Integrative Network-based Association Studies: Leveraging cell regulatory models in the post-GWAS era.

Atul J. Butte[1,2], Andrea Califano[3,4,5†], Stephen Friend[6], Trey Ideker[7,8,9], Eric Schadt[6,10]

[1]Department of Pediatrics and Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA

[2]Lucile Packard Children's Hospital, Stanford University, Stanford, CA, USA

[3]Columbia Initiative in Systems Biology, Columbia University, New York, NY, USA

[4]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA

[5]Department of Biomedical Informatics, Columbia University, New York, NY, USA

[6]Sage Bionetworks, Seattle, WA, USA

[7]Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA.

[8]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA.

[9]The Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA 92093, USA.

[10]Pacific Biosciences of California, Menlo Park, CA, USA.

[†]Corresponding author

Abstract

After completion of a number of large scale Genome-Wide Association Studies (GWAS), there is still a significant amount of trait and disease variance that cannot be explained by existing genetic variability. This review introduces new, Integrative Network-based Association Study (INAS) approaches that aim to minimize the impact from multiple hypothesis testing statistics, thus allowing the identification of rare variants/alterations and epistatic interactions. In particular we discuss methods that rely on the de novo computational, experimental, and integrative dissection of context specific molecular interaction networks (or interactomes, for short). We provide several examples of how these approaches may be used to tackle discovery of genetic variants and somatic alterations causally related to the presentation of specific traits and diseases. We also discuss how more complex systems, including a variety of non-cell-autonomous traits and diseases will require new multicellular networks that explicitly represent short distance paracrine and long distance endocrine interactions.

Introduction

Over the last ten years, the genome wide study of both heritable and somatic human variability has gone from a theoretical concept to a broadly implemented, practical reality, covering the entire spectrum of human diseases: from cancer to obesity to neurodegenerative disorders. While a number of exciting findings have emerged from these studies[1], the result of such genome wide association studies (GWAS) has been for the most part sobering. For instance, although several genes displaying medium to high penetrance within heritable traits have been inferred by these approaches for certain conditions, other diseases are still missing identification of much of the genetic risk[2-7], and few epistatic interactions or low penetrance genes have been identified due to impractical requirements for cohort sizes[8] as well as a lack of methodological developments that maximize power for such detections[9]. At the other end of the spectrum, the extensive somatic genomic rearrangements observed in solid tumors[10] yield such a broad range of candidate alterations that distinguishing ‘driver’ from ‘passenger’ alterations is difficult.

This begs the question of whether, in a post-GWAS era, existing GWAS datasets

may still hold a trove of hidden value. It has been suggested, for instance, that GWAS data could be more insightful when studied integratively within the context of other data modalities. Indeed, a number of previous studies have integrated significant genotype-phenotype associations with databases of gene annotations, such as the Gene Ontology[11], MSigDB[12], or the Kyoto Encyclopedia of Genes and Genomes[13]. The goal of these studies is to recognize higher-order structure within the data through aggregation of loci that encode genes with similar functions or that are in the same pathway.

A particularly important framework for the integration of genomic, metabonomic, and proteomic data is provided by the context-specific networks of molecular interactions that determine cell behavior. The basic hypothesis is straightforward. Among the entire spectrum of genetic and epigenetic variants, those contributing to a specific trait or disease must have some broad coalescent properties, allowing their effect to be functionally canalized via the cell regulatory machinery or via the cell-communication machinery that allows distinct cell types to interact. Thus, if a comprehensive and accurate map of all intra and inter-cellular molecular interactions were available, then genetic and epigenetic events implicated in a specific trait or disease should cluster within sets of closely interacting genes, within the cell's regulatory network.

Two approaches are then possible. First, if the regulatory networks determining the cell pathophysiological behavior were known *a priori*, e.g. a canonical cancer or functional pathway, one could systematically reduce the number of statistical tests for association between genetic or epigenetic variations and the trait or disease of interest by considering only events that form significant clusters within regulatory networks. This is because events that are closer in the regulatory topology of the cell are more likely to produce related phenotypic effects. Such a Pathway-Wide Association Study (PWAS) strategy[14] may improve our ability to distinguish signals from background noise by mitigating the magnitude of the multiple hypothesis testing correction. In most cases, unfortunately, the set of molecular interactions or pathways necessary to present a trait or a disease-related phenotype are not well characterized at the molecular level. Indeed the entire classical notion of relatively linear and interpretable disease pathways may need to be revisited in light of the dynamic, multi-scale, and context-specific complexity of gene regulatory networks. Thus, a second approach requires the simultaneous reconstruction of

both context-specific (and possibly multi-cell) gene regulatory networks and of the genetic and epigenetic events they harbor. We shall call this second strategy Integrative Network-based Association Studies or INAS and suggest that INAS will become increasingly dominant as the dynamic and cell-context specific nature of gene regulatory networks is further elucidated.

In this perspective, we explore current advances in PWAS and INAS research, the natural corollaries of a regulatory-network-oriented view of traits and disease, and future directions that are being pursued within the emerging community of Systems Genetics. We will explore how networks (and pathway motifs within them) can be reconstructed and validated and how they may provide a valuable integrative framework to interpret GWAS as well as other genetic and epigenetic variability data.

THIS IS NOT MY BEAUTIFUL PATHWAY

An increasing body of evidence suggests that canonical pathways may be woefully inadequate to represent and model the complex interplay of signal transduction, transcriptional, post-transcriptional, metabolic, and other regulatory events that ultimately determines cellular behavior. Rather, they satisfy our need to interpret biological

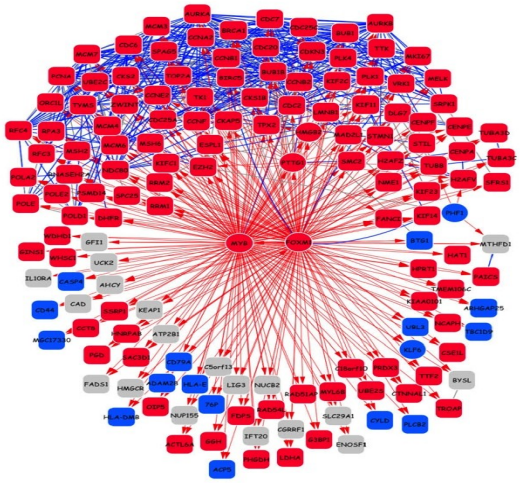

Figure 1a: FOXM1 and MYB co-regulation network from the Human B Cell Interactome[15]. Red and blue represents gene over and under expression, respectively, in germinal centers. Blue arcs represent protein-protein interactions.

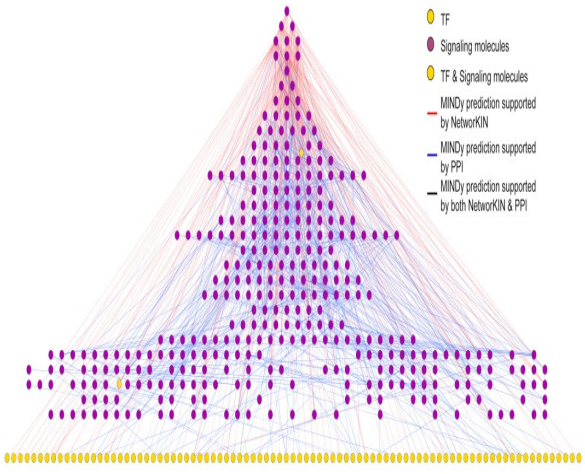

Figure 1b. Visualization of the signalome-transfactome molecular interaction network in human B cells[16].

phenomena as linear chains of events or other simple pathway models that can be easily visually interpreted. Unfortunately, this is a dramatic oversimplification. Biological processes and their regulatory control are anything but linear and cellular processes are determined by complex, multivariate interactions that cannot be visually interpreted. Both require the power of computational modeling to yield valuable biological insight. For instance, it has been shown that individual transcription factors regulate hundreds to thousands of highly cell-context dependent target genes. Indeed, functional specificity is achieved opportunistically by combinatorial interactions between multiple transcription factors. For instance, while FOXM1 and MYB individually regulate transcription of more than a thousand distinct genes, the about roughly 100 targets they co-regulate are exquisitely specific to human B cells during germinal center formation, see Fig. 1a. Conversely, those regulated by either one independently have a wide range of functions and are not specifically differentially expressed[15]. Similarly, transcription factor activity is modulated by hundreds of signal transduction proteins[16], whose availability is again context specific. Fig. 1b, for instance, shows a map of all transcription factors and of their computationally inferred upstream modulators in a human B cell. Many of these interactions have been experimentally validated with low false positive rates, indicating that such a level of complexity is realistic. Additionally, recent large-scale screens for protein-protein interactions in human cells[20] suggest that their number are orders of magnitude larger than the few thousand captured in canonical pathways. Clearly, the concept of a relatively small number of hierarchical and relatively independent signal transduction pathways is not reconcilable with these observations. Finally, adding yet another level of complexity, causal dependencies between the genetic, regulatory, and functional layers provide insight into the mechanisms by which rare germline allele variants and somatic alterations may impact the activity of entire constellations of transcription factors, which in turn regulate thousands of genes , see Fig. 2.

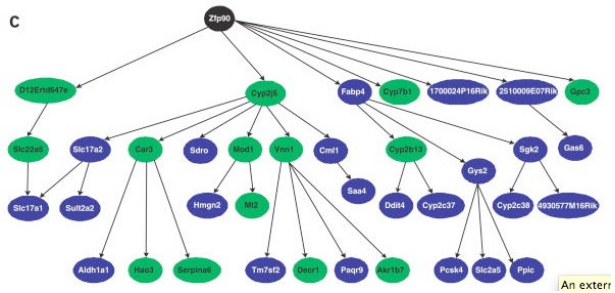

Figure 2. Genetic subnetwork controlled by *Zfp90* (black node) as a central node in the liver transcriptional network. This subnetwork was obtained from a full liver expression network by identifying all nodes that were descended from and within a path length of 3 of the *Zfp90* node. Nodes highlighted in green represent genes testing as causal for fat mass.

As discussed, such intrinsic complexity is made even more

daunting by the exquisite cell-context specific nature of the cell regulatory machinery. In other words, what we learn in one cellular phenotype may not hold true for another. For instance, it is well known that the oncogenic nature of genetic lesions is cell context specific and that depending on the microenvironment signals specific lesions in the same cell may or may not be tumorigenic[27]. Indeed, the paracrine and endocrine molecular interactions that allow distinct cell types and even whole organs to communicate form the highest order networks in a living organism that directly define its physiological and pathological states. Thus, while canonical pathways may provide useful conceptual tools, they cannot recapitulate the full complexity of cellular regulation. To be truly predictive and informative, cellular networks must be reconstructed *de novo* within each cellular phenotype context of interest. Further, we can distinguish between cell autonomous and non cell autonomous processes in thinking about more predictive biological networks. Whereas cancer may make a reasonably good cell autonomous system (given when you profile cancer you get both stroma and cancer at the same time), common non-cancer human diseases like obesity and type II diabetes can result from a failure in multiple organ systems including the central nervous system and tissues involved in partitioning and disposal of nutrients, and so may be best modeled as a non-cell autonomous system. In fact, we have shown that interaction networks constructed between tissues like hypothalamus and adipose tissue collected from an experimental cross population segregating obesity and type II diabetes, may be specific to cross tissue interactions[28]. That is, some subnetworks identified in cross-tissue interaction networks are not visible within single tissue networks, exhibiting a degree of regulation that may go beyond simple cell autonomous systems. Molecular networks constructed from heterogeneous tissues have also exhibited extraordinary context sensitivity, with interactions among different cell types making up a given organ specific to functions associated with that organ. In a model for type II diabetes, molecular interactions between different tissues were observed to be more abundant than interactions within any given tissue (or cell type)[29], and insulin signaling in osteoblasts has been shown to be necessary for whole-body glucose homeostasis[30]. These examples highlight that molecular networks capable of predicting whole system behavior will require modeling approaches that go beyond cellular networks, requiring the explicit representation of interactions at a hierarchy of

scales that provide a path to define the molecular interactions that define physiological states related to disease phenotypes[31].

REVERSE ENGINEERING CELLULAR NETWORKS

Just a few years ago, determining and experimentally validating a single kinase substrate or transcription factor target required a year or more of bench work. Since regulatory and protein-complex networks in eukaryotes appear to be highly complex –hundreds of thousands of interactions, – context-specific[34], and dynamic, how can one possibly reconstruct them in sufficient depth and with sufficient accuracy? Indeed, imagine having not only to elucidate hundreds of thousands of these interactions but also having to assess how they may change and reorganize themselves, under multivariate control, within distinct cellular phenotypes and possibly under distinct stimuli. It is precisely out of this necessity that the fields of high-throughput computational and experimental reverse engineering have blossomed. This is an important and timely effort. Ultimately, our success in elucidating disease related mechanisms on a rational, predictive basis will depend on our ability to use stochastic and kinetic models to accurately map cell regulatory networks and to predict their response to pathophysiological stimuli.

On the experimental side, large-scale, high-throughput efforts have started to release enormous amounts of raw data over the last five years. These data can be used as a scaffold for the assembly of entire regulatory and protein-complex networks, thus providing insight into the architecture of the cell in terms of how direct interactions between molecules may allow assembly of protein complexes, transduction of signals, and control of the transcriptional machinery of the cell[37]. For example, networks of protein–protein interactions in human cells have been assembled using yeast two-hybrid (Y2H) technology or tandem affinity purification coupled with mass spectrometry (TAP–MS)[20]. Similarly, candidate transcriptional targets of specific transcription factors (protein–DNA interactions) have been mapped using the techniques of chromatin immunoprecipitation coupled with DNA microchips (ChIP–chip)[38] or sequencing (ChIP–PET)[39], DNA adenine methylase identification (DamID)[40], or yeast one-hybrid assays[41]. Physical interactions can also be measured *in vitro* using DNA or protein arrays, which have been used to identify transcription factor binding sites and the substrates of kinases[44]. While interactions characterized by high-throughput experimental methods

generally have high false positives and false negative rates and are unlikely to generalize to cellular contexts other than the one in which they were ascertained, they nonetheless provide an initial if sparse snapshot of regulatory networks, especially when integrated with other types of data that can help filter the interactions most appropriate to a given context[25]. By mapping and interpreting changes among snapshots in different contexts we can begin to create a more comprehensive scaffold.

Complementing and extending high-throughput experimental assays, computational reverse-engineering algorithms have recently achieved accuracy and sensitivity comparable with their experimental counterpart, at a fraction of their cost and time requirements. Computational methods for reverse-engineering cellular networks were first developed for the study of prokaryotes and lower eukaryotes[45-47] and have more recently become highly successful in reconstructing the transcriptional[32], post-translational, post-transcriptional[50], metabolic[51], and protein-complex[15] logic of human cells, as well as of their dependence on the genetic information encoded in the DNA molecule, thus paving the road to the regulatory network based study of human disease. Among recent approaches, there has been significant success in using integrative approaches to combine both multiple clues as well as multiple layers of regulation within cellular networks.

Computational methods all rely, in one way or another, on measuring changes in distinct molecular moieties (e.g., mRNAs, microRNAs, proteins, etc.) as a response to either endogenous or exogenous perturbations. The former include, for instance, differences in kinetic constants induced by the genotypic variability between different individuals or the different spectrum of genetic lesions associated with a particular tumor phenotype[54]. The latter include small-molecule[60], RNAi, and environmental perturbations[61], such as differences in temperature, nutrients, or culture serum, among many others. In fact, several methods have been published that specifically use perturbations to infer regulatory networks or to interrogate them to infer drug sensitivity[63], resistance[64], and mechanism of action.

Finally, meta-network information, highlighting functional rather than physical interactions, is provided by genetic interactions, which chart the combinatorial relationships among genes in control of a common phenotype. Genetic interactions are

identified by comparing the phenotypic effect of disrupting (or overexpressing) a gene individually to the effect of disrupting two or more genes in combination[65]. For example, 'synthetic sickness' (or in the extreme 'synthetic lethality') is a genetic interaction in which disrupting both genes has a far more deleterious effect than expected from either disruption alone. 'Epistasis' is a genetic interaction in which one gene disruption masks the phenotypic effect of the other. In model organisms such as yeast, large networks of genetic interactions are being measured through systematically-applied combinations of gene knockouts[66]. In higher eukaryotes including worms, flies, and humans, genetic interactions are presently being explored through the technique of combinatorial RNAi[65] and other RNAi-based screening approaches[67]. Importantly, synthetic-sick and epistatic interactions are also prevalent in GWAS, in which genotypes at multiple loci come together to exert combinatorial control over the phenotypic trait. Alternatively, epistasis can occur when one locus with strong individual linkage to the phenotype is modified by the presence of another 'genetic modifier'. In the absence of prior information, however, de-novo identification of epistatic interactions in GWAS is greatly limited by lack of statistical power, although emerging methods are beginning to address this limitation.

EXAMPLES OF PWAS AND NBAS APPROACHES

In the following, we discuss several approaches that have been successful in identifying genes that are critically involved in the presentation of a phenotype, due to either genetic alteration or functional dysregulation. This list, rather than being comprehensive, is intended to illustrate different approaches in both PWAS and NBAS

Canonical Pathway Analysis: Canonical pathways are compact representations of the knowledge accumulated in a large number of manuscripts and supported by experimental assays about the relationship between multiple proteins, usually in the context of a specific biological process, e.g. embryogenesis, apoptosis, or tumorigenesis. While the knowledge represented within canonical pathways is likely incomplete and may lack context specificity, it does represent an important collection of molecular interactions that have previously resulted in the elucidation of key biological mechanisms.

For instance, integration of NF-κB pathway and targets analysis with GWAS data

from a large collection of Diffuse Large B Cell Lymphoma (DLBCL) samples was successful in the identification of this gene as the key integrator of a spectrum of upstream genetic alterations characterizing the more aggressive ABC subtype of the disease from its GCB counterpart[71]. These included several genes in the BCR and other signal transduction pathways, such as CARD11, A20, TRAF2, TRAF5, TAK1, and RANK, among others. Surprisingly, while Nf-κB itself was not genetically altered in the ABC subtype, it was shown to constitute a key non-oncogene addiction for ABC-DLBCL cells.

There have also been attempts to create more informative pathways by automated data-mining of literature data, using machine-learning approaches. These more complex and non human-interpretable networks have been used to cluster information coming from disease-related human variability data, such as for instance in the study of genetic predisposition to several human diseases[72].

Integrative genomics: There is already a rich literature on methods for using cellular networks, including protein-protein and protein-DNA interaction networks, to interpret gene expression profiles, with the goal of identifying network “hot spots” or “expression-activated modules”. Expression-activated modules are sets of proteins enriched for both interaction and coexpression across several conditions; they provide an important means of distilling the thousands of interactions present in a typical molecular network to arrive at a smaller number of discrete modules of activity. As recent examples, DEGAS (DysrEgulated Gene set Analysis via Subnetworks) and IDEA (Interactome Dysregulation Enrichment Analysis) represent methods for identifying connected gene subnetworks significantly enriched for genes that are dysregulated in specimens of a disease or following a chemical perturbation. In Parkinson's disease, DEGAS found novel evidence for involvement of mRNA splicing, cell proliferation, and the 14-3-3 complex in the disease progression, while in B cell lymphoma, IDEA identified genetic alterations in Chronic Lympocytic Leukemia and Follicular Lymphoma.

In parallel, a set of related methods have been developed for integration of protein networks with the results of genome wide linkage and association studies. For instance, Lage et al.[77] searched for protein complexes whose genes were associated with similar

phenotypes, using a human protein–protein interaction network integrating both human interactions and interactions from model organisms. Proteins were ranked by the phenotype similarity score of their associated diseases and of those of their direct network neighbors. In dmGWAS[78], dense subnetworks of protein-protein interactions were tested for the enrichment of genes harboring low *p*-value SNPs from GWAS studies. Compared with other pathway-based approaches, the method introduces flexibility in defining a gene set through use of local protein-protein interaction information.

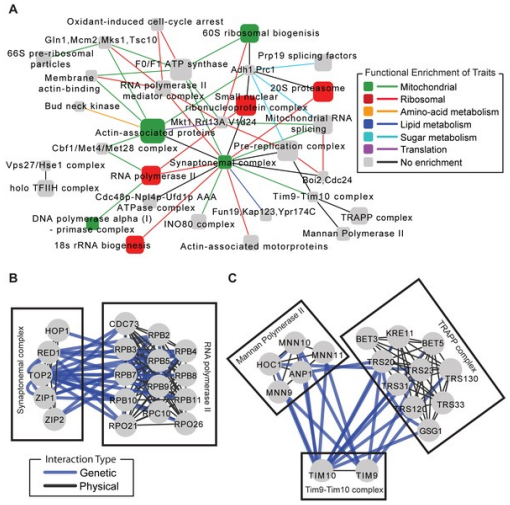

Figure 3. Genetic networks extracted GWAS elucidate pathway architecture. (A) A global map of the top GWAS genetic interactions between protein interaction complexes. Each node represents a protein complex and each interaction represents a significant number of genetic interactions. Node sizes are proportional to the number of proteins in the complex. (B,C) Two specific examples of protein complexes spanned by dense bundles of genetic interactions mined from the GWAS data.

A similar approach integrated a large-scale human protein-protein interaction network and a set of genes linked to ataxia to demonstrate a potential gain in statistical power[37]. Further integrative genomics attempts to boost statistical power to identify genetic interactions in GWAS included the use of pairs of SNP markers whose combined state was associated with the phenotype[79]. A bi-clustering method was used to group SNP-SNP interactions into interactions between larger genomic regions, i.e., reinforced by interactions involving multiple SNPs, which were then projected on a protein-protein interaction network, see Fig. 3. The analysis showed that genetic interactions uncovered by GWAS were strongly enriched within and between complexes of protein-protein interaction. A novel discovery from this GWAS-based method was that the INO80 chromatin remodeling complex has functional links to transcription elongation via RNA polymerase II and vacuolar protein degradation. Finally, related approaches were developed for using prior knowledge to infer epistatic interactions from

GWAS data[38].

Genetical Genomics: Another class of integrative genomics approaches used to reverse engineer regulatory networks is systems genetics, a broad class of approaches that integrate naturally occurring genetic variation with large-scale molecular phenotypic and higher order phenotype data (e.g., clinical trait data) to infer causal relationships among genes and between genes and phenotypes of interest. Variations in DNA can directly affect protein function, transcript structure (alternative splicing), transcription levels of sense strand of genes, antisense transcription of genes, non-coding transcript levels, among other molecular phenotypes. These "cis" changes in gene activity can in turn affect the states of many other genes (in "trans"), so that they can be viewed as perturbations in the same way as artificial perturbations (e.g., gene knockouts, siRNA knockdown, transgenics, and chemical perturbations) commonly employed in biology to establish causal relationships. However, naturally occurring DNA variation is a more relevant perturbation source given common forms of human disease are caused by such variation, and so understanding causality in the context of such variation is perhaps the most relevant context in understanding disease, how best to assess disease risk, how best to track disease progression, and how best to treat disease.

It is now well established that variations in DNA can be used to infer causal relationships among molecular phenotypes and reconstruct entire gene networks by systematically assessing the impact of DNA variation on gene expression, protein expression, metabolite expression, and the interactions between proteins, proteins and DNA, and proteins and metabolites. The construction of large-scale gene networks can elucidate subnetworks comprised of highly interconnected sets of genes driven by common genetic factors that in turn associate with disease (or other phenotypes of interest), without depending on known pathway information (i.e., completely data driven; objective). As an example, Zhong et al. identified a subnetwork from a large-scale gene network constructed from islets isolated from a population of mice segregating a type 2 diabetes phenotype, where genetic loci in this population associated with t2d were very strongly enriched for associating with genes in this network. The integration of these data were then used to show that over half of the genes in this population supported as

causal for t2d were located in this single subnetwork. SNPs in human populations that were associated with the genes comprising this mouse-derived t2d network were then observed to be greater than 8 times enriched for SNPs that associate with t2d in GWAS (one of the most striking pathway enrichments published to date for a common human disease). Interestingly, no similar enrichments were observed in using known pathways defined in the GO and KEGG databases (Zhong et al. AJHG paper).

Along similar lines, module networks approaches[45] were extended to identify genetic determinants of genetic module differential regulation[80] as well as to identify genetic alteration causally related to the presentation of tumor phenotypes[81].

Regulatory Network Analysis: Causal regulatory networks have also been successful in the inference of disease-related genes that have been experimentally validated. In these networks, similar to networks linking genetic variants with regulation, interactions are directed (i.e. causal) rather than undirected as in protein-protein interaction networks. Thus, if the regulatory network is sufficiently accurate and comprehensive one may use it to traverse back the regulatory event to identify the regulators that are most likely to have produced the specific genetic profile (e.g. gene expression signature) within a specific disease-related phenotype. This method was originally proposed for networks reconstructed from DNA binding signatures of transcription factors, without experimental validation[82]. More recently, these Master Regulator genes were inferred and experimentally validated both in disease, for human high-grade glioma[54], and in normal physiologic processes, for formation of the germinal center[15]. For instance, in high-grade glioma, the MARINa (Master Regulator Inference algorithm) was used to identify the key transcription factors that implement the gene expression signature of the mesenchymal subtype of the disease, associated with the worst prognosis. The analysis identified two genes, C/EBP (including the C/EBPβ and δ subunits) and STAT3, as master regulators. Ectopic expression of both genes, but not of each gene in isolation, was sufficient to reprogram neural stem cells along an aberrant mesenchymal lineage. Co-silencing in high-grade glioma lines, but not silencing of either gene in isolation, was sufficient to abrogate the mesenchymal phenotype and tumorigenesis in vivo. Direct exploration of GWAS data from the TCGA study on Glioblastoma in the context of genes

upstream of these master regulators has identified genetic alterations responsible for virtually 100% of the most aggressive high-grade glioma, including the focal amplification of C/EBPδ gene in ~20% of the mesenchymal cases.

Diseaseome Approaches: Generalizing from pathways, sets of related genes, transcripts, and proteins are well known to follow prescribed programs in the context of human diseases. Thus, another approach for the analysis of data from genome-wide association studies is by exploiting prior biological knowledge on the similarities or dissimilarities across diseases.

For example, while there is widespread belief that the immune system is implicated in a variety of pathophysiological phenotypes, suggesting that autoimmune disorders may share causal genetic variants with them, there are also notable differences across these disorders. For example, the G allele of the *rs2076530* polymorphism *BTNL2* (butyrophilin-like 2, a MHC class II associated gene) is more frequent among patients with Type 1 Diabetes and Rheumatoid Arthritis than in healthy controls, while the A allele was more frequent in patients with Systemic Lupus Erythematosus than in healthy individuals[83]. One way to exploit these disease-relations is to study the results of multiple GWAS with each other, to find SNPs commonly predisposing to the entire set of diseases, or more interestingly find SNPs predisposing to some in the set, while significantly protecting against the others. Such "toggleSNPs" could be used to shed light on the molecular details in actual human disease incidence, a key advantage over disease studies in animal models[84].

Phenotype canalization: In many diseases but especially in cancer, there is evidence of an apparent paradox. While the number of distinct genetic and epigenetic alterations, both germline and somatic, associated with the etiology of the disease is generally large, the number of distinct molecular subtypes arising from the analysis of molecular profile data is significantly smaller. For instance, in high-grade glioma, dozens of genetic alterations have been reported[85] and yet there are only three or four distinct molecular subtypes. If both observations are true, then one has to conjecture the existence of an integrative logic, usually at the transcriptional regulation level, that canalizes signals from the

complex spectrum of genetic and epigenetic alterations into a few molecular phenotypes, representing aberrant yet highly stable developmental states of the cell. The existence of this integrative logic has been elucidated in several tumors, including lymphomas[71] and high-grade gliomas[54]. These observations suggest yet another approach to NBAS, based on the identification of the regulatory modules that control the disease subtype signatures followed by interrogations of pathways directly upstream of these modules as well as by association of genetic alterations in the tissue sample with the activity of these modules, for instance using the mutual information, $MI[A_x; M]$, between the presence of a specific alteration A_x and module activity M. These types of approaches may significantly reduce the number of hypotheses that may need to be tested and increase the specificity of the molecular link from alteration → cellular-phenotype to alteration → molecular-phenotype, the latter being far less prone to assessment errors.

A NEED TO REVIEW HOW WE WORK TOGETHER

The power to build better maps of disease in the post-GWAS era clearly leverages emerging "omics" technologies that will benefit from collecting data from large samples of patients over multiple intervals of time. Most of the historic studies that drive our current understanding of diseases have been performed by single institutions often with the primary goal of taking data to build models that are then communicated as the results and conclusions conveyed by citable scientific articles. This current process does not assume that most data might be more useful if it could be accessed by others to build further models and hypotheses, beyond those envisioned by the original scientist. In fact, the absence of a culture of appropriate data sharing in the life and biomedical sciences is perhaps the single greatest impediment to the rapid development of the integrative techniques described above. For instance, GWAS data will no longer be sufficient in isolation to understand the complexity of disease and how best to predict and treat it, but instead will need to be paired with additional molecular profile data as well as with data that may be used to dissect the underlying regulatory model for the specific cellular context of interest.

Even though genomic data is robust and may be successfully used across a wide range of analyses, most investigators involved in clinical genomic studies hold the data

hostage for fear of missing out on the opportunity to extract the last bit of publishable value from it. The net result is that 80% of data is never published and more importantly never shared. Although journals and funding agencies have started to request that data be made publically accessible, investigators rarely provide their data in a format that is easy to access by others to reproduce their original ideas, or even better to explore new models or new hypotheses not originally contemplated. Even more problematic, the review process may delay release of a critical dataset by years. This should be seen as akin to living in a 21st century scientific "hunter-gatherer" society.

For clinical scientists and network biologists to evolve toward a more generative scientific society, where open access to useful data and models is the rule, both technical and cultural changes are necessary. Most data is not annotated in ways that allow others to easily integrate it or even interpret it. Yet other fields such as electronics and economics live in a world of fully shareable standards for data exchange. This integration will thus require new standards and annotations that have become a part of other scientific disciplines such as astronomers, physicists and climatologists who work with large datasets. The cultural barriers to evolving data sharing involve re-examining current reward structures for career advancement and peer recognition that are based on being a first or last author, and the need to own intellectual property around biologic insights. We need to transition to a workplace where scientists are rewarded for their insights, such as the proposing of new disease models, so that they can occur much earlier in the process of working with clinical/genomic data sets.

One example of piloting the advantages of sharing data, models, and tools is called "The Federation". In the summer of 2010 five groups: Sage Bionetworks, the Butte lab, the Califano lab, the Ideker lab, and the Schadt lab decided to test the mechanics of data sharing by jointly working on projects in aging, diabetes, and cancer based on predefined rules on data access and data sharing. Federation rules imply that anyone interested in data, tools, and models produced by any of the five groups would have access to these pooled resources and would implicitly respect publication rights by including data producers in their manuscripts and by notifying each other of pending manuscripts using this data. More importantly, it was set up so that disease models would

be built by teams dynamically formed for the projects in an environment usually only seen for “open source” software projects. Multiple early experiments such as The Federation will be needed to aid in the development of the type of governance rules and processes required to facilitate the sharing in a laboratory environment needed to build the generative disease maps possible in the NBAS and PWAS worlds that follow the large national scale effort in GWAS.

Conclusions

Regulatory networks are emerging as powerful integrative frameworks to understand and interpret the role of genetics and epigenetics in disease predisposition and etiology. By providing the backbone of molecular interactions through which signals are transduced and gene expression is regulated, they dramatically limit the search space of allele variants and alterations that can be causally linked to the presentation of a phenotype. In addition, by providing accurate regulatory models of the cellular machinery that integrates signals that are dysregulated in disease, they yield valuable hypotheses for diagnostic and prognostic biomarkers, for therapeutic targets, and for the understanding of context-specific synthetic lethality.

For regulatory networks to yield their full potential, however, we must understand their variability across cellular context, their dependence on the genetic and epigenetic layer, and their dynamics over time. The latter is particularly important for diseases where the underlying cellular pathophysiology cannot be considered to be close to steady state, such as metabolic and neurological diseases.

Surprisingly, even rough regulatory models that are largely inaccurate and incomplete are starting to show significant value in dissecting the genetics of disease. Thus, we expect that as these models progress and become better able to deal with the dynamic, cell context-specific nature of biological process regulation, they will dramatically increase their ability to yield key insight into both normal cell physiology and its dysregulation in disease. We herald network reverse-engineering and interrogation as one of the most critical challenges of quantitative biology.

References:

1. Stranger BE, Stahl EA, Raj T. Progress and Promise of Genome-wide Association Studies for Human Complex Trait Genetics. Genetics 2010.
2. Kraft P, Hunter DJ. Genetic Risk Prediction -- Are We There Yet? N Engl J Med 2009.
3. Hardy J, Singleton A. Genomewide Association Studies and Human Disease. N Engl J Med 2009:NEJMra0808700.
4. Goldstein DB. Common Genetic Variation and Human Traits. N Engl J Med 2009.
5. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 2008;40:638-45.
6. Lyssenko V, Jonsson A, Almgren P, et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 2008;359:2220-32.
7. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008;322:881-8.
8. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.
9. Narayanan M, Vetta A, Schadt EE, Zhu J. Simultaneous clustering of multiple gene expression and physical interaction datasets. PLoS Comput Biol 2010;6:e1000742.
10. Stephens PJ, Greenman CD, Fu B, et al. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell 2011;144:27-40.
11. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. Am J Hum Genet 2010;86:581-91.
12. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545-50.
13. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 1999;27:29-34.
14. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet 2010;11:843-54.
15. Lefebvre C, Rajbhandari P, Alvarez MJ, et al. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Mol Syst Biol 2010;6:377.
16. Wang K, Alvarez MJ, Bisikirska BC, et al. Dissecting the interface between signaling and transcriptional regulation in human B cells. Pac Symp Biocomput 2009:264-75.
17. Consortium EP, Birney E, Stamatoyannopoulos JA, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007;447:799-816.
18. Margolin AA, Palomero T, Sumazin P, Califano A, Ferrando AA, Stolovitzky G. ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. Proc Natl Acad Sci U S A 2009;106:244-9.

19. Ravasi T, Suzuki H, Cannistraci CV, et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell 2010;140:744-52.
20. Rual JF, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005;437:1173-8.
21. Chen Y, Zhu J, Lum PY, et al. Variations in DNA elucidate molecular networks that cause disease. Nature 2008;452:429-35.
22. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. Nature 2008;452:423-8.
23. Schadt EE, Lamb J, Yang X, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 2005;37:710-7.
24. Yang X, Deignan JL, Qi H, et al. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. Nat Genet 2009.
25. Zhu J, Zhang B, Smith EN, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 2008;40:854-61.
26. Zhong H, Beaulaurier J, Lum PY, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. PLoS Genet 2010;6:e1000932.
27. Kinzler KW, Vogelstein B. Landscaping the cancer terrain. Science 1998;280:1036-7.
28. Dobrin R, Zhu J, Molony C, et al. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. Genome Biol 2009;10:R55.
29. Keller MP, Choi Y, Wang P, et al. A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. Genome Res 2008;18:706-16.
30. Ferron M, Wei J, Yoshizawa T, et al. Insulin signaling in osteoblasts integrates bone remodeling and energy metabolism. Cell 2010;142:296-308.
31. Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature 2009;461:218-23.
32. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet 2005;37:382-90.
33. Wang K, Saito M, Bisikirska BC, et al. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nat Biotechnol 2009;27:829-39.
34. Mani KM, Lefebvre C, Wang K, et al. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Mol Syst Biol 2008;4:169.
35. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 2004;431:308-12.
36. Bandyopadhyay S, Mehta M, Kuo D, et al. Rewiring of genetic networks in response to DNA damage. Science 2010;330:1385-9.
37. Pan W. Network-based model weighting to detect multiple loci influencing complex diseases. Hum Genet 2008;124:225-34.
38. Chen GK, Thomas DC. Using biological knowledge to discover higher order interactions in genetic association studies. Genet Epidemiol 2010;34:863-78.
39. Calvano SE, Xiao W, Richards DR, et al. A network-based analysis of systemic

inflammation in humans. Nature 2005;437:1032-7.
40. Elkon R, Rashi-Elkeles S, Lerenthal Y, et al. Dissection of a DNA-damage-induced transcriptional network using a combination of microarrays, RNA interference and computational promoter analysis. Genome Biol 2005;6:R43.
41. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science 2005;308:523-9.
42. Quayle AP, Siddiqui AS, Jones SJ. Perturbation of interaction networks for application to cancer therapy. Cancer Inform 2007;5:45-65.
43. Birtwistle MR, Hatakeyama M, Yumoto N, Ogunnaike BA, Hoek JB, Kholodenko BN. Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. Mol Syst Biol 2007;3:144.
44. Nelander S, Wang W, Nilsson B, et al. Models from experiments: combinatorial drug perturbations of cancer cells. Mol Syst Biol 2008;4:216.
45. Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 2003;34:166-76.
46. Friedman N. Inferring cellular networks using probabilistic graphical models. Science 2004;303:799-805.
47. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science 2003;301:102-5.
48. Linding R, Jensen LJ, Ostheimer GJ, et al. Systematic discovery of in vivo phosphorylation networks. Cell 2007;129:1415-26.
49. Bandyopadhyay S, Chiang CY, Srivastava J, et al. A human MAP kinase interactome. Nat Methods 2010;7:801-5.
50. Huang Y, Zou Q, Song H, et al. A study of miRNAs targets prediction and experimental validation. Protein Cell 2010;1:979-86.
51. Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E. Network-based prediction of human tissue-specific metabolism. Nat Biotechnol 2008;26:1003-10.
52. Zhu J, Lum PY, Lamb J, et al. An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet Genome Res 2004;105:363-74.
53. Yang X, Deignan JL, Qi H, et al. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. Nat Genet 2009;41:415-23.
54. Carro MS, Lim WK, Alvarez MJ, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature 2010;463:318-25.
55. Zhao X, D DA, Lim WK, et al. The N-Myc-DLL3 cascade is suppressed by the ubiquitin ligase Huwe1 to inhibit proliferation and promote neurogenesis in the developing brain. Dev Cell 2009;17:210-21.
56. Yang X, Peterson L, Thieringer R, et al. Identification and validation of genes affecting aortic lesions in mice. J Clin Invest 2010;120:2414-22.
57. Konig R, Stertz S, Zhou Y, et al. Human host factors required for influenza virus replication. Nature 2010;463:813-7.
58. Jansen R, Yu H, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003;302:449-53.
59. Sharan R, Suthram S, Kelley RM, et al. Conserved patterns of protein interaction

in multiple species. Proc Natl Acad Sci U S A 2005;102:1974-9.
60. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006;313:1929-35.
61. Wang W, Cherry JM, Botstein D, Li H. A systematic approach to reconstructing transcription networks in Saccharomycescerevisiae. Proc Natl Acad Sci U S A 2002;99:16893-8.
62. Pe'er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. Bioinformatics 2001;17 Suppl 1:S215-24.
63. di Bernardo D, Thompson MJ, Gardner TS, et al. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nat Biotechnol 2005;23:377-83.
64. Chen BJ, Causton HC, Mancenido D, Goddard NL, Perlstein EO, Pe'er D. Harnessing gene expression to identify the genetic basis of drug resistance. Mol Syst Biol 2009;5:310.
65. Jia J, Zhu F, Ma X, Cao Z, Li Y, Chen YZ. Mechanisms of drug combinations: interaction and network perspectives. Nat Rev Drug Discov 2009;8:111-28.
66. Chautard E, Thierry-Mieg N, Ricard-Blum S. Interaction networks: from protein functions to drug discovery. A review. Pathol Biol (Paris) 2009;57:324-33.
67. Xie L, Li J, Bourne PE. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. PLoS Comput Biol 2009;5:e1000387.
68. Berger SI, Iyengar R. Network analyses in systems pharmacology. Bioinformatics 2009;25:2466-72.
69. Iadevaia S, Lu Y, Morales FC, Mills GB, Ram PT. Identification of optimal drug combinations targeting cellular networks: integrating phospho-proteomics and computational network analysis. Cancer Res 2010;70:6704-14.
70. Pandey G, Zhang B, Chang AN, et al. An integrative multi-network and multi-classifier approach to predict genetic interactions. PLoS Comput Biol 2010;6.
71. Compagno M, Lim WK, Grunn A, et al. Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. Nature 2009;459:717-21.
72. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A 2008;105:4323-8.
73. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 2002;18 Suppl 1:S233-40.
74. Zien A, Kuffner R, Zimmer R, Lengauer T. Analysis of gene expression data with pathway scores. Proc Int Conf Intell Syst Mol Biol 2000;8:407-17.
75. Faust K, Dupont P, Callut J, van Helden J. Pathway discovery in metabolic networks by subgraph extraction. Bioinformatics 2010;26:1211-8.
76. Ulitsky I, Krishnamurthy A, Karp RM, Shamir R. DEGAS: de novo discovery of dysregulated pathways in human diseases. PLoS ONE 2010;5:e13367.
77. Lage K, Karlberg EO, Storling ZM, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 2007;25:309-16.
78. Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics

2011;27:95-102.
79. Hannum G, Srivas R, Guenole A, et al. Genome-wide association data reveal a global map of genetic interactions among protein complexes. PLoS Genet 2009;5:e1000782.
80. Lee SI, Pe'er D, Dudley AM, Church GM, Koller D. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proc Natl Acad Sci U S A 2006;103:14062-7.
81. Akavia UD, Litvin O, Kim J, et al. An integrated approach to uncover drivers of cancer. Cell 2010;143:1005-17.
82. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, Chinnaiyan AM. Mining for regulatory programs in the cancer transcriptome. Nat Genet 2005;37:579-83.
83. Orozco G, Eerligh P, Sanchez E, et al. Analysis of a functional BTNL2 polymorphism in type 1 diabetes, rheumatoid arthritis, and systemic lupus erythematosus. Hum Immunol 2005;66:1235-41.
84. Sirota M, Schaub MA, Batzoglou S, Robinson WH, Butte AJ. Autoimmune disease classification by inverse association with SNP alleles. PLoS Genet 2009;5:e1000792.
85. TCGA-Consortium. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061-8.
86. Phillips HS, Kharbanda S, Chen R, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 2006;9:157-73.
87. Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010;17:98-110.