms. Extraction of RNA from cells followed by high-throughput sequencing (RNA-seq) is a powerful method to quantify transcriptomes and identify splice variants^{39,40}. Applied to immune cells, this technology has revealed the transcriptional profiles of primary B cells and monocytes⁴¹ as well as the genetic programming underlying helper T cell development⁴², T_H17 cell lineage entry⁴³ and dendritic cell maturation after stimulation with lipopolysaccharide⁴⁴. Several software packages exist that perform differential gene expression analysis of RNA-seq data^{40,45–48}. These packages follow a similar protocol, which includes normalizing the RNA count data, identifying differentially expressed genes and estimating false discovery rates. In addition, the programs Cufflinks and DEXSeq⁴⁹ test differential exon expression to examine splice variation, whereas ERANGE⁴⁰ analyzes single-nucleotide polymorphisms (SNPs).

Given that transcriptional profiling of microarray or RNA-seq data often yields many differentially expressed genes, enrichment analysis is a common computational approach

that helps with the functional interpretation of these gene lists. The popular gene-set enrichment analysis (GSEA) algorithm⁵⁰—as well as extensions of the technique⁵¹—evaluates ranked gene lists for statistical enrichment of genes involved in defined pathways and cellular processes (for example, cytokine signaling, antigen presentation and inflammatory response). This technique has become a useful tool for immunological discovery with the proliferation of publicly available data⁵². For example, the Molecular Signatures Database (MSigDB⁵³) contains thousands of gene sets (for example, biological pathways and oncogene signatures) gathered from analyses of transcriptional profiling experiments that researchers can use to annotate and interpret results from expression data, including an ‘immunological signatures’ gene set collection from data provided by the Human Immunology Project Consortium. Two popular packages that implement GSEA include GenePattern⁵⁴ and database for annotation, visualization and integrated discovery (DAVID)^{55,56}. A software package called Enricher extends the capabilities of GSEA by offering an interactive application with visualization tools that facilitate collaborative analysis through the web or mobile devices⁵⁷.

One approach for understanding genetic drivers of immune-system processes and related diseases is to measure DNA variation and transcriptional profiles from the same sample and then examine the statistical relationships between DNA variation and gene-expression traits. Expression quantitative trait loci (eQTL) analysis is a robust statistical technique that matches variation in a quantitative trait (for example, mRNA expression) to DNA variation at specific genomic loci. eQTL relationships suggest DNA-RNA regulatory interactions. However, genome-wide analysis can require statistical testing of billions of possible eQTL associations (transcript to SNP), which is computationally intensive and necessitates efficient analytical tools. One software package for such analysis, MatrixEQTL⁵⁸ (others are listed in Table 1), uses efficient mathematical operations in modern computer programs, achieving 2–3 orders of magnitude faster run times over other software tools (for example, a difference between ten minutes and ~3.5 days for a moderately sized data set).

eQTL analysis provides a means to understand immunological processes and prioritize key drivers of disease. System-wide genetic profiling (for example, GWAS) of immune traits often identifies associations with numerous genomic loci. A recent GWAS analysis that looked at the largest collection of patients with inflammatory bowel disease identified more than 70 new loci associated with Crohn’s disease or ulcerative colitis⁵⁹. In that study, eQTL associations were used to help prioritize causal genes in loci associated with inflammatory bowel disease. eQTL analysis has also been recently used to identify the downstream regulatory targets of expression-modifying SNPs associated with systemic lupus erythematosus and type 1 diabetes⁶⁰. Such integrative analysis highlights the pathways and common connections in complex diseases. Finally, the genotype-tissue expression (GTEx) project is building a massive biospecimen repository to establish tissue-specific eQTL profiles for more than 40 tissues across roughly 1,000 samples⁶¹. This resource will be an invaluable tool for identifying eQTLs conserved across cell types and tissues, and also for identifying tissue-specific, and eventually cell type-specific, eQTLs that could clarify how SNPs alter expression or binding of specific transcription factors and influence epigenetic regulation.

Cytometry

Single-cell mass cytometry (using the CyTOF mass cytometer) has increased the amount of phenotypic characterization that can be obtained from populations of cells⁶². Currently, ~40 surface markers or intracellular proteins can be measured simultaneously, providing an unprecedented number of possible parameters to characterize different cell types (for example, 190 to 780 biaxial plots for 20 to 40 markers, respectively). High-throughput cytometric data now require automated strategies to process the high-dimensional information and identify cell types based on the extent of similarity among a specified set of markers. Numerous programs exist for automated cytometry analysis⁶³. One software program capable of analyzing mass and flow cytometry data is spanning-tree progression analysis of density-normalized events (SPADE)⁶⁴. SPADE clusters cells with similar marker intensities and connects clusters into a tree structure, which compresses the multidimensional cytometric data into two dimensions for visualization and enables discovery of new cell subsets and states. This algorithm was applied to reconstruct the hematopoietic lineage and monitor cell phenotype and signaling changes after perturbations with cytokines or small molecules⁶².

Automated approaches should ideally retain the single-cell resolution while using an unbiased, data-driven modality to select the most informative combinations of markers to determine distinct cell types or classify samples while filtering out artifacts⁶³. Following this strategy, principal-component analysis was used to identify combinations of cytokines present in different T cell subsets⁶⁵. In that work, subsets were displayed on a two-dimensional plot and color-coded based on antigen specificity, which revealed that viral specificities colocalized to different locations (i.e., differing subsets) in this plot⁶⁵. A nonlinear dimensionality-reduction technique was applied to classify cancerous and noncancerous bone marrow samples, providing a mechanism to identify abnormal cells after treatment as a means to identify minimal residual disease⁶⁶. These computational techniques reduce complexity in visualizing multidimensional data and can highlight differences in cell subsets. Two areas for which there are unmet needs for high-throughput cytometric techniques are (i) computational methods that integrate cytometric data with other genome-wide profiling and (ii) statistical methods for data transformation and hypothesis testing between samples.

Deconvolving cell-specific profiles

Ideally, each cell population of interest would be carefully sorted and transcriptionally profiled, but this is not always feasible because of protocol, sample or resource constraints. Computational methods can be used to estimate cell-specific transcriptional profiles from genome-wide expression data measured from heterogeneous samples (for example, whole-blood gene expression values), using measured cell-type proportions such as those obtained from a complete blood count or flow cytometry analysis. Regression-based approaches can estimate cell type-specific expression values from measured proportions and identify cell type-specific gene expression profiles⁶⁷. Publicly available data (for example, Gene Expression Omnibus or GTEX) can provide cell type-specific gene sets that can then be used to estimate the proportions of cell subsets in a tissue sample. As demonstrated recently,

this approach could be used to characterize lymphocytic infiltrate in a tumor or to approximate the contribution of different cell types in inflamed tissue⁶⁸.

Informatics approaches to integrate molecular data sets

Informatics approaches that integrate high-throughput data sets across multiple time points or cell types have been used to map the topological organization of immune-system networks and elucidate the molecular mechanisms that regulate immune function. Chromatin immunoprecipitation followed by high-throughput genomic sequencing (ChIP-seq) was combined with RNA-seq to measure the regulatory programming of transcriptional and epigenetic markers that drive early T cell development⁴². Multiple studies have explored T helper cell lineage differentiation⁶⁹. For example, time-course data from genome-wide transcriptional profiling were integrated with publicly available data of transcription-factor binding sites to generate a regulatory network model of T_H17 cell differentiation⁴³. This model identified novel control elements and two opposing subnetworks that regulate T_H17 cellular fate. Similarly, ChIP-seq was combined with RNA-seq and transcriptional-profiling (microarray) platforms to construct a predictive regulatory network model for T_H17 cells⁷⁰. Combining these data recovered known T_H17 cell genes and uncovered novel regulatory interactions that control T_H17 cell plasticity⁷⁰. The primary data, resulting networks and analytical tools from this effort are available to the scientific community (<http://th17.bio.nyu.edu/>).

Additionally, methods that integrate gene variants associated with disease in population-based studies, genomic occupancy of transcription factors and genome-wide expression profiling have been used to uncover potential functional links between genetic polymorphisms and phenotypic disease traits. For example, ~800 disease-associated SNPs were linked with NF- κ B DNA-binding profiles from ten individuals and mRNA expression profiles to connect variation in regulatory elements and gene expression levels with disease⁷¹. Such multi-data set integration provides a method to compare transcription-factor binding across individuals and led to the finding that inflammatory disease-associated SNPs are enriched at the NF- κ B binding site⁷¹. Two key challenges for connecting information from genomic variation to allele-specific binding and expression data include constructing a model of a personal diploid genome sequence and then correlating RNA-seq and ChIP-seq data onto the sequence. The software tools AlleleSeq⁷² and the ACT⁷³ (aggregation and correlation toolbox) offer solutions to the challenges of coordinating and comparing genomic, transcriptomic and ChIP-seq data.

Network biology and building global models for immunology

One aim of systems immunology is to construct comprehensive, multiscale network models that accurately capture all of the regulatory elements in the immune system^{74,75}. Network models can be generated by a number of different methods, but they all have similar features. These mathematical models represent maps for potential connections among intracellular and intercellular components, suggesting new functional roles for specific genes, proteins or metabolites (Fig. 1). For immunologists already focused on a particular molecule of interest, networks will potentially place the molecule in the context of new

pathways, molecular interactions and/or even an unanticipated tissue or disease link, potentially leading to specific hypotheses to be tested in experimental models.

The procedure for constructing a network uses statistical techniques to estimate the most likely relationships between molecular entities (for example, transcription factors and noncoding RNAs and mRNAs)⁷⁶. After a preliminary network model is constructed, predicted connections are perturbed to test and validate the model through *in vitro* assays⁷⁷. Applications of network modeling to immunology have uncovered several known mechanisms and identified new mechanisms in regulation of immune system cells. Network modeling has provided high-resolution maps of the regulatory circuitry that control differentiation of hematopoietic cell lineage⁷⁸, differentiation of T_H17 cells⁴³ and the transcriptional regulators for 249 cell types of the mouse immune system⁷⁹. Furthermore, viral-sensing transcriptional networks in dendritic cells have been constructed for pathogen recognition through signaling of Toll-like receptors^{80,81}.

Weighted gene coexpression network analysis can depict the system-wide influence of mRNA expression as well. Constructing this type of network requires soft thresholding of correlations in defining connections between two nodes⁸². This analysis can provide data-defined groups of genes (modules) that offer a new modality for biological interpretability^{59,83–85}. Coexpression network analysis determines modules by organizing genes that correlate more with one another than with other sets of genes across individual samples within an analysis group. Coexpression network analysis can also be used as a framework to evaluate a common denominator in a meta-analysis of patient studies or even cross-species disease models⁸⁶. Likewise, differential connectivity can be used to analyze altered correlation in comparing two coexpression networks from different groups (for example, healthy versus disease)⁸⁴.

Probabilistic causal networks are becoming increasingly more prevalent in modeling the immune system. This computational technique integrates genomic variation (for example, genome-wide SNPs) with quantitative expression levels of mRNA (for example, transcriptional profiling of patient tissues) to infer causal directionality of correlated genes in the networks⁸⁷. Causal regulatory networks are constructed using conditional correlations inherent in eQTLs. ‘Priors’ such as data collected about known protein-protein interactions, transcription-factor binding sites and other molecular relationships can also aid in inference of causality from correlation. Normal genetic variation observed in a population offers a natural perturbation to the immune system from which to construct predictive network models for human disease. Bayesian methods are used to infer causal relationships between molecular interactions by randomly generating many possible network models and using statistical techniques to select a consensus model that best fits the data. A causal regulatory network constructed from multiple tissue samples from 1,000 obese patients undergoing gastric bypass surgery was previously described⁸⁸. The resulting network model identified a macrophage-biased metabolic network enriched for genes associated with a number of other autoimmune diseases. This computational approach has been applied to data collected from multiple tissue types. The analyses have revealed a number of new genetic associations for type 2 diabetes⁸⁹, identified key drivers in inflammatory disease⁸⁶ and uncovered a new gene (*TYROBP*) associated with Alzheimer’s disease⁸⁴.

To explore how an immunological network can bridge therapeutic targets with disease pathophysiology, we selected more than 6,800 drugs (US Food and Drug Administration–approved small molecules and biologics as well as experimental compounds) curated from DrugBank⁹⁰ and display connections where specific and/or nonspecific drug targets converge with genes featured in the network. In addition, we mapped diseases from GWAS to their respective genes on the network (Fig. 2). To refine the drug to immune system–related genes to disease axis, we selected connections between drugs and diseases that were mediated through a common gene or two (Fig. 2a). Although the network is at molecular resolution such that single drugs map to individual genes and distinct diseases, we aggregated drugs, genes and diseases into higher-level categories based on anatomical therapeutic chemical classification, Gene Ontology enrichment terms and disease group, respectively (Fig. 2b).

This approach illustrates how genes in an immunological network connect diverse diseases and provides a window into potential therapeutic targets for using existing drugs to potentially treat diseases both related and unrelated to the immune system. A detailed map of the interactions between drug targets and what immune-system cell types are most likely to express these targets remains largely unexplored and thus offers opportunities for improved understanding through computational approaches. One such opportunity is to leverage transcriptional profiles from cells perturbed with small molecules *in vitro* to repurpose drugs as more effective therapeutics for immune system–mediated diseases⁹¹. Using this approach, the antiseizure drug topiramate was shown to be a potential candidate for the treatment of inflammatory bowel disease⁹². A further step is to integrate this computational framework with network modeling as a general strategy for personalized medicine, as the ability to sequence a tumor cell’s genome may suggest alternative drugs or immune system–targeted therapy tailored specifically to an individual (Fig. 3).

Representing immune network models

Software packages that read network models and display the web of connections in an interactive mode are essential tools for augmenting analysis, integrating additional biological or clinical information in a network and interpreting the myriad nonlinear relationships between data sets. The most common computational tool for visualizing and analyzing molecular network data is Cytoscape⁹³. Cytoscape is an open-source platform, which means anyone can develop computational tools to work within the Cytoscape framework⁹⁴. Hundreds of software tools have been developed to enhance the features and interoperability of Cytoscape, and more packages are continuously being added. For example, PanGIA is a tool for integrating physical and genetic interaction data into hierarchical maps to infer functional relationships between data sets⁹⁵. Software to enhance the features and operability of Cytoscape are in the hundreds, and more are in development.

Human immunology and public repositories

Although mathematical models of the immune system will require further development and computing power to appreciably mimic the dynamics happening *in vivo*, recent systems approaches have shown great progress^{3–8}. The Human Immunology Project Consortium

(HIPC; <http://www.immuneprofilng.org/>) is a large-scale initiative by the US National Institute of Allergy and Infectious Diseases to generate a sizeable repository of immunological data (as of 30 September 2013 more than 16,000 subjects from 69 studies and over 250,000 results across multiple data types such as enzyme-linked immunosorbent assay, enzyme-linked immunospot, flow cytometry and gene expression) and create a computational interface for researchers to use this resource (<https://import.niaid.nih.gov/>). As additional methods and data sets become available, more researchers will be able to leverage these resources to address specific hypotheses or construct new models that will improve our understanding of immunological processes and ultimately suggest better strategies for addressing the biggest challenges facing clinical and basic immunology.

Challenges for the future

Two major challenges for immune monitoring are to determine baseline immunological states from molecular profiles, despite inherent heterogeneity both over time within an individual and across populations, and to associate these molecular states with clinical outcomes or disease. Progress on these issues was made through the integration of multiple data sources collected in medium-sized cohorts (<http://www.immuneprofilng.org/>). However, bigger sample sizes from diverse demographics are needed to define normal ranges and corresponding deviations that result in clinically meaningful conclusions. Large-scale consortia offer one mechanism for building up sample sizes, and leveraging public data can provide another such mechanism.

There is a need for tools that do not require specialized training to allow investigators to build and explore integrated network models. Although advances in computing algorithms and infrastructure offer a range of solutions for analyzing large-scale data sets^{96–98}, the computational models constructed by biomedical engineers require proficiency in computer programming, which often precludes the intended end users, such as biologists and clinicians, from using and integrating these technologies with scientific and clinical practice. To address this problem, we need interactive software solutions that allow researchers to access information in computational models at appropriate levels of detail. An unresolved issue is how to visualize and mine high-dimensional data in ways that help determine how complex systems operate. When examining ‘omics’ data, researchers routinely suffer from ‘data fatigue’. Visualization tools are needed that can display data through multiple perspectives, which could then be cycled through rapidly and intuitively⁹⁹. Tools such as Iris¹⁰⁰ and Cytoscape (others are listed in Table 1) are important steps toward satisfying these needs.

In this process, the key is to annotate networks with specific pathway or cell-type associations and with a possibility to move to alternative scale network views (for example, proteomics) seamlessly. A tool that enables clicking on genes of interest for synthesis of related literature and other forms of regulation (for example, epigenetic or miRNA-mediated) would be useful, as would flexible network models that synthesize data from parent models and represent molecular interactions present in phenotypic or clinical data clusters sourced from medical records, histologic data and magnetic resonance images or other clinical images.

High-throughput technologies such as high-throughput sequencing and mass cytometry are enabling the generation of massive amounts of data, which require computational tools to process, analyze and visualize. Some suggestions for how to use existing tools to support immunological discovery are provided in Box 2. Accurate and comprehensive mathematical models of complex immune system dynamics can be enhanced through computational approaches that integrate multiple layers of profiling data. Building more complete and predictive models of drug-immune system interactions will be useful for enabling more safe and effective drug discovery and development. Network models and analytics will provide the foundation for understanding how to modulate the immune system and develop diagnostic and prognostic indicators for disease. Immunologists and clinicians are now working with systems biologists to take advantage of numerous computational tools to generate and analyze predictive models of immune function, which can be applied to improve biological understanding of the immune system and provide more effective patient care.

ACKNOWLEDGMENTS

We thank C. Berin, B. Brown, R. Kosoy, B. Readhead and C. Tato for critical reading and feedback on the manuscript. This work was supported by funding from the National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK098242) and the Pharmaceutical Research and Manufacturers of America Foundation.

References

1. Pascual V, Chaussabel D, Banchereau J. A genomic approach to human autoimmune diseases. *Annu. Rev. Immunol.* 2010; 28:535–571. [PubMed: 20192809]
2. Boisson B, et al. Immunodeficiency, autoinflammation and amylopectinosis in humans with inherited HOIL-1 and LUBAC deficiency. *Nat. Immunol.* 2012; 13:1178–1186. [PubMed: 23104095]
3. Chaussabel D, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity.* 2008; 29:150–164. [PubMed: 18631455]
4. Querec TD, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat. Immunol.* 2009; 10:116–125. [PubMed: 19029902]
5. Nakaya HI, et al. Systems biology of vaccination for seasonal influenza in humans. *Nat. Immunol.* 2011; 12:786–795. [PubMed: 21743478]
6. Furman D, et al. Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. *Mol. Syst. Biol.* 2013; 9:659. [PubMed: 23591775]
7. Obermoser G, et al. Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. *Immunity.* 2013; 38:831–844. [PubMed: 23601689]
8. Berry MP, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature.* 2010; 466:973–977. [PubMed: 20725040]
9. Cliff JM, et al. Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. *J. Infect. Dis.* 2013; 207:18–29. [PubMed: 22872737]
10. Bloom CI, et al. Detectable changes in the blood transcriptome are present after two weeks of antituberculosis therapy. *PLoS ONE.* 2012; 7:e46191. [PubMed: 23056259]
11. Law GL, Korth M, Benecke A, Katze M. Systems virology: host-directed approaches to viral pathogenesis and drug targeting. *Nat. Rev. Microbiol.* 2013; 11:455–466. [PubMed: 23728212]
12. Chiche L, Jourde-Chiche N, Pascual V, Chaussabel D. Current perspectives on systems immunology approaches to rheumatic diseases. *Arthritis Rheum.* 2013; 65:1407–1417. [PubMed: 23450649]

13. Hummel M, et al. A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N. Engl. J. Med.* 2006; 354:2419–2430. [PubMed: 16760442]
14. Casanova J-L, Abel L, Quintana-Murci L. Immunology taught by human genetics. *Cold Spring Harb. Symp. Quant. Biol.* 2013; 4:a007260.
15. Xavier RJ, Rioux JD. Genome-wide association studies: a new window into immune-mediated diseases. *Nat. Rev. Immunol.* 2008; 8:631–643. [PubMed: 18654571]
16. Orrù V, et al. Genetic variants regulating immune cell levels in health and disease. *Cell.* 2013; 155:242–256. [PubMed: 24074872]
17. Visscher PM, Brown M, McCarthy M, Yang J. Five years of GWAS discovery. *Am. J. Hum. Genet.* 2012; 90:7–24. [PubMed: 22243964]
18. Cho JH, Gregersen P. Genomics and the multifactorial nature of human autoimmune disease. *N. Engl. J. Med.* 2011; 365:1612–1623. [PubMed: 22029983]
19. Cotsapas C, et al. Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* 2011; 7:e1002254. [PubMed: 21852963]
20. Goris A, Liston A. The immunogenetic architecture of autoimmune disease. *Cold Spring Harb. Perspect. Biol.* 2012; 4:a007260. [PubMed: 22383754]
21. Voight BF, Cotsapas C. Human genetics offers an emerging picture of common pathways and mechanisms in autoimmunity. *Curr. Opin. Immunol.* 2012; 24:552–557. [PubMed: 23041452]
22. Bolze A, et al. Ribosomal protein SA haploinsufficiency in humans with isolated congenital asplenia. *Science.* 2013; 340:976–978. [PubMed: 23579497]
23. Flomenberg N, et al. Impact of HLA class I and class II high-resolution matching on outcomes of unrelated donor bone marrow transplantation: HLA-C mismatching is associated with a strong adverse effect on transplantation outcome. *Blood.* 2004; 104:1923–1930. [PubMed: 15191952]
24. Spellman SR, et al. A perspective on the selection of unrelated donors and cord blood units for transplantation. *Blood.* 2012; 120:259–265. [PubMed: 22596257]
25. Wang C, et al. High-throughput, high-fidelity HLA genotyping with deep sequencing. *Proc. Natl. Acad. Sci. USA.* 2012; 109:8676–8681. [PubMed: 22589303]
26. Trowsdale J. The MHC, disease and selection. *Immunol. Lett.* 2011; 137:1–8. [PubMed: 21262263]
27. Prugnolle F, et al. Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* 2005; 15:1022–1027. [PubMed: 15936272]
28. Newell EW, et al. Combinatorial tetramer staining and mass cytometry analysis facilitate T-cell epitope mapping and characterization. *Nat. Biotechnol.* 2013; 31:623–629. [PubMed: 23748502]
29. DeKosky BJ, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* 2013; 31:166–169. [PubMed: 23334449]
30. Jiang N, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* 2013; 5:171ra19.
31. Wu X, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science.* 2011; 333:1593–1602. [PubMed: 21835983]
32. Wu D, et al. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci. Transl. Med.* 2012; 4:134ra63.
33. Robins HS, et al. Overlap and effective size of the human CD8⁺ T cell receptor repertoire. *Sci. Transl. Med.* 2010; 2:47ra64.
34. Boyd SD, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* 2009; 1:12ra23.
35. Jiang N, et al. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci. USA.* 2011; 108:5348–5353. [PubMed: 21393572]
36. Bolotin DA, et al. MiTCR: software for T-cell receptor sequencing data analysis. *Nat. Methods.* 2013; 10:813–814. [PubMed: 23892897]
37. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc M-P. IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Res.* 2012; 882:569–604.

38. Boyd SD. Diagnostic applications of high-throughput DNA sequencing. *Annu. Rev. Pathol.* 2013; 8:381–410. [PubMed: 23121054]
39. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 2009; 10:57–63. [PubMed: 19015660]
40. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods.* 2008; 5:621–628. [PubMed: 18516045]
41. Fairfax BP, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* 2012; 44:502–510. [PubMed: 22446964]
42. Zhang JA, Mortazavi A, Williams B, Wold B, Rothenberg E. Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell.* 2012; 149:467–482. [PubMed: 22500808]
43. Yosef N, et al. Dynamic regulatory network controlling TH17 cell differentiation. *Nature.* 2013; 496:461–468. [PubMed: 23467089]
44. Shalek AK, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature.* 2013; 498:236–240. [PubMed: 23685454]
45. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 2010; 28:511–515. [PubMed: 20436464]
46. Robinson MD, McCarthy D, Smyth G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010; 26:139–140. [PubMed: 19910308]
47. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* 2010; 11:R106. [PubMed: 20979621]
48. Li J, Witten D, Johnstone I, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics.* 2012; 13:523–538. [PubMed: 22003245]
49. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* 2012; 22:2008–2017. [PubMed: 22722343]
50. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA.* 2005; 102:15545–15550. [PubMed: 16199517]
51. Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann. Appl. Stat.* 2007; 1:107–129.
52. Haining WN, Wherry EJ. Integrating genomic signatures for immunologic discovery. *Immunity.* 2010; 32:152–161. [PubMed: 20189480]
53. Liberzon A, et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* 2011; 27:1739–1740. [PubMed: 21546393]
54. Reich M, et al. GenePattern 2.0. *Nat. Genet.* 2006; 38:500–501. [PubMed: 16642009]
55. Huang W, Sherman B, Lempicki R. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 2009; 4:44–57. [PubMed: 19131956]
56. Huang W, Sherman B, Lempicki R. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009; 37:1–13. [PubMed: 19033363]
57. Chen EY, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics.* 2013; 14:128. [PubMed: 23586463]
58. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics.* 2012; 28:1353–1358. [PubMed: 22492648]
59. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature.* 2012; 491:119–124. [PubMed: 23128233]
60. Westra H-J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* 2013; 45:1238–1243. [PubMed: 24013639]
61. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 2013; 45:580–585. [PubMed: 23715323]
62. Bendall SC, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science.* 2011; 332:687–696. [PubMed: 21551058]

63. Aghaeepour N, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods*. 2013; 10:228–238. [PubMed: 23396282]
64. Qiu P, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* 2011; 29:886–891. [PubMed: 21964415]
65. Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. *Immunity*. 2012; 36:142–152. [PubMed: 22265676]
66. Amir AD, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* 2013; 31:545–552. [PubMed: 23685480]
67. Shen-Orr SS, et al. Cell type-specific gene expression differences in complex tissues. *Nat. Methods*. 2010; 7:287–289. [PubMed: 20208531]
68. Ahn J, et al. DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*. 2013; 29:1865–1871. [PubMed: 23712657]
69. Vahedi G, et al. Helper T-cell identity and evolution of differential transcriptomes and epigenomes. *Immunol. Rev.* 2013; 252:24–40. [PubMed: 23405893]
70. Ciofani M, et al. A validated regulatory network for Th17 cell specification. *Cell*. 2012; 151:289–303. [PubMed: 23021777]
71. Karczewski KJ, et al. Systematic functional regulatory assessment of disease-associated variants. *Proc. Natl. Acad. Sci. USA*. 2013; 110:9607–9612. [PubMed: 23690573]
72. Rozowsky J, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* 2011; 7:522. [PubMed: 21811232]
73. Jee J, et al. ACT: aggregation and correlation toolbox for analyses of genome tracks. *Bioinformatics*. 2011; 27:1152–1154. [PubMed: 21349863]
74. Arazi A, Pendergraft W, Ribeiro R, Perelson A, Hachohen N. Human systems immunology: hypothesis-based modeling and unbiased data-driven approaches. *Semin. Immunol.* 2013; 25:193–200. [PubMed: 23375135]
75. Germain RN, Meier-Schellersheim M, Nita-Lazar A, Fraser I. Systems biology in immunology: a computational modeling perspective. *Annu. Rev. Immunol.* 2011; 29:527–585. [PubMed: 21219182]
76. Segal E, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 2003; 34:166–176. [PubMed: 12740579]
77. Amit I, Regev A, Hachohen N. Strategies to discover regulatory circuits of the mammalian immune system. *Nat. Rev. Immunol.* 2011; 11:873–880. [PubMed: 22094988]
78. Novershtern N, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*. 2011; 144:296–309. [PubMed: 21241896]
79. Jovic V, et al. Identification of transcriptional regulators in the mouse immune system. *Nat. Immunol.* 2013; 14:633–643. [PubMed: 23624555]
80. Amit I, et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*. 2009; 326:257–263. [PubMed: 19729616]
81. Chevrier N, et al. Systematic discovery of TLR signaling components delineates viral-sensing circuits. *Cell*. 2011; 147:853–867. [PubMed: 22078882]
82. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 2005; 4 Article17.
83. Voineagu I, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*. 2011; 474:380–384. [PubMed: 21614001]
84. Zhang B, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell*. 2013; 153:707–720. [PubMed: 23622250]
85. Chen Y, et al. Variations in DNA elucidate molecular networks that cause disease. *Nature*. 2008; 452:429–435. [PubMed: 18344982]
86. Wang IM, et al. Systems analysis of eleven rodent disease models reveals an inflammatoric signature and key drivers. *Mol. Syst. Biol.* 2012; 8:594. [PubMed: 22806142]
87. Schadt EE, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 2005; 37:710–717. [PubMed: 15965475]

88. Greenawalt DM, et al. A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res.* 2011; 21:1008–1016. [PubMed: 21602305]
89. Zhong H, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet.* 2010; 6:e1000932. [PubMed: 20463879]
90. Wishart DS, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 2006; 34:D668–D672. [PubMed: 16381955]
91. Sirota M, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 2011; 3:96ra77.
92. Dudley JT, et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* 2011; 3:96ra76.
93. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003; 13:2498–2504. [PubMed: 14597658]
94. Saito R, et al. A travel guide to Cytoscape plugins. *Nat. Methods.* 2012; 9:1069–1076. [PubMed: 23132118]
95. Srivas R, et al. Assembling global maps of cellular function through integrative analysis of physical and genetic networks. *Nat. Protoc.* 2011; 6:1308–1323. [PubMed: 21886098]
96. Schadt EE, Linderman M, Sorenson J, Lee L, Nolan G. Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.* 2010; 11:647–657. [PubMed: 20717155]
97. Dudley JT, Butte A. In silico research in the era of cloud computing. *Nat. Biotechnol.* 2010; 28:1181–1185. [PubMed: 21057489]
98. Dudley JT, Pouliot Y, Chen R, Morgan AA, Butte AJ. Translational bioinformatics in the cloud: an affordable alternative. *Genome Med.* 2010; 2:51. [PubMed: 20691073]
99. Kotecha N, Krutzik P, Irish J. Web-based analysis and publication of flow cytometry experiments. *Curr. Protoc. Cytom.* 2010; 53:10.17.
100. Lum PY, et al. Extracting insights from the shape of complex data using topology. *Sci. Rep.* 2013; 3:1236. [PubMed: 23393618]
101. Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics.* 2013; 29:542–550. [PubMed: 23303508]
102. Warren RL, Nelson B, Holt R. Profiling model T-cell metagenomes with short reads. *Bioinformatics.* 2009; 25:458–464. [PubMed: 19136549]
103. Barak M, Zuckerman N, Edelman H, Unger R, Mehr R. IgTree: creating Immunoglobulin variable region gene lineage trees. *J. Immunol. Methods.* 2008; 338:67–74. [PubMed: 18706908]
104. Glanville J, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. USA.* 2009; 106:20216–20221. [PubMed: 19875695]
105. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods.* 2009; 6:S22–S32. [PubMed: 19844228]
106. Koboldt DC, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012; 22:568–576. [PubMed: 22300766]
107. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 2011; 43:491–498. [PubMed: 21478889]
108. Li H, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25:2078–2079. [PubMed: 19505943]
109. Guttman M, et al. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 2010; 28:503–510. [PubMed: 20436462]
110. Langmead B, Schatz M, Lin J, Pop M, Salzberg S. Searching for SNPs with cloud computing. *Genome Biol.* 2009; 10:R134. [PubMed: 19930550]
111. Wilbanks EG, Facciotti M. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE.* 2010; 5:e11471. [PubMed: 20628599]
112. Ji H, et al. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* 2008; 26:1293–1300. [PubMed: 18978777]

113. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008; 9:R137. [PubMed: 18798982]
114. Rozowsky J, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* 2009; 27:66–75. [PubMed: 19122651]
115. Kharchenko PV, Tolstorukov M, Park P. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* 2008; 26:1351–1359. [PubMed: 19029915]
116. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 2007; 81:559–575. [PubMed: 17701901]
117. Broman KW, Wu H, Sen S, Churchill G. R/qtl: QTL mapping in experimental crosses. *Bioinformatics.* 2003; 19:889–890. [PubMed: 12724300]
118. Clayton D, Leung H-T. An R package for analysis of whole-genome association studies. *Hum. Hered.* 2007; 64:45–51. [PubMed: 17483596]
119. Watson M. CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics.* 2006; 7:509. [PubMed: 17116249]
120. Bonneau R, et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol.* 2006; 7:R36. [PubMed: 16686963]
121. Basso K, et al. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 2005; 37:382–390. [PubMed: 15778709]
122. Aghaeepour N, Nikolic R, Hoos H, Brinkman R. Rapid cell population identification in flow cytometry data. *Cytometry.* 2011; 79A:6–13. [PubMed: 21182178]
123. Pyne S, et al. Automated high-dimensional flow cytometric data analysis. *Proc. Natl. Acad. Sci. USA.* 2009; 106:8519–8524. [PubMed: 19443687]
124. Qian Y, et al. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry Clin. Cytom.* 2010; 78V(suppl. 1):S69–S82.
125. Zare H, Shoostari P, Gupta A, Brinkman R. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics.* 2010; 11:403. [PubMed: 20667133]
126. Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* 2009; 19:1639–1645. [PubMed: 19541911]
127. Lefranc MP, et al. IMGT, the international ImmunoGeneTics database. *Nucleic Acids Res.* 1999; 27:209–212. [PubMed: 9847182]
128. Siebert JC, Munsil W, Rosenberg-Hasson Y, Davis M, Maecker H. The Stanford Data Miner: a novel approach for integrating and exploring heterogeneous immunological data. *J. Transl. Med.* 2012; 10:62. [PubMed: 22452993]
129. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000; 155:945–959. [PubMed: 10835412]
130. Kumar S, Sanderford M, Gray V, Ye J, Liu L. Evolutionary diagnosis method for variants in personal exomes. *Nat. Methods.* 2012; 9:855–856. [PubMed: 22936163]
131. Li MJ, et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* 2012; 40:D1047–D1054. [PubMed: 22139925]
132. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 2012; 40:D930–D934. [PubMed: 22064851]
133. Boyle AP, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 2012; 22:1790–1797. [PubMed: 22955989]
134. Raychaudhuri S, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* 2009; 5:e1000534. [PubMed: 19557189]
135. Griffith M, et al. DGIdb: mining the druggable genome. *Nat. Methods.* 2013; 10:1209–1210. [PubMed: 24122041]
136. Engreitz JM, et al. ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression. *Bioinformatics.* 2011; 27:3317–3318. [PubMed: 21967760]

137. Heiser LM, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. USA*. 2012; 109:2724–2729. [PubMed: 22003129]
138. Tan CM, Chen E, Dannenfelser R, Clark N, Ma'ayan A. Network2Canvas: network visualization on a canvas with enrichment analysis. *Bioinformatics*. 2013; 29:1872–1878. [PubMed: 23749960]

Key terms

State. Collection of molecular parameters (for example, transcription levels of a gene and protein states) that describe the configuration of an immune cell or system.

High-dimensional data. Data set that includes many variables or factors (for example, a microarray is a collection of mRNA expression data on thousands of genes, i.e., 'dimensions').

Informatics. Field that stores, processes, analyzes and communicates information.

Systems immunology. Field that aims to integrate how all the components (molecules, cells and tissues) interact to maintain immunessystem function.

Multiscale. Diverse data sets that span different locations, sizes (for example, molecules, cells or tissues) or time points.

Data-driven. Knowledge and models learned from patterns in the data rather than a preconception or a prior hypothesis.

Bayesian network. A network that captures causal relationships between variables or nodes of interest (for example, transcription levels of a gene, protein states, etc.). Bayesian networks enable the incorporation of prior information in establishing relationships between nodes.

Omics. Collection of all the parts (e.g., genes, proteins, metabolites) and their interactions.

Suggestions for analyzing high-dimensional data

Below are some ‘rules of thumb’ for computational analysis of data generated by high-throughput technologies.

The examples provided suggest one computational tool, but in many cases, a number of alternatives exist (see Table 1).

Genetic variation

Evaluate allele frequency spectra in the context of population structure and local ancestry using tools such as Structure (<http://pritchardlab.stanford.edu/structure.html>) and PCAdmix (<https://sites.google.com/site/pcadmixon/home/>). Evaluate loci and alleles for evolutionary expectations of functional importance using tools such as myPEG (<http://www.megasoftware.net/myPEG/myPEG.php>).

Use a resource like GWASdb (<http://jjwanglab.org/gwasdb/>) to map known genotype-phenotype associations. Use HaploReg (<http://www.broadinstitute.org/mammals/haploreg/haploreg.php>) or RegulomeDB (<http://regulomedb.org/>) to integrate with functional genomics data sets (for example, encyclopedia of DNA elements (ENCODE) and eQTL reference databases) to understand how genetic variation might lead to functional differences in expression or regulation. Explore gene-set or pathway enrichment in genomic loci using GRAIL (gene relationships across implicated loci; <http://www.broadinstitute.org/mpg/grail/>). Look for genetic variation in druggable targets using the drug-gene interaction database (<http://dgidb.genome.wustl.edu/>).

Transcriptomics

If starting with RNA-seq, process the data with TopHat (<http://tophat.cbcb.umd.edu>) and Bowtie (<http://bowtie-bio.sourceforge.net>). Use cufflinks tools to call differential expression and to call variants from RNA reads. If the study sample size is more than 30, consider building gene coexpression networks using an R package for weighted-correlation network analysis (WGCNA).

If the same study incorporates genetic variation data or genetic variants are called from RNA-seq reads, then compute eQTL associations using MatrixEQTL. If the study sample size is >200, consider constructing Bayesian networks using the software package for reconstructing integrative molecular Bayesian networks (RIMBANet). Evaluate differentially expressed genes or subnetwork modules for functional enrichment using tools such as GSEA (<http://www.broadinstitute.org/gsea/index.jsp>) or ProfileChaser (<http://profilechaser.stanford.edu/>). Search for drugs that might shift the transcriptional profile in the gene sets or pathways of interest using the LINCS (library of integrated network-based cellular signatures) browser tools and databases (<http://lincs.hms.harvard.edu/explore/pathway/> and http://lincs.hms.harvard.edu/explore_/canvasbrowser/).

Cytometry

Although manual gating is standard practice for identifying cell populations from flow cytometry data, consider using an automated gating strategy such as FlowMeans (many

other options are listed in Table 1) when the sample size increases beyond 40. If starting with mass cytometry, explore the relationships among the cells in this high-dimensional data by reducing the dimensionality and visualizing the cells on a two-dimensional plot using viSNE or SPADE.

Figure 1.

Integrating biological data from multiple sources to construct regulatory network models. State-of-the-art computational techniques combine information from large-scale public data sets with omics measurements collected from the samples under study to generate multiscale causal models. Whenever possible, samples should be collected from multiple tissues or states in an individual (for example, diseased and healthy) at several time points. Measurements from single tissues miss important context-specific regulatory interactions that are responsible for disease. Single time points (cross-sectional studies) fail to capture the system dynamics and often require larger sample sizes to detect intersample variability. WGS, whole-genome sequencing. Ab, antibody.

Figure 2.

Identifying drugs to treat diseases by using networks. **(a)** A drug can target the product of a gene within the network and influence disease if the drug acts directly on the gene product from the disease-associated gene (left) or by modulating a gene product that then influences the gene associated with disease (right). **(b)** Schematic network of drug-immune system-disease interactions. Individual genes from a Bayesian causal regulatory network (macrophage-enriched) have been aggregated into modules using Gene Ontology (GO) terms and are shown schematically as light blue circles. Diseases associated with genes

through GWAS are shown as triangles and connected to a gene module if the SNP resides on one of the genes inside it. Drugs from DrugBank have been organized into anatomical therapeutic chemical classes, and are displayed as colored squares. Categories are connected to gene modules if at least one gene product is the known target for one of the drugs in the class. Size of each shape is proportional to the number of elements (genes or drugs) it contains.

Figure 3.

Constructing causal regulatory networks to understand the immunological basis of disease and advance precision medicine. **(a)** Cohorts of patients provide data that lead to tissue-specific and cell type-specific networks for health and various diseases. These regulatory networks allow for cross-condition comparisons to understand the molecular basis of immunological diseases. **(b)** Measurements collected from omics networks can be used to estimate the dynamic range for each molecular measurement and to develop the parameter landscape for various healthy and diseased conditions. **(c)** Data collected from individual patients can be projected onto molecular networks and parameter landscapes to construct a personal profile for informing medical decisions.

Table 1

Analysis and visualization tools for systems immunology

Data types or tools	Program name	Weblink
TCR repertoire	MiTCR ³⁶	http://mitcr.milaboratory.com/
	Decombinator ¹⁰¹	https://github.com/uclinfectionimmunity/Decombinator
	iSSAKE ¹⁰²	ftp://ftp.bcgsc.ca/supplementary/iSSAKE/
	IMGT or HIGHV-Quest ³⁷	http://www.imgt.org/HighV-QUEST/
Antibody repertoire	IMGT or HIGHV-Quest ³⁷	http://www.imgt.org/HighV-QUEST/
	IgTree ¹⁰³	http://immsilico2.lnx.biu.ac.il/Software.html
	VDJFasta ¹⁰⁴	http://sourceforge.net/projects/vdjfasta/
RNA-seq ¹⁰⁵	VarScan ¹⁰⁶	http://varscan.sourceforge.net/
	GATK ¹⁰⁷	http://www.broadinstitute.org/gatk/
	SAMtools ¹⁰⁸	http://samtools.sourceforge.net/
	ERANGE ⁴⁰	http://woldlab.caltech.edu/rnaseq
	Scripture ¹⁰⁹	http://www.broadinstitute.org/software/scripture/
	Cufflinks ⁴⁵	http://cufflinks.cbc.umd.edu/
	CuffDiff ⁴⁵	http://cufflinks.cbc.umd.edu/
	EdgeR ⁴⁶	http://www.bioconductor.org/packages/2.12/bioc/html/edgeR.html
	DESeq ⁴⁹	http://www.bioconductor.org/packages/release/bioc/html/DESeq.html
	Myrna ¹¹⁰	http://bowtie-bio.sourceforge.net/myrna/index.shtml
	PoissonSeq ⁴⁸	http://cran.r-project.org/web/packages/PoissonSeq/index.html
	ERANGE ⁴⁰	http://woldlab.caltech.edu/rnaseq
ChIP-seq ^{105,111}	CisGenome ¹¹²	http://www.biostat.jhsph.edu/~hji/cisgenome/
	MACS ¹¹³	http://liulab.dfci.harvard.edu/MACS/
	PeakSeq ¹¹⁴	http://info.gersteinlab.org/PeakSeq/
	SPP ¹¹⁵	http://compbio.med.harvard.edu/Supplements/ChIP-seq/
	MatrixEQTL ⁵⁸	http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/
eQTL	PLINK ¹¹⁶	http://pngu.mgh.harvard.edu/~purcell/plink/
	R/qtl ¹¹⁷	http://www.rqtl.org/
	snpMatrix ¹¹⁸	http://www.bioconductor.org/packages/2.3/bioc/html/snpMatrix.html
Networks	WGCNA ⁸²	http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/
	coXpress ¹¹⁹	http://coxpress.sourceforge.net/
	Inferelator ¹²⁰	http://bonneaulab.bio.nyu.edu/networks.html
	ARACNE ¹²¹	http://wiki.c2b2.columbia.edu/workbench/index.php/ARACNE/
Cytometry ⁶³	RimbaNET ⁸⁷	http://icahn.mssm.edu/departments-and-institutes/genomics/about/software/rimbanet/
	FlowMeans ¹²²	http://www.bioconductor.org/packages/release/bioc/html/flowMeans.html
	FLAME ¹²³	http://www.broadinstitute.org/cancer/software/genepattern/modules/FLAME/
	FLOCK ¹²⁴	https://import.niaid.nih.gov/importWeb/analysis/flow/displayOverview.do?method=displayFlockOverview
	SamSPECTRAL ¹²⁵	http://www.bioconductor.org/packages/devel/bioc/html/SamSPECTRAL.html

Data types or tools	Program name	Weblink
Visualization	SPADE ⁶⁴	http://www.bioconductor.org/packages/release/bioc/html/spade.html http://cytospade.org/
	viSNE ⁶⁶	http://www.c2b2.columbia.edu/danapeerlab/html/cyt.html
	Cytoscape ⁹³	http://www.cytoscape.org/
	Gephi	https://gephi.org/
	Circos ¹²⁶	http://circos.ca/
Databases or online tools	ImMunoGeneTics ¹²⁷	http://www.imgt.org/
	ImmPort	https://immpart.niaid.nih.gov/
	MsigDB ⁵³	http://www.broadinstitute.org/gsea/msigdb/index.jsp
	GenePattern ⁵⁴	http://www.broadinstitute.org/cancer/software/genepattern/
	DAVID ^{55,56}	http://david.abcc.ncifcrf.gov/
	Stanford Data Miner ¹²⁸	
	Structure ¹²⁹	http://pritchardlab.stanford.edu/structure.html
	PCAdmix	https://sites.google.com/site/pcadmixon/home
	myPEG ¹³⁰	http://www.megasoftware.net/myPEG/myPEG.php
	GWASdb ¹³¹	http://jjwanglab.org/gwasdb/
	HaploReg ¹³²	http://www.broadinstitute.org/mammals/haploreg/haploreg.php
	RegulomeDB ¹³³	http://regulomedb.org/
	GRAIL ¹³⁴	http://www.broadinstitute.org/mpg/grail/
	DGIdb ¹³⁵	http://dgidb.genome.wustl.edu/
	GSEA ⁵⁰	http://www.broadinstitute.org/gsea/index.jsp
	ProfileChaser ¹³⁶	http://profilechaser.stanford.edu/
	LINCS browser ^{137,138}	http://lincs.hms.harvard.edu/explore/pathway/ http://lincs.hms.harvard.edu/explore/_/canvasbrowser/