COMPUTATIONAL FUNCTIONAL GENOMICS FOR DIRECTING BIOLOGICAL DISCOVERIES

CHRISTOPHER YOUNG PARK

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF

COMPUTER SCIENCE

ADVISER: OLGA G. TROYANSKAYA

APRIL 2014

Abstract

Biological systems have been extensively studied for over a century, however we still only have a partial functional and mechanistic understanding of the interplay between genes and pathways. Recently, there has been an exponential increase in experimental datasets generated. However, the complexity of data types and the ambiguity of dataset relevance to biological processes and pathways have limited the integrated usage of this vast knowledge base for directing biological discoveries in human and model organisms. In this thesis, I develop several approaches of utilizing such public effort to address the challenges of inferring gene function and diverse biomolecular interaction networks and of improving the transfer of functional knowledge between organisms to facilitate the investigation of understudied biological processes.

Specifically, in the first part of the thesis, I show that computational functional genomics can be used to improve the transfer of gene annotations between organisms. Furthermore, I demonstrate that functional knowledge transfer, when coupled with machine learning algorithms, can improve the coverage and accuracy of gene function prediction in a diverse set of organisms. In the second part of the thesis, I provide a general method for simultaneous prediction of many interaction types genome-wide and present the results of applying this methodology in *S. cerevisiae.* By incrementally overlaying different interaction types as suggested by our results, investigators can make specific and testable novel hypotheses about new pathways, new pathway components, or new interconnections between existing pathways. Finally, I extend our interaction inference work in *S. cerevisiae* to mammalian organisms, by methodologically addressing the largest source of biological variation in the metazoan data compendium: tissue and cell-lineage heterogeneity.

Acknowledgements

First of all, I would like to thank my adviser Olga Troyanskaya for leading me the way to this point. Without her support, all would not have been possible. I would also like to thank the current and former members of the Troyanskaya lab, Curtis Huttenhower, Aaron Wong, Ana Pop, Young-suk Lee, Dima Gorenshteyn, Arjun Krishnan, Casey Greene, Yuanfang Guan, Qian Zhu, Jian Zhou, Vicky Yao, Max Homilius, Jonathan Goya, Matt Hibbs, Chad Myers, Maria Chikina, Patrick Bradley and David Hess, for all the great moments and conversations we shared together as a group.

I would like to thank my committee members, Mona Singh, Kai Li, Robert Schapire and Moses Charikar for their feedback, advice and time. I also thank Melissa Lawson, Marybeth Fedele and John Wiggins for their excellent guidance and technical support.

Also, to my friends I met at Princeton, Wonho Kim, Younghan Park, Aditya, Aravindan, CJ, Anirudh, Esung Yoon, Sunghwan Ihm, Rajsekar, Soyoung and Keun, I thank you for the friendship and encouragement.

Last but not least, I would like to thank my parents and sister for their love and support. Finally, Soyi, my wife, thank you for being my life long companion and for your unconditional love.

I was supported by Princeton graduate fellowship, NSF CAREER award DBI-0546275, NIH grants R01 GM071966 and T32 HG003284, and NIGMS Center of Excellence grant P50 GM071508.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Biological background of computational biology

Cells are the building blocks of life. Within a cell lies DNA (deoxyribonucleic acid), the genetic code that contains all the necessary instructions for building a cell and a living organism. DNA takes the form of a double helix, where each strand is a chain of nucleotides of four variations (A, G, C, T). The composition and order of these four nucleotides encompasses the genetic information of the organism. Specific regions of the DNA are called genes that are transcribed into messenger RNA (mRNA) and encode the amino acid sequences for all proteins, the functional machinery of the cell. The passing of information from a gene to a mRNA molecule is termed transcription, and a protein assembled from amino acids by the ribosome with a mRNA template molecule is called translation.

The amount and collection of proteins required at any given time for the cell can vary drastically. Some proteins are required during the different developmental stages of the organisms while others are produced in response to the environment (e.g. heat shock or starvation). Also, in multi-cellular organisms such as mammals, different tissues and cell-types have specialized functions and thus the composition of proteins can be drastically different. This regulation of the functional machinery unit proteins can happen at all level of the information cascade. Transcription of a gene can be regulated by proteins that activate or suppress RNA-polymerases from producing mRNA molecules. A transcribed mRNA molecule can be selectively degraded through the cellular RNAi machinery and thus preventing the translation process of producing proteins.

Finally, post-translational regulation also can occur where a protein can be covalently modified to be biologically “active” or “inactive” and even be degraded.

This complex level of regulation is ultimately required for maintaining the homeostasis of the organism, however requires an extensive network of cooperation between multiple proteins. Proteins can physically bind and work together as operating units called complexes. Multiple complexes and proteins work together to form the cellular circuitry called pathways to provide multiple tasks, such as response to stress or signal transduction. Much like electric circuits, cellular pathways can execute complex level of logic. The large number of genes and gene products (e.g. >20,000 genes in human) allow the capacity to encode such logic, however also diverse types of pairwise interactions, ranging from physical protein-protein interactions and modifications to indirect regulatory relationships, provide a significant toolbox to carry out the pathway’s overarching cellular role. In table 1, I describe the variety of biomolecular interactions that occur at multiple level of specificity in the cell.

Interaction type	**Description**
Complex	Any macromolecular complex composed of two or more polypeptide subunits.
Covalent modification	Proteins regulated by transfer or remove of a molecule or atom from a donor to an amino acid side chain that serves as the acceptor of the transferred molecule or regulating an enzyme by altering the amino acid sequence itself by proteolytic cleavage.
Functional group transfer	Transfer or removal of a functional group.
Functional relationship	Two proteins function in same biological process.
Interaction pathway	Two genes interact within a pathway level, including post-transcriptional, transcriptional and post-translational regulation or the functional dependency such as synthetic interactions.
Isoenzyme	Enzymes that differ in amino acid sequence but catalyze the same

	chemical reaction.
Mediated by small molecule	Two proteins where a small molecule is involved as part of a protein modification.
Metabolic interaction	Functionally associated at the metabolic level.
Non covalent binding	Two proteins interact in a non covalent nature.
Peptide transfer	Transfer peptide to protein.
Phenotypic aggravation	Mutation or over expression of one gene results in suppression of any phenotype (other than lethality/growth defect) associated with mutation or over expression of another gene.
Phenotypic alleviation	Mutation or over expression of one gene results in enhancement of any phenotype (other than lethality/growth defect) associated with mutation or over expression of another gene.
Phenotypic interaction	Mutation or over expression of one gene results in alteration of any phenotype (other than lethality/growth defect) associated with mutation or over expression of another gene.
Phosphate transfer	Addition/removal of a phosphate (PO4) group to/off a protein.
Phosphorylation	Addition of a phosphate (PO4) group to a protein.
Physical interaction	Two proteins physically interact.
Posttranscriptional regulation	Post-transcriptional regulation is the control of gene expression at the RNA level.
Posttranslational regulation	Post-translational regulation refers to the control of the levels of active proteins.
Regulatory interaction	A gene regulates a gene either at the RNA, protein or transcription level.
Same enzyme class	Two enzymes that share the same enzyme class.
Shared pathway	Two proteins are closely involved in a pathway
Synthetic aggravation	Mutation or deletion of one gene aggravates the effect of a strain mutated/deleted for another gene.
Synthetic alleviation	Mutation or deletion of one gene alleviates the effect of a strain mutated/deleted for another gene.
Synthetic growth defect	Interaction is inferred when mutations in separate genes, each of which alone causes a minimal phenotype, result in a significant growth defect under a given condition when combined in the same cell.
Synthetic interaction	Interaction in which a combination of mutations in two or more genes of a single strain results in a phenotype that is different in degree or nature from the phenotypes conferred by the individual mutations.
Synthetic lethal	Mutations or deletions in separate genes, each of which alone causes a minimal phenotype, result in lethality when combined in the same cell under a given condition.
Synthetic rescue	Mutation or deletion of one gene rescues the lethality or growth defect of a strain mutated/deleted for another gene.

Transcriptional regulation	Transcriptional regulation is the change in gene expression levels by altering transcription rates.
Ubiquitination	The post-translational modification of a protein by the covalent attachment of one or more ubiquitin monomers.
Ubiquitin transfer	The post-translational modification of a protein by the covalent attachment or removal of one or more ubiquitin monomers.

Table 1. Description of the variety of biomolecular interactions.

Here I list the functional and biochemical description of each interaction type commonly occurring in biomolecular pathways.

Fundamentally, to start unraveling the complexity of a living organism systematic experimental effort is required. Ever since the earliest DNA sequencing technology was developed by Sanger and colleagues, there has been large advancement in biomolecular technology that allows biologist to monitor the cellular state at multiple levels. Sanger sequencing and now next-generation sequencing allowed researchers to complete human genome project, but now investigators can sequence multiple genomes within a week and continue to advance in efficiency and cost [1,2]. Also, the invention of gene expression microarrays and recently developed RNA-Seq allows biologist to quantitatively measure the transcription level of thousands of genes simultaneously from any collection of cells [3].

In addition, there have been a number of high-throughput assays that have been developed to detect variety of interactions such as gene-gene, protein-protein, protein-DNA and protein-RNA interactions. Gene-gene interactions, often referred as genetic interactions, have been extensively mapped in yeast through SGA technology [4] by creating double-mutant strains of yeast and measuring the growth-rate compared to the single mutant. In mammalian organisms, shRNA pooled libraries allows the detection of genetic interactions [5] and the recent development of the

CRISPR-Cas9 system provides great opportunities for extensive mapping of such interactions [6]. Affinity mass spectrometry or yeast two-hybrid methods have allowed biologist to detect thousands of protein-protein interactions [7,8]. Also, the development of techniques for cross-link biochemistry (e.g. formaldehyde and U.V) has allowed biologist to detect interactions between protein bound to DNA or RNA, by cross-linking the protein to the substrate molecule and conducting follow-up sequencing to identify the exact binding locations of these proteins [9]. Such techniques have been applied to multiple cell and tissue samples providing *in vivo* depiction of RNA and DNA bound protein profiles. Overall, with the development of such high-throughput assays there has been an exponential increase in genomic datasets generated. However, biological assays are often very noisy with poor signal-to-noise ratios leading to high number of false positive results. Additionally, cell-cultures or clinical specimens that are experimentally tested consist of large biological variation and are almost always subject to batch effects due to the sample preparation steps. Thus, the high-throughput experiments has consistently confronted with the challenge of accurately interoperating this wealth of data.

1.2 The challenges of computational biology

Over the last half-century, biologists have been attempting to unveil the vast complexity of biological systems [10,11]. As high-throughput experiments become increasingly common [2-4,12,13], biologists face substantial challenges effectively leveraging genome-scale data from diverse organisms to inform new hypotheses [14,15]. On an individual gene level, functional annotations have improved, but still there exist a large gap between our knowledge and

biological systems. Most significantly, the interplay of relationship between proteins has remained a significant hurdle in understanding complex biological systems [16]. Genes do not usually work as a single unit. Proteins, which are encoded by genes, work under a complex network of interactions between other genes to maintain the complexity of the living organism. Even more dauntingly, genes interact in many different biomolecular and functional manners with multiple partners [17]. Thus, one key challenge of computational and systems biology is to bridge three aspects of this complexity: the growing body of high-throughput data assaying these interactions; the specific interactions in which individual genes participate; and the genome-wide patterns of interactions in a system of interest.

Concretely, as a biologist, in order to navigate the landscape of biological systems for hypothesis generating exploration, one would like to know how a gene set of interest fit within the complex gene cellular interaction network in the investigating organism. However, due to the vast intricacy of the cell, it would be of great benefit that the cell would be presented through subdivided functional networks that would each represent a critical slice of the cellular system. More specifically, one would like to explore how the query set of genes interacts with each other and between other genes through the diverse set of possible interaction types at varying levels of biomolecular specificity.

For instance, lets imagine an investigator studying cyclin proteins in yeast, which are key regulators in cell-cycle control[18]. Understanding the full functional context of cyclins would require decomposition into many different functional interaction layers with full coverage at genome scale. For example, at the post-translational level, cyclins activate cyclin-dependent kinases by forming protein complexes that subsequently phosphorylates many key cell cycle

proteins [18]. At the transcriptional level, cyclins are regulated by a heterodimeric transcription factor complex SBF [19]. Finally, at the repressor and activator level, the inhibition of cyclins is through multiple SCF ubiquitin ligase complexes that lead to protein degradation [20]. As illustrated, the heterogeneity surrounding the few cyclin proteins is extremely complex. With only limited gene function annotations or a simplistic gene physical contact network, it would be insufficient for understanding the full interplay of proteins, which together enable many cellular processes. Unfortunately, unlike this relatively small network, which has taken over 50 years of laborious research to map out, very little is known of most genes functional associations and the type of biomolecular interaction. Thus, there is a great need to develop algorithmic and statistical approaches to efficiently identify pathways from the burgeoning abundance of data from high-throughput molecular techniques.

Another key challenge in computational biology is identifying the functional roles of genes in human diseases. Due to technical and ethical challenges many human diseases or biological processes are studied in model organisms [21-23]. However, experimental data coverage for a model organism can be sparse and prior functional knowledge (i.e. low-throughput experiments validating a gene's function) can be notably limited [24]. These impediments affect the breadth and accuracy of bioinformatic methods (e.g. machine-learning algorithms) that apply prior knowledge in learning novel biology [25,26]. As a consequence, the applicability of these methods is often limited to biological processes and pathways that are already well characterized for an organism.

For example, a common challenge for biological researchers is interpreting the results of a genome-wide experiment (e.g. a list of candidate genes from a microarray experiment) and

generating hypotheses for experimental follow-up. There are several effective resources, some network-based, for researchers to analyze their gene sets [27-33]. These resources cover a wide range of organisms and address different needs of biologists by applying a variety of methods: from pathway analysis of a gene list to machine learning algorithms that predict a gene's function. All of these resources' methods require known examples (i.e. pathways with at least a few annotated genes) in an organism. For example, the most widely utilized methods Gene Set Enrichment Analysis (GSEA) [27] and hypergeometric test based GO term enrichment analysis [34] both measure the significance of overlap or rank of known biological gene sets. Consequently, the effectiveness of these applications is constrained by the extent of prior knowledge and available experimental data in the queried organism.

Other resources address the problem of disparate data coverage among organisms by focusing on methods to transfer high-throughput data (e.g. microarray, physical interaction experiments) between organisms [35-38]. However, these efforts are limited to learning gene association networks, and none of them solve the problem of making accurate functional predictions and associations for biological processes that have not been well studied in a given organism. For example, most of the discovered genes involved in neuromuscular process have been in mouse (65 known genes according to Gene Ontology [39]). Relatively few genes are definitively known in mammalian systems outside of this model organism. Consequently, many existing methods will not be able to predict genes to that biological process in rat (where only one such gene is experimentally annotated), and a biologist using a rat model system with existing resources will not be able to leverage the known biology in mouse. Thus, there is a great need for biologists the technology that allows the systematic application of prior functional knowledge from other

organisms to their organism of study, at multiple points in an analytic workflow: from interpreting experimental results to generating hypotheses for functional assays.

1.3 Contributions

In this thesis, I tackle the challenges laid out in the previous chapter 1.2. First, I address the challenge of transferring discoveries in model organisms back to human or other model organisms. Traditional methods for transferring novel gene function annotations have relied on finding genes with high sequence similarity believed to share evolutionary ancestry. However, sequence similarity does not guarantee a shared functional role in molecular pathways. In chapter 2, I show that functional genomics can complement traditional sequence similarity measures to improve the transfer of gene annotations between organisms. I coupled our knowledge transfer method with current state-of-art machine learning algorithms (SVM [40] and Random Forest [41]) that enabled us predicted gene functions to 8,091 biological processes currently without any experimental annotations across six organisms. In addition, I demonstrate our method can help biologists systematically integrate prior knowledge from diverse model organisms to direct targeted experiments for understudied processes in their organism of study. In collaboration with molecular biologist Dr. Rebecca Burdine at Princeton University, we have experimentally demonstrated the power of our method by conducting an *in vivo* experiment validating our predicted role of gene *wnt5b* in establishing correct heart asymmetry during development in zebrafish. Together, our results show that functional knowledge transfer can improve the coverage and accuracy of machine learning methods used for gene function prediction in a

diverse set of organisms. Such an approach can be applied to additional organisms, and will be especially beneficial in organisms that have high-throughput genomic data with sparse annotations. In addition, our computational method resulted in an interactive web portal, IMP (imp.princeton.edu), which contains the integration results of over 45,000 diverse heterogeneous genomic experiments across seven organisms (human, mouse, rat, zebrafish, fly, worm, yeast) and enables biologists to analyze their experiments in the functional contexts of sets of genes and gene-gene probabilistic functional networks from across these organisms. This work has been published at PLoS Computational Biology [42] and Nucleic Acids Research [43].

Next, I address the challenge of inferring biomolecular pathways that are built from diverse types of pairwise interactions. In chapter 3, I describe a methodology for simultaneously predicting specific types of biomolecular interactions using high-throughput genomic data in the model organism *Saccharomyces cerevisiae* (baker's yeast). This algorithm uniquely uses a statistical model to interrogate the hierarchical information embedded between various types of bimolecular interaction types and hidden signals in high-throughput data. This results in a comprehensive compendium of whole-genome networks for yeast, derived from ~3,500 experimental conditions and describing 30 interaction types, which range from general (e.g. physical or regulatory) to specific (e.g. phosphorylation or transcriptional regulation). I used these networks to investigate molecular pathways in carbon metabolism and cellular transport, proposing a novel connection between glycogen breakdown and glucose utilization supported by recent publications. Additionally, in collaboration with Dr. David Hess at Santa Clara University, we experimentally confirmed the accuracy and capacity of the algorithm by validating 14 out of

20 gene pairs predicted to have a synthetically lethal interaction (>100 fold increase over baseline discovery rate). This work has been published at PLoS Computational Biology [44].

Finally in chapter 4, I extend our work in chapter 3 of predicting Biomolecular interactions in unicellular organisms (e.g. yeast), and developed a strategy of inferring such interaction networks in metazoan mammalian orgnaisms. Uniquely compared to unicellular orgnaisms, tissue-specific expression underlies the development and maintenance of diverse cell types in metazoan organisms. Therefore, the biological and technical heterogeneity of the functional genomic data originating from the diverse tissues and cell-lineages compounds the difficulty in translating the vast amount of human genomic data into specific interaction-level hypotheses. To address such challenge, I developed an integrated approach for inferring the individual types of human biomolecular interactions, scalable to the whole-genome and robust to any biological dataset of diverse tissue contexts (i.e. experimental results drawn from differing tissues). I demonstrate that directly incorporating tissue contextual information significantly improves the accuracy of our predictions and show that such whole-genome scale predictions can be used to highlight active regulatory interactions from condition-specific primary experimental datasets (e.g. ChIP-Seq, mass spectrometry). A version of this work will soon be submitted for publication.

2 Functional knowledge transfer for the investigation of under-studied biological processes

This chapter describes joint work with Aaron Wong and Casey Greene, who together contributed in the computational aspects and assembly of related manuscripts. Also, Jessica Rolland contributed to this work for the validation experiments.

2.1 Introduction

Defining the role of proteins in pathways is among the key challenges of human genomics. Many successful approaches have been developed for prediction of protein function and pathway membership [30,45-49], however they rely on prior knowledge in the organism of interest to make new predictions (i.e. at least some genes in the organism already annotated to the pathway) [25,50-53]. These approaches rely on identifying characteristic behavioral patterns, in functional genomic datasets, phylogenetic profiles, or genomic feature studies of genes that are known to participate in a pathway, then use these patterns to predict additional pathway members [54-56].

For example, gene expression and protein interaction profiles can be used by machine learning methods to associate novel genes to pathways based on previously known pathway members [57,58]. The potential of such computational approaches to direct experiments has been demonstrated in studies investigating mitochondrial biogenesis [59] and seed pigmentation [60]. Other common exploratory methods, such as hierarchical clustering [61], don't directly use known gene annotations to learn a prediction classifier, however they often use existing annotations to interpret the resulting cluster of genes (e.g. gene enrichment analysis) [27]. However in many organisms including human, pathways and processes where functional annotations of genes are most needed often have few or no prior experimentally confirmed annotations, making novel predictions of genes that may participate in such a process difficult or impossible. Thus, our study describes a method to robustly increase the set of prior gene annotations, which has the potential to improve all function prediction methods by increasing the accuracy of their predictions and enabling wider coverage of pathways and biological processes.

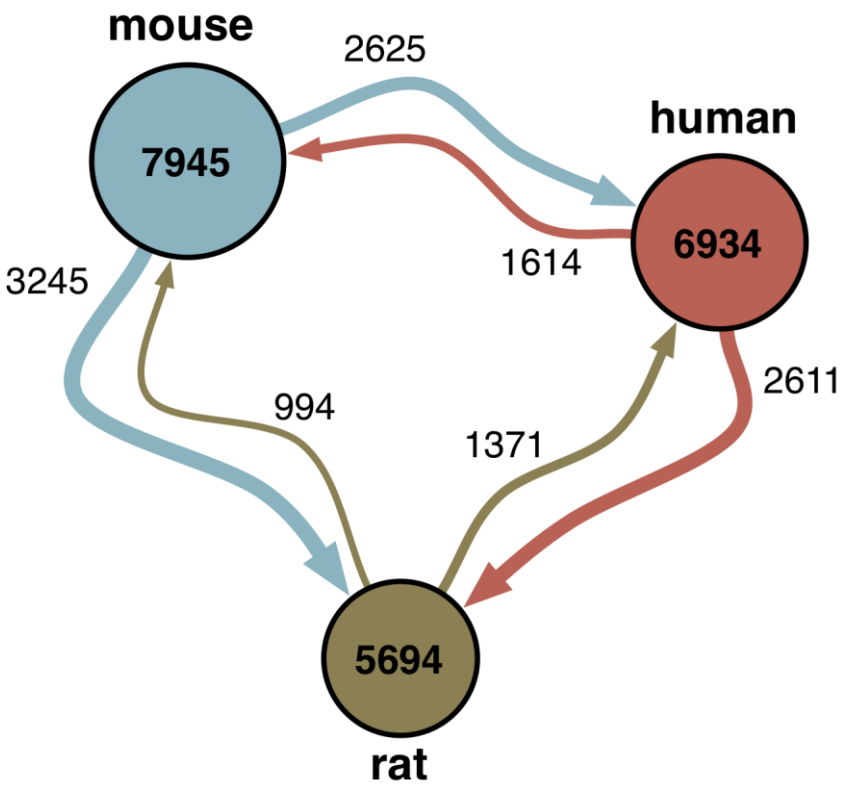

Figure 1. Functional knowledge of biological processes is far from uniform, even among closely related organisms.

Each node represents the number of experimentally annotated biological processes in an organism. Each edge value corresponds to the number of experimentally annotated processes in the source organism that lack any experimental annotations in the target organism. Thus, the directed edges between nodes indicate the direction of potential annotation transfer between organisms. For example, 3,245 processes with annotations in mouse have no experimentally annotated genes in rat.

Many of these processes are well studied in *some* model organism, but not necessarily in an investigator's organism of interest. Even when applying a conservative examination of only the closely related and heavily studied mammalian species human, mouse, and rat, processes represented in one species are often not well-characterized in another (summarized in Figure 1). For example, the process *cellular glucose homeostasis,* an increasingly important process due to

the role of cellular metabolism in cancer development, has less than 5 gene annotations in human, yet has 31 in mouse, a commonly used model organism for cancer studies. These processes (referred to in the text as *understudied processes*) are not well studied in a particular organism of interest (i.e. very few genes are annotated) but might be well characterized in some other organism.

A longstanding solution to improving the prediction accuracy of understudied processes has been to transfer functional annotations from organisms where the process is better characterized [62]. The critical challenge in accurately transferring functional knowledge between organisms is identifying the appropriate genes for the transfer: those genes that share not only sequence similarity, but also conserved pathway roles. Large-scale automated methods have so far exclusively used sequence homology to identify functionally conserved genes [63,64]. However, the relationship between sequence similarity and function is not trivial. For example, human angiopoietin-4 (ANGPT4), an important angiogenesis growth factor, has been shown to activate TEK (tyrosine-protein kinase receptor), while the mouse sequence-ortholog (Angpt4) has been shown to inhibit TEK [65].

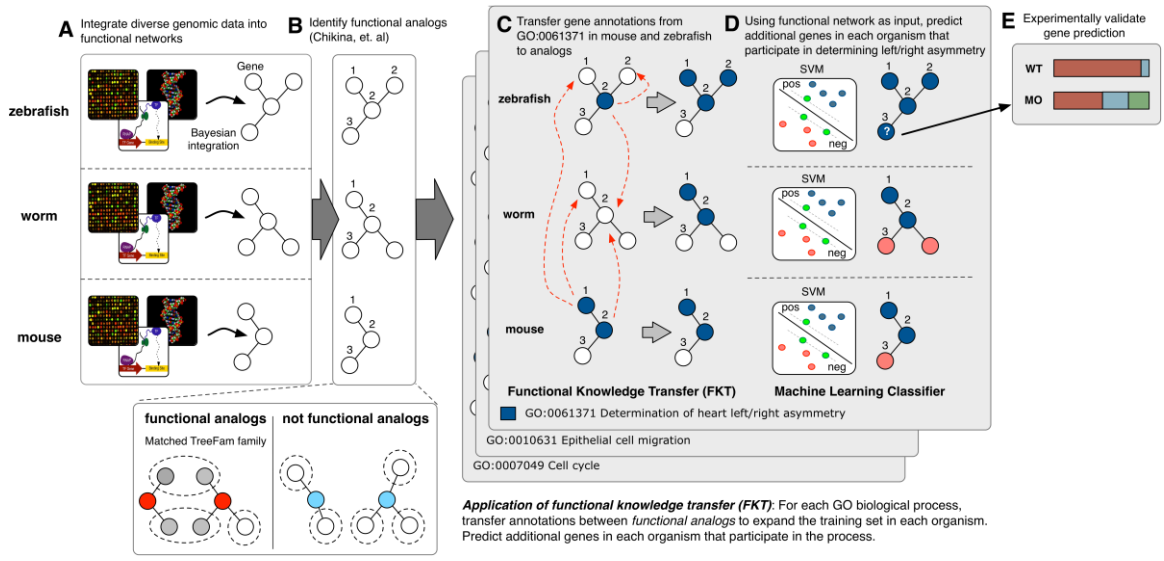

Figure 2. Schematic of the functional knowledge transfer.

A. A functional relationship network is constructed for each organism through Bayesian integration of heterogeneous genomic data (e.g. expression, TF motif binding, physical interaction assays). B. Functionally analogous gene pairs (i.e. functional analogs) are identified by computing a gene pairwise functional similarity score introduced in Chikina et. al between all sequence homologs. Functional similarity is measured by the statistical significance of the number of common TreeFam gene families in the functional relationship network neighbors of each homologous gene pair. C. Next, experimentally confirmed biological process annotations for each gene are transferred to its functional analogs. D. For each biological process the extended set of gene annotations (which include direct gene annotations, if available, and cross-annotated genes) can be used as training examples for machine learning methods (SVM used in this study) to make novel gene membership predictions. E. Top predicted genes are carried over for experimental validation.

In our previous work [24], we developed a cross-organism gene functional similarity measure, which relied on the concept that functional genomics data can be used to resolve homologous relationships among closely related genes. The approach summarizes the compendium of genomics data in each organism into functional relationship networks to identify genes that do not simply share sequence similarity but also functional behavior in large collections of

heterogeneous functional data, and are thus functionally analogous (referred to in text as *functional analogs*). In this current study, we present a novel knowledge transfer method, Functional Knowledge Transfer (also referred to in text as FKT and outlined in Figure 2), which leverages the mapping of functional analogs to direct cross-organism annotation transfer for function prediction. FKT can be especially beneficial for existing and future machine learning methods studying biological processes with sparse annotations in any given organism of interest. By transferring experimental knowledge between genes that have been identified as functional analogs, our method extends beyond simple annotation transfer by sequence similarity. Experimental functional annotations are only transferred for genes that are not just similar in sequence, but also in their functional behavior derived from a large and relatively comprehensive compendium of genomic data.

In this study, we show that FKT improves the prediction accuracy of machine learning algorithms, particularly for biological processes with few existing annotations in an organism of study. We compare FKT to annotation transfer by sequence similarity (BLAST) and demonstrate the superior performance of our method in improving gene function prediction performance. The consistent improvement and high performance across various state-of-the-art classification algorithms demonstrates our approach is robust to different learning models, which is crucial for wide applicability.

We apply FKT to gene function (i.e. biological process) prediction in six metazoan organisms (*Homo sapiens, Mus musculus, Rattus novegicus, Drosophila melanogaster, Danio rerio* and *Caenorhabditis elegans*) and show that FKT is robust enough for the automated transfer of annotations among these diverse organisms and accurate function prediction. Finally, we

demonstrate an application of FKT to discovering novel biology by coupling the knowledge transfer with a Support Vector Machine (SVM) to predict proteins involved in left-right asymmetry regulation during heart development in *Danio rerio.* We correctly predict several proteins in the pathway and experimentally confirm the first evidence of *wnt5b*'s role in the process. A comprehensive application of FKT to 11,000 biological processes, along with the functional relationship networks for all six organisms, are available through the IMP web-server portal accessible at http://imp.princeton.edu [43].

2.2 Results

In Figure 2, we outline the pipeline for FKT and the subsequent gene function predictions (details provided in the Methods section below). Briefly, we first integrated high and low-throughput experimental data such as gene expression data, protein-protein interaction data, protein domain and transcription factor binding motif information into functional networks for each of seven organisms (*Saccharomyces cerevisiae* was also included as an annotation source). Next, we calculated a network-based functional similarity score as described in our prior work [24] but extended here to additional organisms and data sources, between all ortholog and paralog pairs in a Treefam [63] gene family to identify the targets for annotation transfer. Homologs with high functional similarity scores were determined to be functional analogs. Next, we applied FKT by transferring all gene-process annotations between functional analogs and merge these with existing annotations (if available) in an organism. To test the predictive power of FKT, the set of transferred and organism-specific annotations were used to train a Support

Vector Machine (SVM) classifier [40] and predict new genes to all biological processes in six metazoan organisms. Functional network connection weights (i.e. the inferred probability that two genes co-function in the same biological process), were treated as input features to the classifier (see Methods). Additional state-of-art machine learning methods (L1-regularized logistic regression [66] and Random forest [41]) were trained and evaluated to test the robustness of FKT performance improvement. Finally, we demonstrate the power of our approach with an *in vivo* experiment validating the predicted role of wnt5b in establishing correct heart asymmetry in *Danio rerio*.

2.2.1 Functional knowledge transfer enables accurate gene prediction for pathways with few or no known genes

Most modern machine-learning methods that predict novel members of a biological pathway require a set of genes already known to participate in the pathway. These approaches are therefore limited to predicting genes to biological processes with sufficient prior knowledge in an organism [67]. For example, in the MouseFunc competition [25] (a broad competition focused on the performance of biological process prediction approaches), terms with less than three gene annotations were considered infeasible to predict and not included.

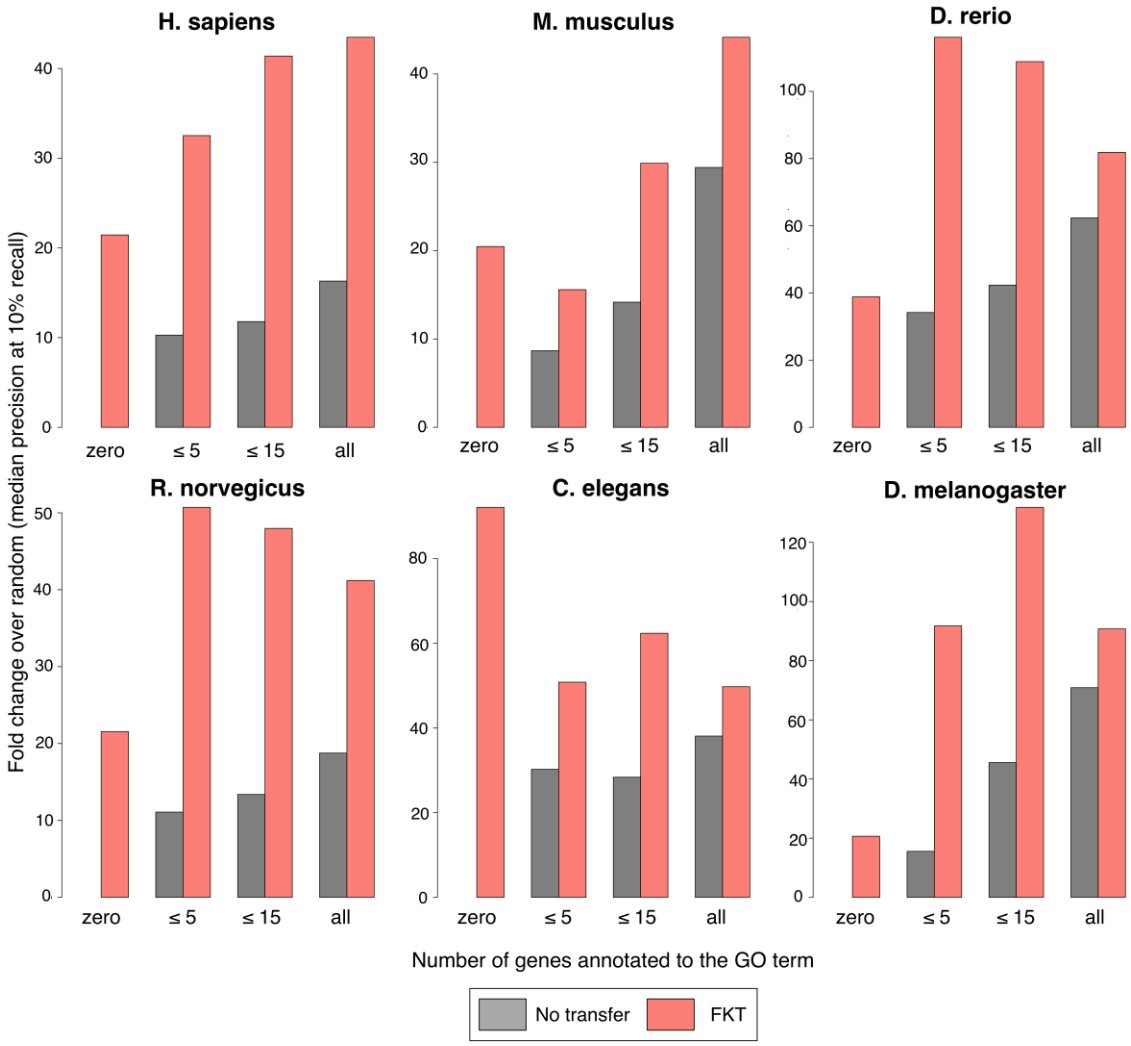

Figure 3. Cross annotation allows accurate recovery of small and unannotated terms.

All annotations accumulated after 5/11/2008 are held out from our prediction pipeline (as outlined in Figure 2) and are used for evaluation of prediction performance. 3,207 GO biological processes terms that acquired new annotations subsequent to our holdout date are grouped by organism and by the number of annotations at 5/11/2008 (zero, <=5, <=15, all). Performances at recapitulating future annotations are compared for a machine learning method (SVM) without (gray) and with (pink) including functional knowledge transfer (FKT) derived examples. For processes with zero annotations before 5/11/2008, no predictions can be made without cross-annotation (shown as absent performance bar). In all six metazoan organisms and for all process sizes, FKT improves prediction performance.

We address this constraint by leveraging knowledge across species, which allows us to take advantage of known biology from a model organism where the pathway of interest may be better studied. We applied our functional cross-annotation strategy (FKT) to biological processes with few known genes (annotation sizes of <= 5 and <= 15) in six metazoans and evaluated the predictive performance of an SVM trained with these annotations. To evaluate our performance, we constructed a three-year temporal holdout of experimental annotations. We used only biological process annotations added to Gene Ontology [68] before 5/11/2008 in learning the functional networks, transferring annotations across organisms, and predicting gene-process participation. New experimental annotations added to Gene Ontology between 5/11/2008 to 5/11/2011 were held-out and used for evaluation. In total, 3,207 GO biological process terms across the six organisms acquired new gene annotations in the subsequent three years. We evaluated the accuracy of our predictions with the gene-process assignments made during the hold-out time period in Figure 3. We observed substantial improvement using FKT when compared with only using the direct annotations for an organism. Improvement was evident across all six organisms, suggesting that even well characterized model organisms (e.g. mouse) can benefit from genomic-data-driven knowledge transfer. In addition, by holding out gene-process annotations acquired within the last three years, we could evaluate our ability to predict genes to processes which had no known genes in an organism prior to the hold out date (i.e. before 5/11/2008). Even though these processes were uncharacterized at that time, they subsequently became the focus of a directed experiment and thus were deemed biologically relevant and experimentally feasible in the organism. As shown in Figure 3, FKT gene predictions to these processes performed competitively even compared to biological processes with known gene annotations.

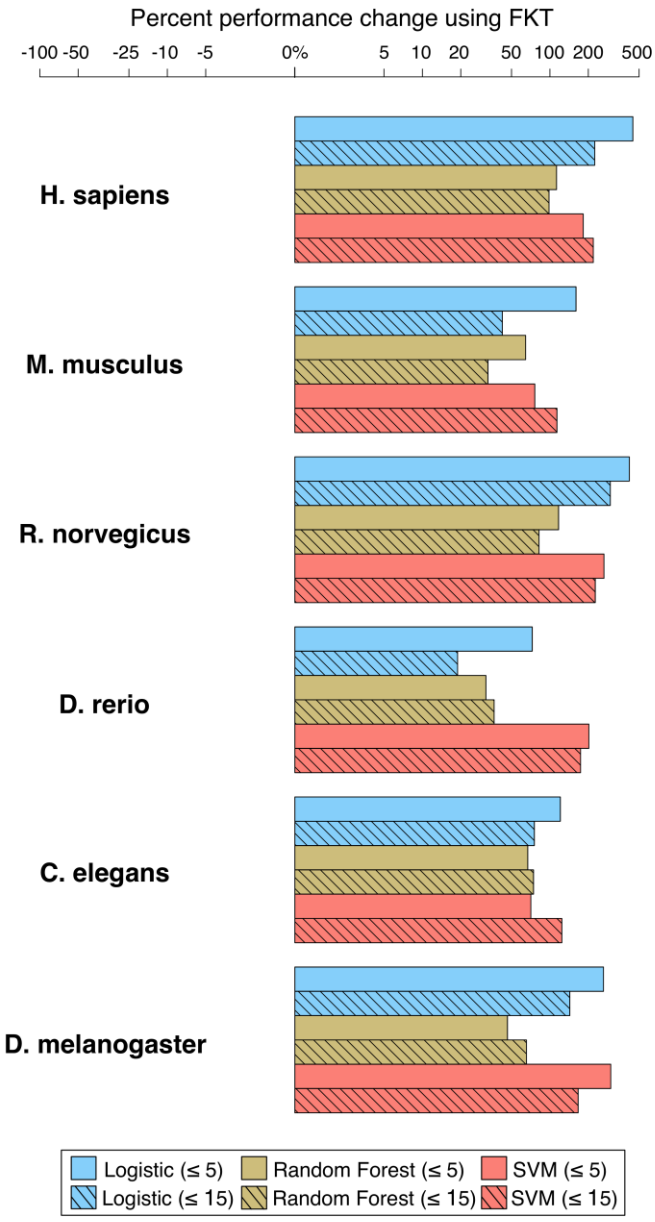

Figure 4. Functional knowledge transfer (FKT) improves prediction accuracy for a wide range of state-of-the-art classification algorithms.

The performance change when applying FKT are compared for each of three machine learning algorithms: L1-regularized logistic regression, Random forest and SVM (evaluated based on the ability to recapitulate held-out annotations accumulated after 5/11/2008). 3,207 GO biological process terms are shown, grouped by annotation size at 5/11/2008 (<=5, <=15). The percent change in performance (median fold over random) when applying FKT compared to no FKT with each machine learning algorithm is shown for six diverse organisms. All bars are to the

right of zero, indicating a performance improvement when FKT is applied for each machine learning algorithm.

We hypothesized that our transfer method could improve prediction performance for a wide range of machine learning methods. Machine learning algorithms are often based on distinct learning models and assumptions, thus any widely applicable annotation transfer method must be robust to not only the biological variability (e.g. different organisms or pathways) but also to this modeling variability. Thus in addition to SVM, we evaluated two widely used state-of-the-art learning methods: L1-regularized logistic regression [66] and Random forest [41]. We trained both classification methods with and without FKT and evaluated on the held-out set of annotations. FKT improved prediction accuracy across each machine-learning algorithm and organism (Figure 4). In particular, these improvements were consistent across biological process annotation sizes (<=5 and <=15). Altogether, these results indicated that FKT could recover biological processes that would be otherwise missed by most prediction methods, and that the transfer had wide applicability - improving performance across diverse organisms and machine learning algorithms.

2.2.2 Genes predicted to processes with no prior annotations in the study organism reflect subsequent experimental findings

We coupled FKT with an SVM and applied the machine learning classifier to predicting novel gene functions in six organisms. These predictions included gene-process membership for 8,091 GO biological processes currently without experimental annotations in at least one organism.

Supervised machine learning methods would be unable to predict novel genes to these biological processes without annotation transfer. They represent a wide range of biological pathways and processes ranging from development and metabolism to immune response and response to various stimuli.

For example, the biological process *regulation of exit from mitosis* (GO:0007096) represents a crucial mitotic cell cycle process that enables cells to regulate their exit from M phase. This process had no experimental annotations in *Danio rerio* at the time of our study, however had been extensively studied in the model organisms *Saccharomyces cerevisiae* [69], *Mus musculus* [70] and *Drosophila melanogaster* [71]. Our functional cross-annotation method has identified a total of 18 genes in *Danio rerio* with functional analogs annotated to this process (11 from yeast, 5 fly, 1 mouse and 1 rat), enabling novel predictions of gene membership to this process.

Our top gene prediction for this process, *cdh2*, has been experimentally confirmed in a recent study examining cell cycle progression in *cdh2* mutant retina cells [72]. Interestingly, *cdh2* is not only a novel prediction in *Danio Rerio* (i.e. this gene function was unknown at the time of our study), but also no *cdh2* homologs are known to be involved in the *regulation of exit from mitosis* in other organisms. *Cdh2* is a member of the cadherin protein family, which are important transmembrane proteins that play a crucial role in cell adhesion in multi-cellular organisms. Methods that employ only sequence similarity would be unable to predict this because *cdh2* homologs have not been annotated to this process in other model organisms. Furthermore, prediction methods without FKT will miss this finding because there are no existing *Danio rerio* annotations to this process. Only methods coupling FKT with a machine learning algorithm can take advantage of information from the single cell model organism

Saccharomyces cerevisiae, where cell-cycle checkpoints have been extensively studied [73], and successfully predict this finding in the multicellular model organism *Danio rerio*. This *in vivo* experimental result demonstrates FKT's utility for predicting novel genes to understudied processes. In addition, by coupling functional transfer to machine learning methods that leverage organism-specific functional data collections, we can make reliable gene-process predictions even without an annotated sequence-homolog.

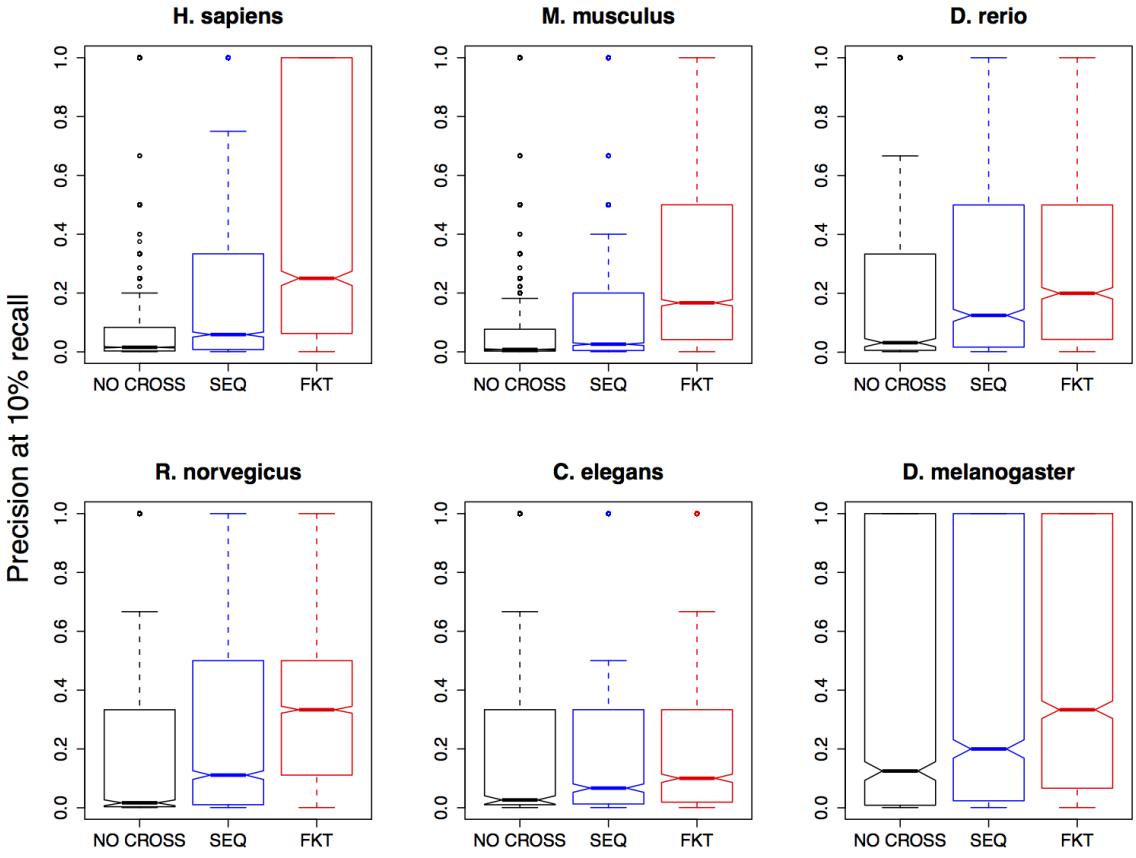

Figure 5. Functional knowledge transfer (FKT) improves performance for predicting small processes.

The performance of two knowledge transfer methods (FKT and sequence-only) and a baseline method (with no cross-annotation) are compared. Shown here are results of threefold cross-

validation for small processes (<= 15) that represent specific or understudied pathways. FKT paired prediction method shows improved performance compared to both sequence-only transfer and the baseline method.

2.2.3 Cross-annotation among functional analogs improves prediction accuracy for small processes

To compare our functional transfer method, which applied a more specific annotation transfer, to commonly used methods that used only sequence homology, we evaluated a method that did not leverage functional similarity and a baseline method without any cross-annotation. In this sequence-only method, all homologous gene pairs (reciprocal BLAST best hit gene pairs) were targets for annotation transfer - any biological process annotated to a gene was transferred to its reciprocal best-hit gene in all organisms. To obtain a representative set of gene-process annotations for evaluation, we conducted a threefold cross-validation on genes that had experimental biological process annotations, and evaluated the SVM classifier prediction performance on each corresponding held-out set of biological process annotations. The results of the comparison showed that although both methods improved performance for small processes, FKT showed greater performance gains (Figure 5). In a few organisms, the performance gains were substantial - for example, in human and mouse, the median performance (precision at 10% recall) increased more than fivefold.

Upon examining the processes that improved the most when compared to a sequence-only method, many pathways and processes with transcriptional based regulatory control showed improved performance using FKT. *Response to mechanical stimulus*, *ameboidal cell migration*, *regulation of neuron differentiation* and *striated muscle cell development* were among the top

improved processes in all organisms using FKT compared to sequence-only. Unsurprisingly, these processes have been well known to be tightly regulated through transcriptional programs (e.g. stress response, developmental TF gradients) [74-76] and have multiple datasets measuring the transcriptional profiles incorporated in our functional networks [77-79].

We expect that FKT will continue to improve as the functional genomics compendia for many organisms continue to grow, including expression and other types of measurements across multiple perturbations. An additional advantage of a functional genomics similarity approach, as shown in [24], is the ability to differentiate functional differences in tissue specificity between sequence homologs. The example of mouse RNA polymerase II elongation factor *Supt5h* and its direct sequence-ortholog *C. elegans spt-5* highlight this issue. FKT determined these sequence-orthologs as not being functional analogs. Indeed, mouse Supt5h is predominantly neuronal, while *C. elegans* SPT-5 is non-neuronal and primarily expressed in the intestine and pharynx [80-82]. Even though these sequence-orthologs have diverged in tissue specificity, they still share high sequence similarity and a sequence-only method would inappropriately transfer all functional annotations between them.

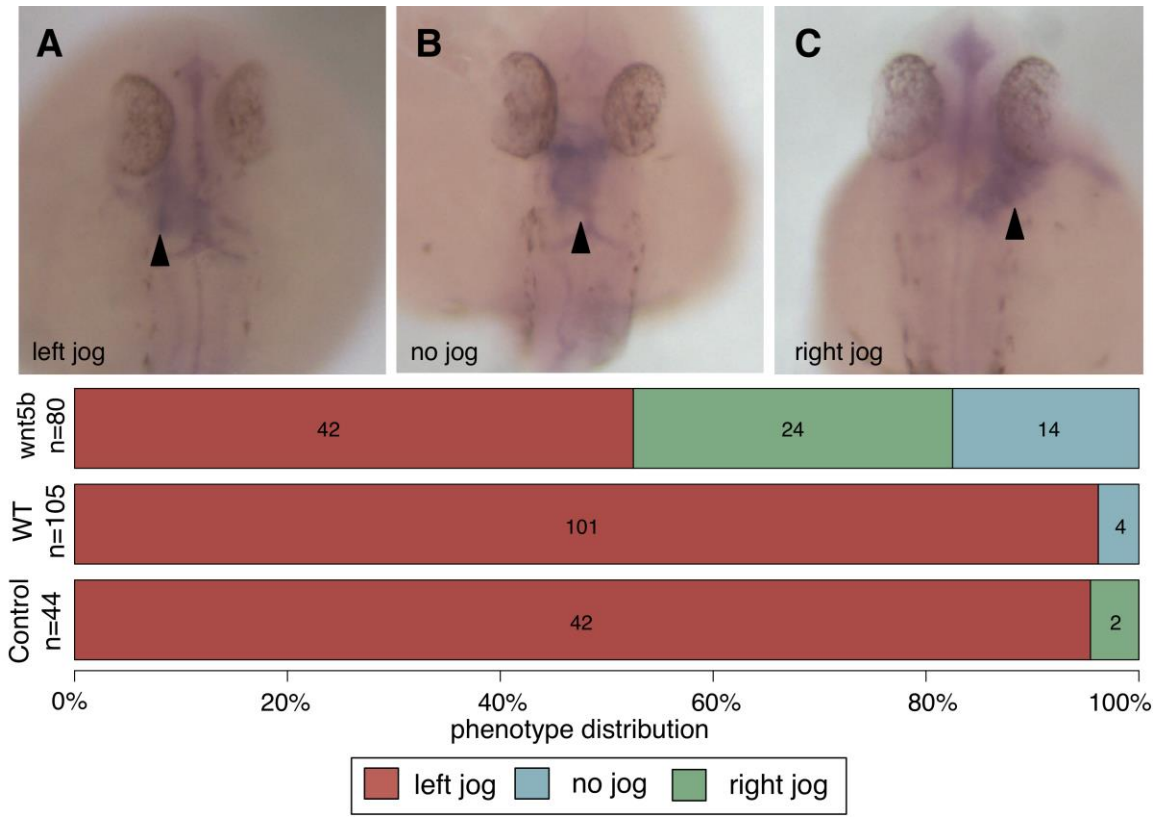

Figure 6. Knockdown of wnt5b leads to defects in zebrafish heart asymmetry.

Morpholinos (MO) against *wnt5b* were injected into zebrafish embryos at the 1-2 cell stage. Embryos were evaluated for heart jogging at 27 hour post fertilization and scored as either left (C), right (B), or no jog (A). While control MO injected embryos had predominantly left-jogged hearts, embryos injected with the wnt5bMO displayed randomized heart jogging with 48% of embryos displaying right or midline jog.

2.2.4 *In vivo* validation of *Danio rerio* gene *wnt5b* involvement in the establishment of heart asymmetry

In all vertebrates, the heart develops with a distinct left-right (L-R) asymmetry during embryonic morphogenesis. Deviations in left-right heart development can lead to complex congenital heart defects that are among the most common human neonatal diseases [83,84]. During cardiac morphogenesis in *Danio rerio*, two distinct stages of cell migrations lead to the final

asymmetries of the heart. In the first stage, called “heart jogging”, myocardial cell migration within the cardiac cone place the ventricular cells to the left side, while atrial cells remain near the midline. In the second stage of “heart looping”, the heart tube bends and forms a loop that places the ventricle to the right of the atrium. Although the steps of cell migration progression leading to left-right heart asymmetry are beginning to be explored [85-88], an understanding of how it is achieved mechanistically is still lacking.

In Gene Ontology, the biological process term “determination of heart left right asymmetry” (GO:0061371) represents the developmental pathways regulating heart jogging and looping. To validate our prediction method (FKT coupled with SVM), we investigated the top five predicted genes that had not already been annotated to this GO term: *sox32*, *wnt5b*, *ndr1*, *tbx1* and *lft1*. We found existing literature evidence confirming the involvement of four of the five genes (*sox32* [89-92], *ndr1* [93], *tbx1* [94], and *lft1* [95-97]). Although there existed experimental results confirming the role of these genes in influencing heart asymmetry, these results had not yet been curated by GO annotators. For example, in a knock-out experiment of our top predicted gene (*sox32*/*casanova*), *Danio rerio* embryos had fewer dorsal forerunner cells which led to defects in Kupffer’s vesicle formation and subsequent left-right patterning of the heart, confirming that *sox32* was required for proper establishment of heart asymmetry. The only gene among the top five without existing experimental support was *wnt5b*, our second ranked prediction after *sox32*. Previous work had shown the involvement of *wnt5b* in cell migration during gastrulation [98] but the gene had not been specifically associated with heart left-right asymmetry regulation. To experimentally validate our prediction of *wnt5b* to left-right heart determination, we knocked down its function by means of morpholino antisense oligonucleotides (MO) [99].

A significantly greater proportion of embryos where *wnt5b* was knocked down with a morpholino (Figure 6) had a defective heart jogging phenotype (Fisher's exact test p-value < 0.001). In total, 48% of morpholino treated embryos showed either right-sided heart jog or midline jog comparable to previous genes known to be involved in this biological process [100-102]. Only 4% of wild type and control-MO treated embryos exhibited this phenotype. This phenotype is likely due to the disruption of asymmetric expression of the TGFbeta member Nodal (data not shown), which is typically asymmetrically expressed on the left side of vertebrate embryos during somitogenesis. Left-sided Nodal in *Danio rerio* myocardial cells directs their subsequent migration during asymmetric cardiac morphogenesis [85,88]. Further investigation would be necessary to understand the mechanistic role of *wnt5b* in left-right heart determination, however our *in vivo* experiment confirmed the regulatory role of *wnt5b* in *Danio rerio* left-right asymmetry determination in heart development, as our method predicted.

2.3 Discussion

This study demonstrates that state-of-the-art machine learning methods coupled with our functional knowledge transfer method can accurately prioritize novel genes of understudied processes. Previous methods have focused on incorporating functional genomic data primarily as input data [103-106]. In contrast, here we demonstrate that the prevalence of understudied processes and the abundance of genomic data provide an opportunity to improve the accuracy of cross-organism annotation transfer and extend prediction coverage to processes with no prior annotations. We now integrate FKT into our IMP web-server [43]. This makes IMP a web

interface for exploratory analysis covering all organisms included in this study across 10,653 biological processes (http://imp.princeton.edu). Functional knowledge transfer allows IMP to also include gene predictions for processes currently unannotated in an organism. Although in our current study we have experimentally followed up on our top predicted gene, all of our predictions in IMP are shown with estimated probabilities allowing biologist to draw a threshold dependent on how much the assay costs, and how important it is to find true positives (versus not finding false positives). In addition, the website includes the Bayesian functional relationship networks that were used for mapping functional analogs and used as input features to the machine learning methods. In particular, to the best of our knowledge, we include the first zebrafish (*Danio rerio*) functional relationship network.

We anticipate that our approach can be extended beyond the six organisms shown in this study, as it is especially beneficial in organisms that have high-throughput genomic data with sparse annotations (e.g. frog, slime mold). Next-generation sequencing is further increasing the diversity of organisms that are measured on the genome-scale, and functional knowledge transfer can help us understand and annotate the roles of genes in such emerging model systems. Functional knowledge transfer allows for accurate hypothesis generation and experiment guidance even for pathways with no previous experimental knowledge in a given organism, thus benefiting human biology, broadly studied organisms such as mouse and fly, and newly adopted model systems.

2.4 Methods

We developed a functional knowledge transfer method and applied this method to predicting gene functions in six organisms using a functional network based classification strategy. In summary, data integration was performed using a regularized naive Bayes classifier, which summarized the data compendium into organism specific function relationship networks. Edges in functional relationship networks represented, given all collected data from that organism, the posterior probability of a gene pair co-functioning in the same biological process. Next, a collection of organism specific experimental annotations supplemented with cross-annotated gene annotations (based on both sequence and functional similarity) was used as gold standard for each GO biological process to train a GO term specific SVM with the functional relationship network as features. To test for robustness across different machine learning algorithms, L1-regularized logistic regression and Random forest were also evaluated by coupling both algorithms with the functional knowledge transfer method. Final predictions were made on a total of 10,653 unique biological processes. We experimentally validated our method's predictions for the determination of heart left-right asymmetry in *Danio rerio*. Of our top five predictions, four were validated via existing but un-curated experiments from the literature. We validated the fifth, *wnt5b*, using a morpholino knock-down assay.

2.4.1 Integration and Summary of Organismal Data Compendia

Data source and pre-processing

We collected 2,444 microarray datasets from NCBI Gene Expression Omnibus (GEO) covering a total of 43,865 conditions across seven model organisms. Probes were collapsed and

normalized according to the procedure described in [106] and the Fisher's z-transformed pearson correlation were calculated for each gene-pair as described in [52].

Physical and genetic interaction data from BioGRID [107], IntAct [108], Mint [109], and MIPS [110] were processed as counts of experimental assays that support an interaction between two genes (e.g. a gene pair with evidence from two-hybrid and western blot would receive two counts). Potential transcription factor (TF) to target gene associations were obtained from Yeastract [111] and TF binding site motifs retrieved from Jaspar [112]. Yeastract's predicted TF-gene relations were treated as pair-wise binary scores. For Jaspar, we searched for possible transcription factor binding sites by scanning each TF profile in 1 kb upstream sequence of all genes using FIMO [113]. Motif matches were treated as a binary score (present if p-value < .001 and not-present otherwise) and the final gene pair score was obtained by calculating the pearson correlation between the two genes' binary score vectors.

Phenotype and disease data from SGD [114], MGI [115], Wormbase [116], Flybase [117], GSEA [27], Zfin [118] were incorporated into our functional networks by summing the co-occurrences of gene pairs in all phenotypes/diseases and normalizing by the size of the phenotype/disease. For gene pair, *i, j* the scoring function is the following:

$$S(i, j) = \sum_{k=1}^{n} \frac{I_k(i) I_k(j)}{N_k}$$

where function $I_k(i)$ and $I_k(j)$ are the indicator functions that have the value 1 when gene *i* or *j* is annotated to the phenotype or disease, *n* is the total number of phenotypes/diseases, and N_k is the total number of genes associated with the phenotype or disease *k*. Protein sequence similarity

between genes was obtained from Biomart [119], and protein domain data were treated as binary evidence from PfamA [120] and Prosite [121].

Generating functional relationship networks

To summarize the processed heterogeneous genomic data, we generated one global functional relationship network per organism. We applied Bayesian integration, specifically a naïve Bayes classifier, to systematically deal with differences in accuracy and relevance of each data source for predicting gene functional relations. Gene pairs co-annotated to a set of 433 expert selected Gene Ontology [68] biological process fringe terms were used as known functionally related genes (i.e. positive examples) [106,122]. Gene pairs not co-annotated to any terms in the GO fringe, KEGG [123], PID [124] or Biocyc [125] were considered as unrelated (i.e. negative examples) except in the following cases:

1. A gene pair was annotated to terms overlapping with a hypergeometric P-value below 0.05
2. A gene pair was annotated to a set of 'negative' GO terms that define minimal relatedness (as described in [106])

If a gene pair met either of the two conditions, it was excluded from unrelated pair generation (i.e., they were neither related nor unrelated for training). Thus this formed a set of global related and unrelated gene pairs to be used for training and evaluation.

One binary regularized naive Bayes classifier was trained per Gene Ontology fringe term (i.e. biological process/context). Each naive Bayes classifier contains one class node determining the membership of a gene-pair to the biological process and organism specific dataset nodes

conditioned on the class node. When integrating large number of genomic datasets, the naive Bayes assumption of conditional independence among datasets can no longer be justified. We have shown that a mutual information based parameter regularization for naive Bayes classifiers can alleviate the conditional dependency among datasets [106]. In this work, we make modifications to our prior method by directly estimating the conditional dependency between a dataset by limiting the mutual information calculation between datasets to gene-pairs that are not functionally related. This heuristic enables us to estimate the conditional dependency between datasets without having to regress out the incomplete functional relation class node information. Specifically, the heuristic sum of shared information U_k is now:

$$U_k = \frac{\sum_{i \neq k} I_{pairs \in negative}(D_k, D_i)}{H(D_k)}$$

$$\alpha_k = 2^{U_k} - 1$$

where $I_{pairs \in negatives}$ is the mutual information between dataset D_k and D_i among gene pairs not known to have a functional relationship (i.e. negative gene pair examples) and H is the single dataset entropy. Then we use the exponential decreased ratio (α_k) to weight a given dataset's likelihood function. Finally, the naive Bayes functional relationship posterior probability for gene pair *i,j* is the following:

$$P^*(D_k = d_k(g_i, g_j) \mid FR = 1) = P(D_k = d_k(g_i, g_j) \mid FR = 1)\left(\frac{n_s}{n_s + \alpha_k}\right) + \frac{1}{|D_k|}\left(\frac{\alpha_k}{n_s + \alpha_k}\right)$$

$$P_{g_i,g_j}(FR=1 \mid D) = \frac{P(FR=1)\prod_{k=1}^{n} P^*(D_k = d_k(g_i, g_j) \mid FR=1)}{P_{g_i,g_j}(D)}$$

where the weighted dataset likelihood function is P^*, $d_k(g_i, g_j)$ is the experimental value for gene pair i,j, $|D_k|$ is the total number of discretization levels and n_s is a pseudocount set to 3 in our integration based on cross-validation results.

Finally, with biological process specific functional relation networks predicted by each naive Bayes classifier, we averaged the edge probabilities from each process specific functional network to generate the final global functional relationship network.

2.4.2 GO biological process gold standard construction through cross-annotation

In total, 10,653 GO biological process terms were predicted for new gene annotations covering six organisms, *Homo sapiens, Mus musculus, Rattus novegicus, Drosophila melanogaster, Danio rerio* and *Caenorhabditis elegans.* We limited the positive examples for each GO term to propagated experimental GO annotations with GO evidence codes EXP, IDA, IPI, IMP, IGI and IEP (all "NOT" annotations were removed). In addition, to leverage the research strengths across organisms, we transferred gene annotations among six organisms plus yeast, first based on sequence similarity and second filtered by function similarity. In detail, we start with all sequence paralog and ortholog gene relations within each TreeFam [63] gene family. Next, based on our previous algorithm [24], we filtered for functional analogs among all paralog and ortholog gene pairs using our functional relationship networks. We define a functional analog to be a gene pair that has a significant number of overlapping TreeFam gene families among its closest gene

neighbors in the global functional relationship network (a functional network is converted into a binary network by using a probability cutoff of 0.5). We defined a gene pair's score as the following:

$$S_{G1,G2} = \sum_{i=k}^{\min(m,n)} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$$

where *m* and *n* are the number of TreeFam gene families in each gene *G1* and *G2*'s direct neighborhood in the functional network, *k* is the number of overlapping TreeFam gene families between gene *G1* and *G2* gene neighbors and *N* is the total number of TreeFam gene families around gene *G1* and *G2*. The functional similarity score is the probability of observing greater or equal to the number of overlapping gene families by chance, thus can be interpreted as a hypergeometric p-value. We used a score cutoff of <= 0.01 to consider a gene pair as functional analogs.

Finally, all experimental annotations are propagated between functional analogs. In total, our supervised functional knowledge transfer allowed us to make predictions for 8,091 additional GO biological processes, thus extending our predictions beyond simply well-studied and well annotated processes and pathways.

2.4.3 Biological process prediction with network based SVM

We used the augmented gold standard genes by functional knowledge transfer and functional relation network as features into state-of-art machine learning algorithm Support Vector Machine

(SVM) to predict novel biological process gene annotations. Our functional relation network based SVM method has shown to outperform methods that directly input the raw data into the SVM or a simplistic sum of the functional networks to the positive examples [126].

For each biological process, the feature space was constructed as the weights in the functional relation network. Thus for each gene example, all gene edge weights connecting to the example gene were used to create the feature vector. Therefore, each organism feature count will be equal to the number of genes in the organism. The set of feature vectors for training examples were used to train a linear SVM according to the standard formulation:

$$\min_{w,\xi_i \geq 0} \frac{1}{2} w^T w + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

$$\forall i : y_i \left(w^T x_i \right) \geq 1 - \xi_i$$

where n is the of training example genes, w is the gene weight vector, y_i is the training label of gene i and x_i is the edge weight vector connecting gene i to all genes in the functional network.

Finally, the unbounded SVM prediction scores were transformed into probabilities based on a maximum likelihood sigmoid fit to the SVM outputs [127].

2.4.4 Additional machine learning algorithms

To validate that the observed performance improvement was not specific to any single learning algorithm, we applied the functional knowledge transfer to two additional widely used machine learning methods: L1-regularized logistic regression and Random forest. Regression analysis coupled with regularization has been a broadly used approach to control for the bias-variance

trade-off [128]. In particular, L1-regularization has been successfully used in many methods for shrinkage and feature selection applications, most famously in the works of LASSO [129]. By coupling L1-regularization with a logit link function, conditional probabilities of a gene membership to a biological process can be computed based on selected genes of predictive power. The predictive gene weight vector *w* was obtained by the following regression problem:

$$\arg\min_{w} \sum_{i=1}^{m} \log\left(1 + e^{-y_i w^T x_i}\right) + \lambda \sum_{i=1}^{n} |w_i|$$

where $\lambda > 0$ is the regularization parameter, y_i is the training label of gene i and x_i is the edge weight vector connecting gene i to all genes in the functional network.

Random forest classifiers are a combination of decision trees that are aggregated to make a final prediction. Random forest algorithms have been shown to produce improved prediction accuracy compared to a single decision tree by better estimating the contribution of each predictor through random sampling [41]. In genomic applications, Random forest has gained interest due to the many high-dimensional genomic learning problems [130]. Formally, random forest is defined by the following:

$$RF = \{h(X, d_i), i = 1, \ldots, n\}$$

where the random forest *RF* is a set of $h(\cdot)$ decision tree functions, trained on training examples *X* and a bootstrap sample d_i from the original feature space of *D*. For classification, the votes of each *n* decision trees are averaged as shown in the following:

$$av_n \sum_{i=1}^{n} I(h_i(X))$$

where $I(\cdot)$ is the indicator function for the prediction class of interest. In our study, for each GO term 61 decision trees were trained on independent bootstrap samples of our original genomic training data.

2.4.5 Performance evaluation

For performance evaluations for GO terms with no prior annotation, we used a three-year temporal holdout set of gene annotations for each GO biological process. The held-out gene annotations were preserved throughout the prediction pipeline (functional network integration and SVM predictions) to avoid any overestimation of performance. Although we train our SVM classifiers using the augmented cross-annotated gold standard, only the non-transferred experimental GO term annotations were used for evaluation with all transferred annotations excluded.

The GO gene association files used to create our gold standard was downloaded from Gene Ontology [68] on 5/11/2011. To generate an accurate temporal three-year holdout we downloaded the GO gene association version archived at 5/11/2008. All annotations were propagated and only experimental examples newly annotated after 5/11/2008 for each GO term was used in the temporal evaluation. Accordingly, any GO term that had no gene annotations on 5/11/2008, but subsequently accumulated new annotations were used to evaluate our performance in predicting terms with no-known prior annotations.

To compare performance between knowledge transfer methods, we conducted an evaluation by performing a threefold cross-validation among genes that had experimental biological process annotations. This set of evaluation annotations represents a random sampling of the current knowledge base as of 5/11/2011. Identical to our temporal holdout, all evaluation annotations for

each holdout were withheld from our prediction pipeline to avoid any performance over-estimation.

2.4.6 Implementation

All software used in this study has been implemented in the open source and publicly available Sleipnir library [131] available from http://libsleipnir.bitbucket.org, which interfaces with the SVMperf library [132] for linear kernel SVM classifiers (the error parameter C was set to 100 for these experiments through cross-validation). L1-regularized logistic regression used the LIBLINEAR [66] and Random forest used the MILK (Machine Learning Toolkit) python package implementation with 61 decision trees per GO term.

2.4.7 Morpholino Microinjections and Scoring of Heart Left-right Asymmetry

The *wnt5b* morpholino (MO) and standard control MO were purchased from GeneTools. The sequence of the *wnt5b* MO used is as follows: 5'-GTCCTTGGTTCATTCTCACATCCAT-3'. Morpholinos were injected at a concentration of 6ng/uL into the yolk of one-cell stage embryos for whole knockdown in the embryonic cells. Initial assessment (Figure 6) was performed via *in situ* hybridizations on fixed embryos using the standard protocol [133] with *cmlc2/myl7* used as a probe. Images were captured at 4× or 10× magnification using a ProgressC14 digital camera (Jenoptik) on a Leica MZFLIII microscope.

Heart laterality for each treatment (*wnt5b* MO, control MO, wild type) was evaluated in live Tg(*cmlc2::GFP*) embryos at 27 hours post fertilization. Embryos were scored as left/right/no

jog based on expression of GFP driven by *cmlc2*'s heart specific promoter using a Leica SP5 confocal microscope.

3 Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components

This work was conducted with the experimental collaboration with David C. Hess.

3.1 Introduction

The complexity of cellular activity is driven not only by interactions among genes and gene products, but also by the timing and dynamics of these interactions, the conditions under which they occur, and the many forms that they can take. Proteins interact in many different functional manners with multiple partners - physically in complexes[134] and through modifications[135,136], synthetically when employed in parallel pathways[4], and in regulatory roles as activators or repressors[137] - and these interaction types combine to form complete molecular pathways. Functional assays such as gene expression, localization, and binding each capture individual aspects of this molecular activity at a global level, but translating the vast amount of resulting genomic data into specific hypotheses at the molecular pathway level has

proven challenging. The heterogeneity of gene interactions within each pathway has compounded this difficulty by preventing any one assay from providing a complete biological picture. It is thus critical to integrate large genomic data collections to describe not only the membership of gene products within pathways, but also their construction from the building blocks of individual types of biomolecular interactions.

In this work, we provide the means for investigators to study complete molecular pathways at a whole-genome level as generated from integrated functional genomic data. First, we relate 30 general and specific biomolecular interaction types, such as transcriptional regulation, ubiquitination (and other post-translational modifications), or protein complex formation, in an ontology of interaction types. This ontology is hierarchical, in that a phosphate transfer is perforce a covalent post-translational modification, which is in turn by definition a transient physical interaction, and so forth. Next, we combine this ontology with Bayesian hierarchical classification methodology [138], enabling the simultaneous prediction of genome-wide interaction networks of all of these 30 types from integrated heterogeneous experimental data. Finally, we apply this method to a compendium of ~3,500 *Saccharomyces cerevisiae* experimental conditions, experimentally validating several of the resulting predictions in glucose utilization, DNA topological maintenance, and protein biosynthesis as described below. This methodology ensures that investigators can take advantage of all available data to accurately identify the entire range of functional interaction types within specific pathways and across an organism's genome.

It is important to contrast this genome-wide system for predicting diverse biomolecular interaction types with previous work predicting specific individual interaction networks. A

variety of methodologies have been proposed for inferring regulatory networks [139-142], physical interaction networks [35,143], synthetic interaction networks [144,145], and other interaction types[146], generally from their respective primary data types (ChIP-chip and -seq, proteomics, double knockouts/knockdowns, etc.) Likewise, other methods have been proposed for heterogeneous genomic data integration [28,36,106,144,147-151], but these almost uniformly focus on either general functional interactions or on specific bimolecular interaction types. This work combines the strengths of these two bioinformatic areas, providing a simultaneous platform with which all data available for a system can be integrated and focused onto specific interaction types, genome-wide and for individual gene products.

We first applied our yeast network compendium to explore two cellular processes, carbon metabolism and cellular transport. This generated many promising interactions involving Snf1, Cmk2, Glc7, Adr1 and Gph1 supported by recent published work. We also suggest several novel pathway connections, such as the interplay between the glycogen breakdown and glucose utilization pathways, by systematically layering multiple different interaction types. To experimentally validate a collection of our predicted yeast interactions, we focused on the synthetic lethal interactions, where double knockouts result in lethality, predicted among proteins involved in DNA topological change and regulation of protein biosynthesis. Highly ranked 20 protein pairs, 10 pairs from each pathway, were hypothesized to be synthetically lethal, and we experimentally confirmed 14 of these pairs (70%). Furthermore, we evaluated our posttranslational modification predictions based on recent experimental results on 173 protein pairs, resulting in a prediction AUC over 0.8. In an analysis of the systems-level global and local network structures of our interactomes, we observed differential usage of several recurring

subgraphs, providing insight into the functional design principles of pathway components. Finally, we provide a web-based interface to explore all 30 yeast interaction networks at http://function.princeton.edu/bioweaver. This will allow investigators to interactively survey and generate hypotheses from the diverse interaction types comprising the *S. cerevisiae* cellular circuitry.

3.2 Results

We present a general methodology for integrating large, diverse genomic data compendia to simultaneously predict multiple biomolecular interaction network types (physical, genetic, regulatory, etc.; Figure 7). We applied this methodology to ~3,500 *S. cerevisiae* experimental conditions to generate 30 whole-genome networks describing predicted gene and gene product interactions in yeast. We first evaluated these predictions quantitatively using cross-validation, achieving AUCs over 0.7 for most interaction types. More qualitatively, we examined a set of predicted molecular linkages of diverse types between glycogen breakdown and glucose utilization genes, which were validated by recent literature. Finally, we experimentally confirmed 14 of 20 predicted novel synthetic lethal interactions in the DNA topological change pathway.

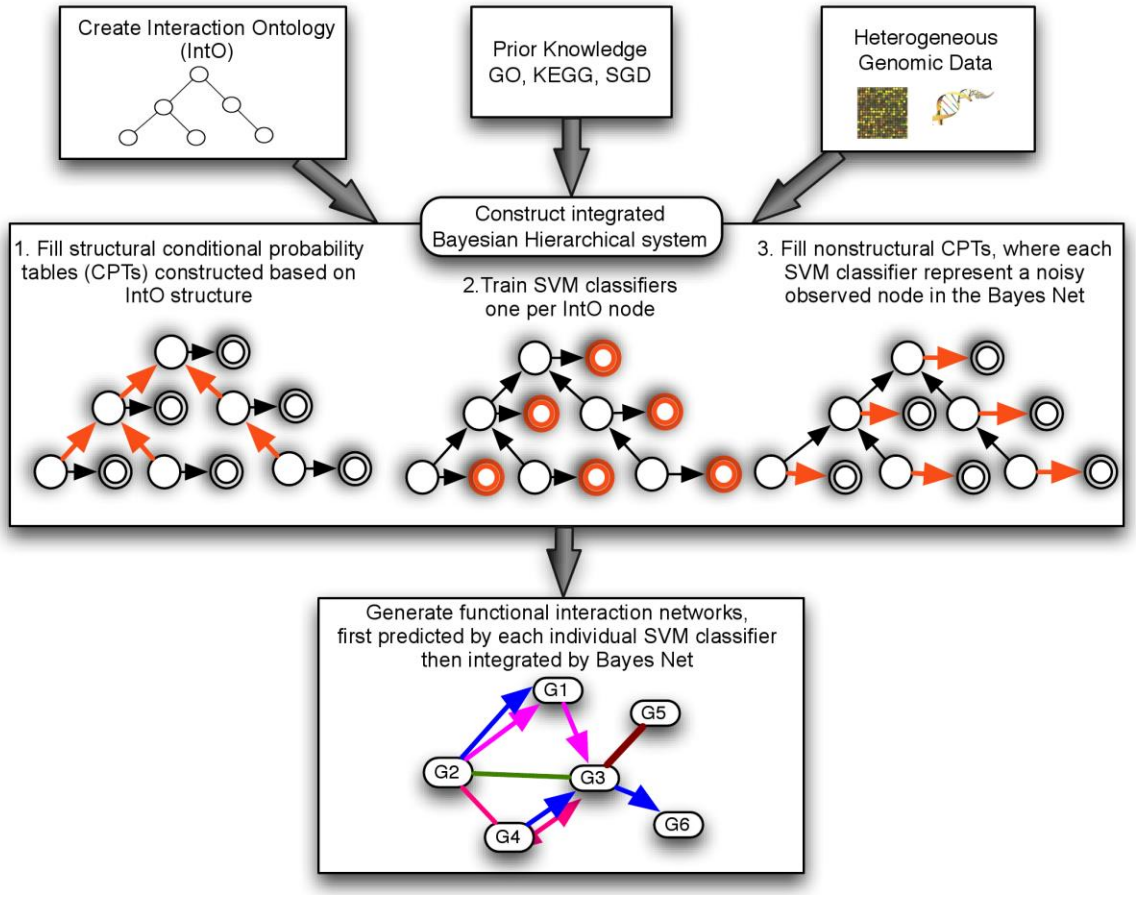

Figure 7. Overview of our integrated Bayesian hierarchical system for inferring diverse interaction networks.

An interaction ontology was constructed categorizing gene interaction types. A corresponding Bayesian network was constructed in which each node represents one interaction type. This network's structural parameters, P[*parent node label* |*child node labels*], were first determined using prior knowledge from GO, KEGG, SGD, and other curated sources. Second, individual SVM classifiers were trained to predict each interaction type in isolation using heterogeneous data sources. Third, the non-structural Bayesian network parameters, P[*true latent node label* |*SVM output*], were filled by relating each observed SVM classifier to a latent interaction type membership node using cross validation. Finally, to generate new predictions, a gene pair's interaction type is first predicted by the SVM classifiers and then hierarchically resolved by finding the most probabilistically consistent set of label assignments corresponding to the latent nodes in our Bayesian network.

3.2.1 Evaluating the accuracy of predicted S. cerevisiae biological networks

We predicted a compendium of biomolecular interaction networks by integrating diverse yeast genomic data using a multi-label hierarchical classification system ([138], Figure 8A). As briefly outlined in Figure 7, we first independently predict each interaction type using specifically trained SVM classifiers. Next, it is desirable to avoid making inconsistent interactome predictions due to noisy data, e.g. predicting that two genes share a regulatory relationship without occurring within the same pathway. In order to share information among classifiers for related interaction types in a principled manner, each SVM's predictions are treated as noisy observations. The final set of labels for each gene pair is then derived by finding the maximum likelihood assignment of interaction labels by integrating these observations in a Bayesian graphical model.

Based on ~30% heldout test data, our average AUC over all 30 interaction types was 0.79, with minimal variations in performance across the interaction ontology (Figure 8A). The most general interaction type, *functional relationship*, also incurred the lowest AUC of 0.63, which remains comparable to state-of-the-art functional interaction prediction systems [152]. In order to quantify the contribution of our hierarchical Bayesian system relative to predicting disparate biomolecular interaction types in isolation, we compared the accuracy of each individual SVM classifier to that of the complete system. For all 30 predicted interactomes, the Bayesian hierarchy showed increased AUC scores, averaging +0.076 and ranging from a minimum of +0.011 to a maximum of +0.166. For example, posttranslational regulation improved from 0.61 to 0.77, while phosphorylation increased from 0.67 to 0.79. In combination, these two evaluations suggest that this methodology can accurately leverage large genomic data collections to simultaneously infer a diversity of genome-wide interaction networks.

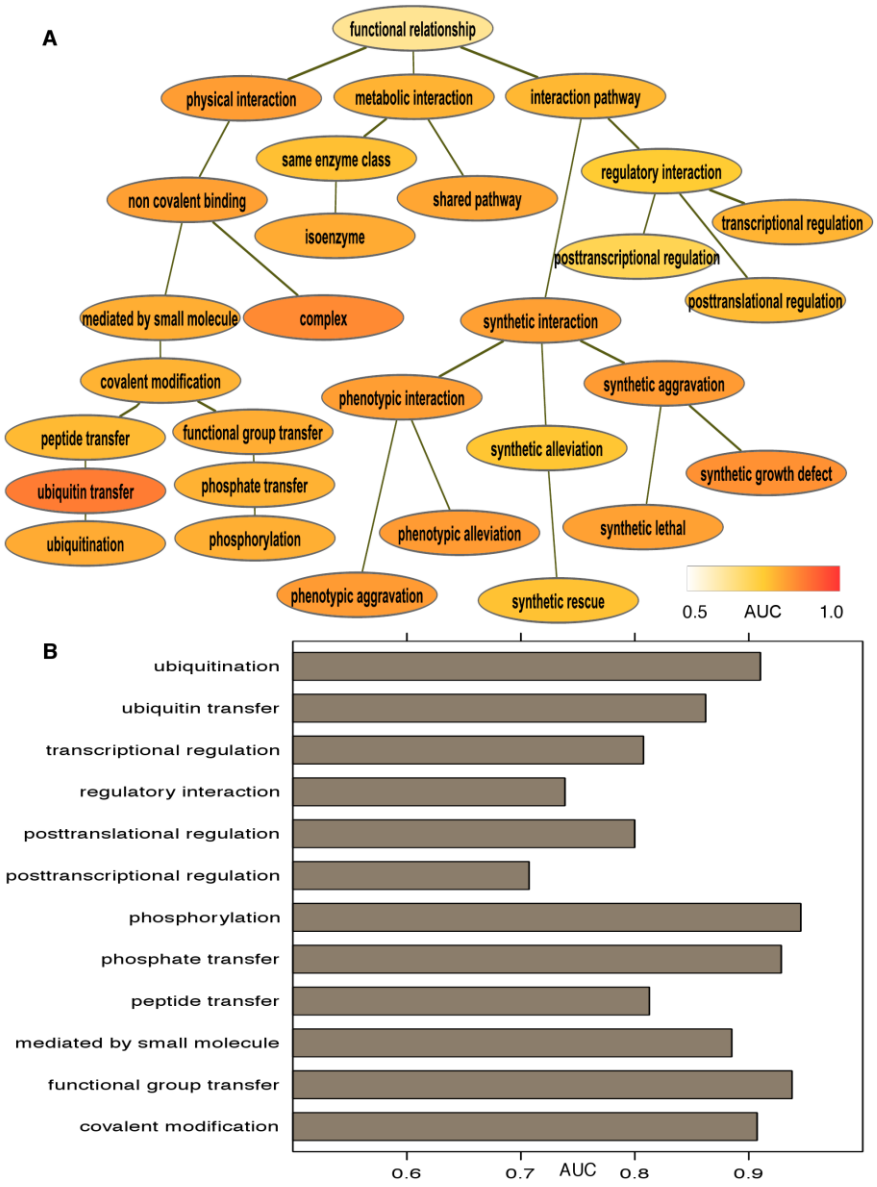

Figure 8. Performance evaluation of inferred networks.

We predicted 30 *S. cerevisiae* interaction networks, each representing one interaction type. A) To evaluate the overall accuracy of these networks, we withheld ~30% of the genes in our gold standard as a test set. Performance on this test set averaged an AUC of 0.79 across all interaction types in the ontology. B) To specifically assess the accuracy with which interaction directionality was predicted (as opposed to the presence/absence of interactions in part A), we tested the frequency with which an interaction's correct direction was ranked above its incorrect direction in each of the 12 directed interaction networks. These results are uniformly well above random (0.5), supporting our ability to accurately predict both the presence and the directionality of many specific types of protein interactions.

3.2.2 Accurate prediction of directed interaction networks

Many gene interactions are directional and thus asymmetric, e.g. phosphorylation or ubiquitination, in which the two interactors take on distinct source and target roles. It is thus important to correctly infer not only the presence or absence of these directed interactions, but also the correct directionality. Specifically, for each directed interaction type, we constructed a list of all edges ranked by predicted probability; we then compared the rank of the correct interaction direction relative to the incorrectly flipped interaction between the same two genes. Using this as a true- and false-positive rate criterion, we were able to predict the correct direction of gene interactions with average AUC of 0.85 over the 12 directed networks (maximum 0.94, minimum 0.70). This indicates that this methodology can accurately recover not only overall pathway structure, but also the upstream and downstream effects of individual gene products within molecular pathways.

3.2.3 Predicted interactions provide mechanistic insight into the yeast glycolysis pathway

Simultaneous inference of biomolecular networks for many different interaction types allows the generation of very specific novel hypotheses. As a first example, we detail a combination of transcriptional, genetic, post-translational, and metabolic interactions among gene products coordinating glycogen breakdown and glucose utilization in yeast.

As shown in Figure 9, Adr1 is an important transcription factor involved in carbon metabolism in *Saccharomyces cerevisiae*. It has many known regulatory inputs [153], one of which is the glucose-responsive kinase Snf1, and what proteins transmit this regulatory information has been under investigation for some time. By examining different classes of predicted interactions with Adr1 and other proteins *not* in our gold standard (Figure 9A), we first identified regulatory and genetic interactions between the protein phosphatase Glc7 and Adr1. Specifically, our prediction of a synthetic alleviating interaction between Glc7 and *adr1* mutants places it upstream of Adr1 in this pathway. This combination of interactions is almost always associated with an upstream inhibitory regulator, consistent with the known biological role of Glc7 as a protein phosphatase that removes activating phosphorylations [154].

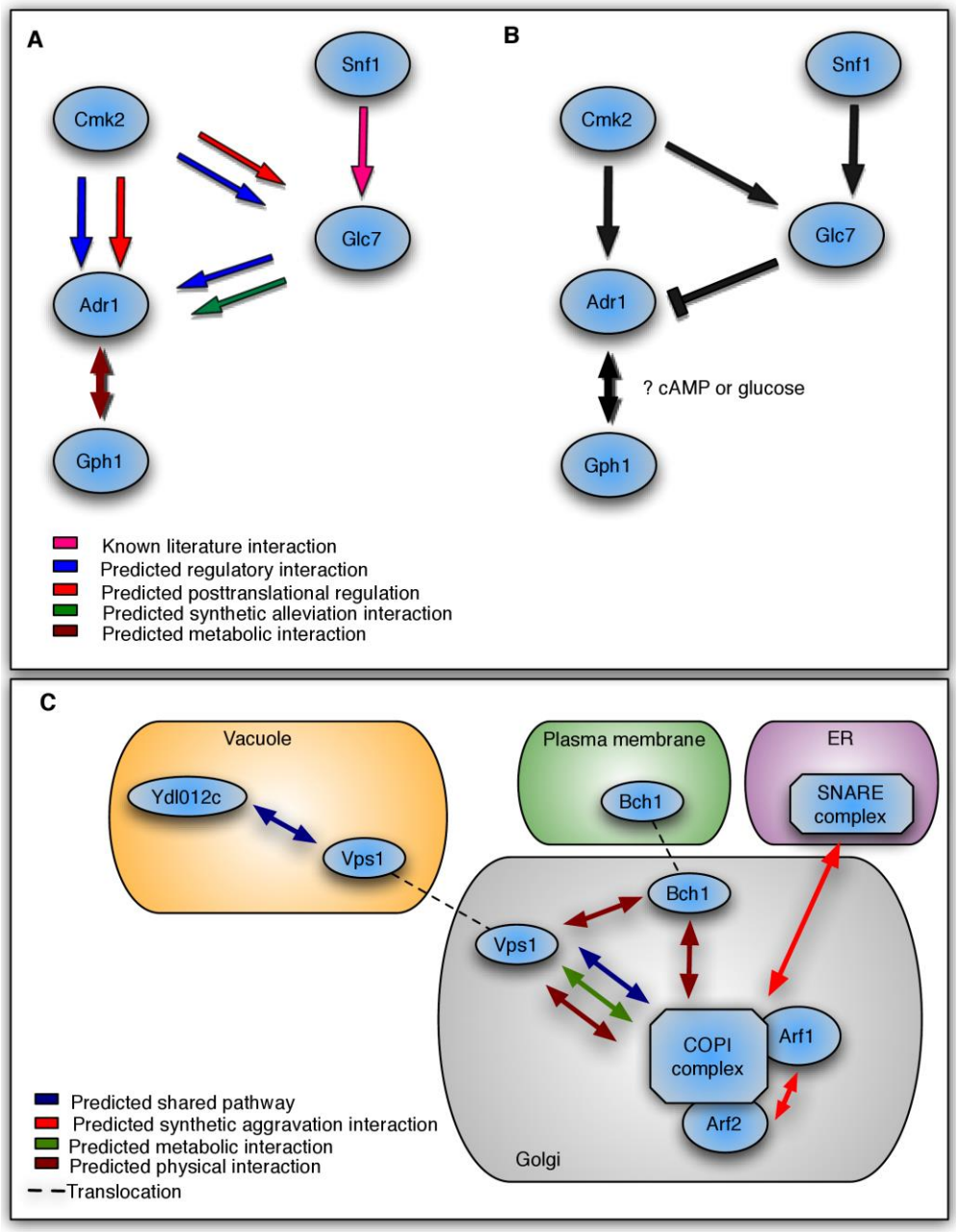

Figure 9. Examining the mechanisms of protein interactions within the yeast carbon metabolism and cellular transport pathways.

A) Predicted interactions of four specific types combined to assemble B) (arrows in black representing our final predicted pathway interactions) a pathway connecting the transcription factor Adr1 involved in carbon metabolism process to its regulatory input Snf1. This generates two concrete hypotheses suggesting, first, cross-talk between the calmodulin- and Snf1-dependent pathways via Cmk2 phosphorylating Glc7. Second, we also predict coordinated regulation between the glycogen breakdown and glucose utilization pathways through a

metabolic interaction between Adr1 and Gph1. C) Previously known and newly predicted interactions in yeast protein transport connecting the plasma membrane, vacuole, golgi and ER. We propose a regulatory competition between the Arf1 and Vsp1 GTPases for Bch1 functionality that is likely regulated by GTP availability, which itself is known to be regulated by protein sorting events in the cell. These predictions also hypothesize that YDL012c may be involved in regulating Vps1 activity.

The predicted yeast networks also hypothesized post-translational regulatory interactions between Cmk2 and both Adr1 and Gkc7 (Figure 9A). This three-protein network creates a feed-forward regulatory motif in which Cmk2 simultaneously activates Adr1 as well as its inhibitor Gkc7, creating a time-delayed inactivation of Adr1. These interactions are supported by a recently published paper [153] linking the calmodulin- and Snf1-dependent pathways to Adr1 regulation. Our predicted pathway takes these results a step further by identifying which of the three calmodulin-dependent kinases (Cmk2) is responsible [155]. Finally, a novel metabolic interaction was predicted between Adr1 and Gph1, the only high scoring interaction of its type for Adr1. Like Adr1, Gph1 is involved in glucose metabolism by glycogen breakdown, and both are regulated by the metabolites glucose and cAMP [156]. A metabolic interaction between Adr1 and Gph1, combined with the known regulation of these genes by glucose and cAMP, suggests that coordinated regulation is occurring between the glycogen breakdown and glucose utilization pathways and is transcriptionally controlled by Adr1.

3.2.4 An inferred pathway incorporating physical, genetic, and metabolic interactions spans cellular compartments in yeast protein transport

Protein sorting and trafficking is an essential function of eukaryotes and requires numerous multi-subunit complexes to ensure the proper localization and secretion of proteins (Figure 9C, [157]). At the early stages of this process, the two major transport pathways from the endoplasmic reticulum (ER) to the Golgi are governed by the SNARE and COPI complexes [157]. We predicted synthetic interactions between these pathways (e.g. synthetic aggravation for Arf1-Sec18 and synthetic alleviation for Sec27-Uso1) that are supported by known genetic interactions[158,159]; Arf1 and Arf2 are a representative example, as they are considered functionally redundant GTPases, and each COPI complex contains either Arf1 or Arf2 [160].

Later in the pathway, Bch1 is a member of the ChAP family of proteins, which direct cargo bound to COPI complexes in the Golgi to their destinations such as the plasma membrane [161]. We predict a physical interaction between Bch1 and the COPI complex that is well established in the literature but was not part of our gold standard. Likewise, Vps1 serves a similar function for vacuole targeting [162], and our predictions of its physical and shared pathway interactions with COPI are supported by the literature [161].

Novel hypotheses in Figure 9C include the predicted physical interaction between Bch1 and Vps1, suggesting competition between the Sec27-Arf1 and Vps1 complexes for the Bch1 sorting function (also supported by a metabolic interaction between Sec27-Arf1 and Vps1). Both Vps1 and Arf1 are GTPases that must hydrolyze GTP to perform their roles in protein sorting [160]. Thus, this predicted pathway hypothesizes a competition between the Arf1 GTPase and Vsp1 GTPase for Bch1 that is likely regulated by GTP availability. Similarly, the uncharacterized membrane-bound protein YDL012c is placed in the same pathway as Vps1, suggesting that the former may be involved in regulating Vps1 activity. By highlighting just a

few of our predicted interactions in the protein sorting pathway, we demonstrate the potential for generating hypotheses used to drive novel biological discoveries.

3.2.5 Experimental validation of predicted interactomes

To experimentally evaluate the accuracy of a subset of our predicted interactions in a directed manner, we focused on the DNA topological change and protein biosynthesis regulation processes in *S. cerevisiae* [163]. 20 synthetic lethality interactions predicted with high probability were experimentally tested using SGA technology [4,144], with the results summarized in Figure 10. 14 gene pairs (70%) were validated, 8 involved in DNA topological change and 6 in the regulation of protein biosynthesis. Several of the remaining 6 unconfirmed interactions may be synthetic lethal under different conditions. For example, GCS1 and SLT2 deletions both individually decreased resistance to ethanol stress [164], and similar conditions might elicit synthetic lethality. Based on a total of ~100,000 pairs estimated to have been synthetically lethal in yeast of a possible ~18 million (0.05%) [144], our predictions are a clear improvement over the baseline rate for novel discovery.

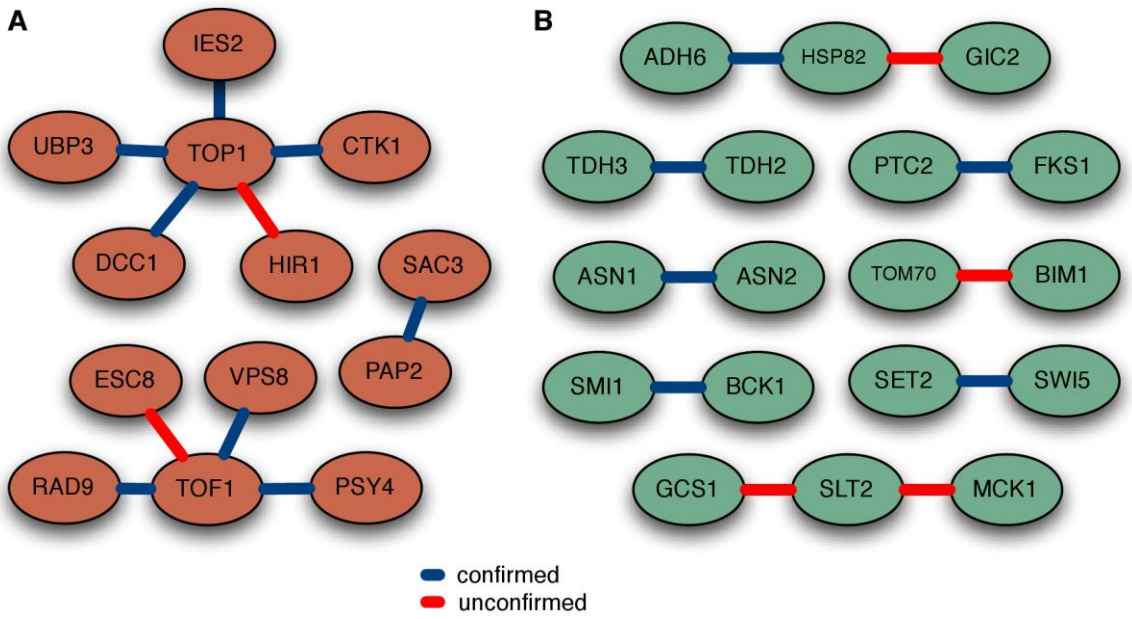

Figure 10. Experimental validation of predicted synthetic lethal interactions.

Experimentally tested synthetic lethal hypotheses in the yeast A) DNA topological change and B) regulation of protein biosynthesis processes. A total of 20 gene pairs from our predicted synthetic lethality networks were experimentally tested using the SGA platform. We confirmed 14 of these interactions (70%), 8 in DNA topological change and 6 in protein biosynthesis. Several of the remaining unconfirmed pairs (e.g. GCS1 and SLT2; see main text) show additional evidence of condition-specific synthetic lethality [4,144].

As an additional evaluation, we collected 24 recent publications containing a total of 173 experimentally confirmed post-translationally regulated protein pairs. None of these interactions was present in our training standard. Evaluating on this set, our Bayesian hierarchical system achieved an AUC of 0.802, demonstrating its ability to accurately predict novel, experimentally verifiable post-translational regulation interactions on a whole-genome scale. This accuracy is comparable to our initial cross-validation AUC of 0.778, indicating that our evaluation provides a reasonable estimate of the expected experimental verification rate for novel predictions.

3.2.6 Systems level view of cellular interactomes

This rich compendium of inferred interaction types provided an opportunity to analyze systems-level network features genome-wide at multiple levels of biomolecular activity. In particular, we examined the network structural characteristics that potentially help to define the functional roles of each interactome. Biological networks have been proposed to exhibit a scale free topology [165], implying a power-law degree distribution. Previous studies have detected such distributions based on partial networks and single interactomes [166]. Here (Figure 11A), we observe a scale-free degree distribution very robustly in all 30 interaction types. However, the high-degree hubs in each interactome do differ, reflecting the distinct functional activities carried out by each network type. To verify this, we analyzed the extent of the overlap of high-connectivity genes (in the top 5% of the degree distribution) between the networks for each pair of interactomes (Figure 11B; directed interactomes were divided into separate in- and out-degree comparisons). The major clusters show distinct functional similarity, correctly reflecting the similarities captured by our interaction ontology: transient and nontransient physical interactions each group together, synthetic interactions cluster, and so forth. Beyond confirming the structure of the ontology, this also captures relationships such as the sharp divide between regulatory in- and out-degree (the most regulated genes are not themselves high-level regulators with many targets) and a tendency for regulatory hubs to incur more synthetic interactions than expected.

Degree distribution captures a global description of each network, while analysis of small recurring subgraphs has been proposed to describe local network properties [167,168]. We investigated the enrichment of two types of subgraphs, network motifs and graphlets, in our

interactomes. First, network motifs are small directed subgraphs that have been found to recur in a growing number of organisms [169-171]. In our 12 directed interaction networks, the feed forward loop motif showed significant enrichment (relative to a random background) consistent with previous studies on the yeast transcription factor network [168]. Feed forward loops are known to accelerate or delay the response of a input signal [172], suggesting in this context a much wider usage of dynamic information processing than has been previously reported in regulatory networks[173-175].

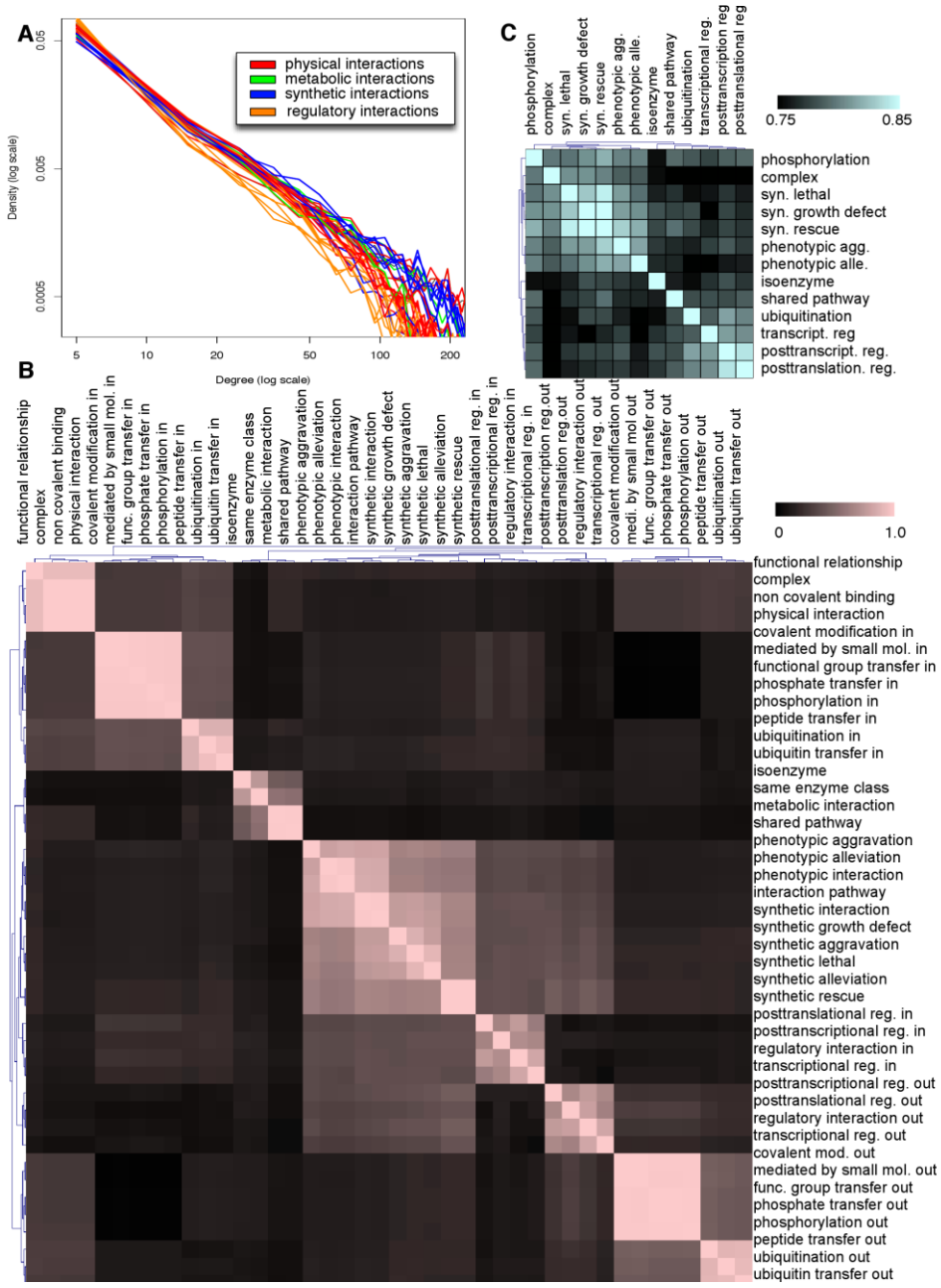

Figure 11. Systems-level analysis of inferred networks.

In all cases, continuously weighted networks were binarized by choosing an edge cutoff three standard deviations above mean, retaining ~1% of all edges. A) The degree distribution for all 30 of our predicted interactomes agrees strongly with a scale-free network topology. B) Conditional probabilities for a gene to appear in the top 5% of each pair of networks' degree distributions. Similarity indicates that a pair of networks share the same high-connectivity genes and thus represent functional activity carried out by similar sets of proteins. C) Graphlet degree distributions compared using the GDD metric between the 13 leaf interactomes in our interaction

ontology. Network pairs with greater similarity demonstrate related local network topologies, suggesting that comparable functional modules might be employed in the two interactomes (e.g. between phosphorylation and synthetic interactions or ubiquitination and post-translational regulation).

A second approach to exploring the local structure of biological networks is to examine graphlet degree distributions [167]. Graphlets are small non-isomorphic subgraphs, and a graphlet's degree for a given node is defined as the number copies of that graphlet to which it is incident. For example, the number of triangle motifs touching a particular node represents its 3-node graphlet degree. Compared to network motifs, for which enrichment can be difficult to detect due to the complexity of the baseline null distribution[176], graphlet analysis may have a higher sensitivity towards infrequent subgraphs. Thus, as a complementary analysis, we computed the graphlet degree distributions for all two to five node graphlets for the 13 specific leaf node interactomes in our interaction ontology (Figure 11C). We compared the graphlet degree distributions between these interactomes, demonstrating a clear divergence in the local network structure between subclasses of metabolic, regulatory and synthetic interactions. Unlike the comparison of high-degree genes, this also captures unexpected similarities between disparate interaction types: phosphorylation and ubiquitination, for example, are siblings in the interaction ontology and represent comparable mechanisms of post-translational modification. The former's local network topology is more similar to that of synthetic interactions, however, while ubiquitination is more strongly regulatory. This differentiating pattern between ubiquitination and phosphorylation provides a base for intriguing network hypotheses for further investigation. One potential explanation could be due to the differing mechanistic activities where ubiquitination is most often employed exclusively as a regulatory mechanism to degrade active

proteins, whereas phosphorylation serves both regulatory and dynamic information processing roles[177].

3.3 Discussion

The increasing abundance of genomic data has opened up countless new possibilities for systems-level biological perspectives, but its increasing complexity impedes the understanding of specific cellular circuitry at a mechanistic level. Here, we provide a method with which very large experimental data compendia can be integrated to predict 30 specific biomolecular interaction types at a genome-wide scale. By applying this to more than ~3,500 experimental conditions in yeast, we have evaluated these predictions at an average AUC of 0.79, validated 70% of experimentally tested synthetic lethal interactions, and proposed novel transcriptional, genetic, post-translational, and metabolic interactions in the yeast carbon metabolism and cellular transport pathways.

As described above, the investigation of specific *S. cerevisiae* biology in the processes of glucose utilization and protein trafficking demonstrates the use of these interactomes to reconstruct complete pathways. In many instances, experimental biologists are faced with the task of designing experiments to target a specific set of genes. By simultaneously hypothesizing all types of biomolecular interactions in which a group of gene products may be involved, this methodology can be used to select both the proteins to be assayed and the assays that may be most informative. Prior approaches inferring these interaction types in isolation mask this information and may even be inconsistent; how might a biologist interpret predictions that two

proteins physically interact, but that they are not part of the same pathway? Such inconsistencies are avoided by simultaneous ontology-based inference, allowing underlying experimental data to be integrated into a consistent description of a cellular system.

To our knowledge, there has been no other method that simultaneously enables researchers to leverage high-throughput data in an interaction-type-specific manner within an ensemble setting. Successful focused attempts to predict specific interaction types have shown comparable AUCs to our results [178,179], which could be incorporated into a framework like this as base classifiers during future work (instead of the SVMs utilized in this study). Recent "functional coupling" predictions [36] are also related, but fall short of pathway-level interaction predictions, mainly due to a lack of the crucial directional information needed to infer bimolecular pathways. These frameworks typically also do not resolve inconsistencies among predicted interaction type labels that can hinder pathway reconstruction and experimental follow up.

Ultimately, compendia of inferred interaction networks can be used to explicitly construct and understand distinct cellular pathways. By investigating and confirming different interaction types suggested by our system, investigators can stitch together both new pathways and new interconnections between existing ones. This process can be applied in any organism for which diverse genome-scale data is available - a situation that is only becoming more common. We believe that our work can leverage this diversity of experimental results that might otherwise be underutilized, helping to spur new functional discoveries in organisms beyond yeast. Finally, all of our predicted networks are made publicly available through an interactive tool at

http://function.princeton.edu/bioweaver for investigators to explore their own biological areas of interest.

3.4 Methods

We developed an integrated method for concurrently predicting multiple protein interaction types. This method integrates large and diverse collections of functional genomic data in the context of a biomolecular interaction ontology. Each gene interaction type in the ontology is first predicted using an SVM classifier by integrating ~3,500 experimental conditions from expression, colocalization, regulatory, and other yeast experimental data (withholding data types directly related to the interaction type being predicted; see below). These isolated interactomes are then reconciled using a hierarchical Bayesian framework to obtain the most probable set of consistent labels for each gene pair within the hierarchy of our interaction ontology. Using this system, we generated 30 *S. cerevisiae* interactomes, with which we validated several mechanistic interaction predictions in carbon metabolism, cellular transport, and 14 new synthetic lethal interactions in DNA topological change and protein biosynthesis.

3.4.1 Interaction ontology construction

We constructed an interaction ontology focused on categorizing gene pair relationships. This is similar in spirit to the Gene Ontology (GO) [163], which curates individual proteins' molecular functions, biological roles, and subcellular localizations. Our interaction ontology contains a

total of 124 terms and integrates information from existing interaction catalogs [180,181], the EBI [182], and SGD [183]. The ontology's three major branches are metabolic, interaction pathway, and physical interactions. Metabolic interactions describe protein pairs linked in metabolic pathways, such as isoenzymes or enzymes that catalyze adjacent reactions. Physical interactions include covalent or non-covalent binding, e.g. stable complexes or transient post-translational modifications. Pathway interactions include more conceptual relationships between genes in a pathway, such as regulation or synthetic interactions. We selected the 30 nodes in our interaction ontology with more than 70 annotations (as described below) to include in this evaluation, and the complete ontology with descriptions of each term is provided in Table 1.

3.4.2 Gold standard construction

There exists no comprehensive curated gold standard repository for all types of gene pair interactions. For the 30 interactomes evaluated here, we assembled a gold standard for each type from various sources. SGD interaction labels were used for all terms under the physical and pathway interaction terms [183]. Additional transcriptional regulation annotations were obtained from the high confidence set from [184]. Co-complex annotations were obtained from gene pairs in the GO Slim term *PROTEIN_COMPLEX* [39]. Pairs included in terms under metabolic interaction were obtained from reactions in the KEGG database [185]. For the topmost node, functional relationships, we used positive examples from the biological process branch of GO [122]. When possible, we further manually curated gene pairs to more specific terms based on literature examination. Manual curation was performed to annotate ubiquitination interactions based on SGD curated interaction publications and also to cross annotate experimentally

validated covalent modification branch examples to regulatory interaction branch terms. The directionality of the gold standards was derived directly from the inherent high throughput experiments (e.g. kinases to targets). All gene pairs annotated to a term were propagated such that they were included as positive interactions for all ancestor terms. This resulted in a total of 1,333,014 unique positive labels across 30 terms.

This process established positive interactions for each term in our interaction ontology. For supervised machine learning (such as our SVM-based method described below), negative examples are also required. As protein interactions are sparse, we randomly selected a number of negative gene pairs for each term's gold standard equal to the number of positive interactions [186]. Additionally, to assess the accuracy of our directed interaction predictions, we used negative gene pairs identical to the positive examples but with inverted directionality.

Evaluation was performed by randomly excluding ~%30 of the genes for each interaction type during training. That leads to a group of genes that are not in the training set and established a test set of interactions containing at least one gene from this exclusive gene set. The remaining pairs were used for SVM training and for parameter estimation in the Bayesian network. We used area under the receiver operator characteristic (ROC) curve (AUC) for evaluation.

3.4.3 Data sources and preprocessing

As training data for each interaction type, we used subsets of a data compendium consisting in total of microarray, colocalization, protein domains, transcription factor binding sites, and sequence similarity. For each interaction type to be predicted, experimental data closely related

to the output was excluded (e.g. TF binding sites for regulatory relationships). 78 yeast microarray datasets were included, comprising 3,516 conditions. Missing values in these datasets were imputed using KNNImpute [187] with k=10, and genes with more than 30% missing values were removed.

For machine learning, one feature was constructed per expression condition as follows. For directional gene pair interaction types such as phosphorylation, we evaluated various methods and found x_i-x_j to provide optimal performance, where x_i and x_j are the expression values of gene i and j in condition x. When predicting non-directional interaction types such as physical interaction, we used $|x_i$-$x_j|$, the absolute value of the subtracted expression values.

Colocalization data for 22 different cell compartments [188] and automatically determined protein family information from Pfam B [189] were both included as binary features (true if both genes in a pairs shared localization or a protein family). TRANSFAC data [190] was incorporated using the Euclidian distance between the two gene's binding site profiles across 211 transcription factors. Sequence similarity between the two genes in each pair's 1,000bp upstream and 1,000bp downstream was scored as the sequence alignment E-values from all-against-all BLAST outputs.

3.4.4 Algorithm

We developed an integrated method for predicting diverse protein interactions, based on a multi-label hierarchical classification formulation we have previously applied to gene function prediction in both yeast and mouse [45,138]. First, for each interaction type, we trained 10

separate SVM classifiers. We use bagging (bootstrap aggregation, [191]) to combine these and improve generalization, training each individual SVM classifier on a bootstrapped subsample of its interaction type's complete gold standard. We thus begin with a total of 300 SVM classifiers for our 30 interaction types in yeast, and each interaction type's group of 10 SVM outputs were averaged (bagged) to produce a non-hierarchically-resolved predicted interactome.

Next, a Bayesian network was constructed based on the structure of the interaction ontology. First, we modeled each interaction type's bagged SVM output i as a random event Y_i taking discrete values binned by five standard deviations above or five below the training set mean. Each SVM's predictions in isolation were treated as a noisy observation of a latent event X_i representing the true, binary interactions and non-interactions of each type i. Each Y_i was considered to be dependent only on its corresponding X_i, and each X_i was dependent only on its set of children $\{X_j, ..., X_k\}$ in the interaction ontology, resulting in the "decorated tree" Bayesian network structure seen in Figure 7 and in [138]. Given this structure, conditional probability table parameters for $P(Y_i|X_i)$ were learned using maximum likelihood from interaction type i's training data. Finally, parameters for $P(X_i|X_j, ..., X_k)$ were fixed to constrain the hierarchical semantics of the ontology. If a pair is annotated to any child in $\{X_j, ..., X_k\}$, it must also be of interaction type i, making $P(X_i=1|X_j=1) = ... = P(X_i=1|X_k=1) = 1$. The remaining parameters $P(X_i=1|X_j=0, ..., X_k=0)$ were inferred using maximum likelihood by counting the corresponding training labels. Finally, Laplace smoothing was used to improve parameter robustness.

3.4.5 System level network analysis

All 30 interactomes were converted into binary interaction networks by setting a threshold of 5 standard deviations above the mean edge probability, retaining ~1% of all edges. The degree of each gene was counted in this binarized network. The overlap between each pair of interactomes' high-connectivity genes was computed as the probability of a gene g being in the top 5% of interactome N_1's degree distribution($Q_i(N_j)$, defined as genes in the top i percent degree distribution of interactome N_j) given that it was in N_2's: P[g in $Q_{0.05}(N_1)|g$ in $Q_{0.05}(N_2)$]. For each of the 30 interactomes N_2, we generated a sorted gene list by edge degree; for directed interactomes, separate lists were generated for in- and out-degree. Next, we counted the number of shared genes in the top 5% of edge degree in the target interactome N_1. Finally, hierarchical clustering was used to generate clusters of shared high degree genes.

Network motif enrichment analysis was carried out using FANMOD [192]. Searches were conducted for 3-node motifs using a sampling method with probability parameters of 0.6, 0.5, 0.4 and compared to 500 random networks generated using an edge swapping process preserving each gene's degree. Computational complexity precluded analysis of 4-node motifs. Graphlet degree distributions were calculated using GraphCrunch [193]. For each interactome, 73 graphlet degree distributions were generated, each representing a unique distribution of 2-5 node graphlets. Comparison between graphlet distributions was performed using the *GDD agreement* metric, defined as the average normalized distance to provide robust comparisons [167,193].

3.4.6 Implementation

All software was implemented using the Sleipnir library [131], which interfaces with the SVM^{perf} software [194] for linear kernel SVM classifiers (the error parameter C was set to 20 for these experiments). Bayesian network inference used the Lauritzen algorithm [195] as implemented in the University of Pittsburgh SMILE library [196].

3.4.7 Experimental validation of synthetic lethal pairs

20 gene pairs predicted to synthetically interact [183] with high probability were selected from the DNA topological change and regulation of protein biosynthesis pathways in yeast (as defined by GO [163]). Synthetic Genetic Array (SGA) technology [4,144] was applied to these pairs by combining either non-essential gene deletion mutants or conditional alleles of essential genes in haploid yeast double mutants. The query mutant strain for each pair of genes (harboring SGA-specific reporters and markers) was crossed to the complementary single mutant strain. Mating to the non-essential gene deletion collection was followed by meiotic recombination and selection of haploid meiotic progeny, resulting in an output array of double mutants grown in rich medium. Fitness was assessed by comparing this double mutant colony size to the sizes of single mutant colonies, which were assessed for significance as described in [4,144]. A p-value threshold of 0.05 was used to determine the final confirmed synthetic lethal pairs.

4 Data integration for the inference of pathway level interactions in metazoan organisms

4.1 Introduction

The molecular activity in a cellular system is maintained by a complex interplay between genes, gene products, metabolites and the environment [197]. In particular, diverse types of mechanistic pairwise interactions, including physical binding in protein-protein complexes [198] and through modifications [135] and regulatory roles as activators or repressors [137], are combined to form intricate biomolecular pathways. Mapping out these cellular pathways at a whole-genome level is a crucial step for the advancement of human systems biology, aiding at every level from deciphering cellular function to understanding the molecular cause of many complex human diseases.

Functional genomic datasets such as gene expression, cellular localization, and DNA/protein binding assays each capture distinct aspects of the cellular activity across multiple cell types and perturbations, however turning these different instances or “views” of a complex system into an

understanding of pathways has proven to be challenging undertaking. Especially in metazoan mammalian organisms, tissue and cell type-specific expression underlie the development, function, and maintenance of diverse cell types [199,200]. Consequently, the biological and technical variation of the functional genomic data, particularly with respect to tissue and cell-lineage heterogeneity compounds the difficulty in translating the vast amount of genomic data into specific pathway-level hypotheses. Thus, it is of great need to develop and apply algorithmic and statistical approaches for inferring the individual types of biomolecular interactions, scalable to the whole-genome and robust to any biological dataset of diverse tissue contexts (i.e. experimental results drawn from differing tissues).

In this work, we developed an overall strategy for predicting and studying multiple types of biomolecular interactions for metazoan mammalian organisms, specifically applying this approach to the human data compendium. Our method consists of three operational steps. First, we catalog tissue and interaction-type specific gold standards (e.g. phosphorylation interactions among brain expressed genes) restricted to protein pairs expressed in total 77 tissues based on curated pathway databases and gene-to-tissue expression profiling. Second, we utilize a state-of-art classifier Support Vector Machine (SVM) [40] to predict each mechanistic interaction type by integrating ~1,600 human experimental datasets (each dataset consisting of multiple conditions) independently in the context of each tissue. Finally, we aggregate the tissue-context based predictions to obtain the most probable set of interaction labels for each gene pair across the set of mechanistic interaction types (in this study we predict transcriptional regulation, phosphorylation, co-complex and post-translational regulation). Our methodology uniquely allows us to harness the wealth of information in high-throughput genomic data collections by

simultaneously separating the heterogeneity originated from tissues while predicting for each individual interaction types.

To our knowledge, prediction of such genome-wide mechanistic networks in metazoans is an open problem. Many recent studies have begun to address the challenges by inferring physical interaction networks [13,143,201,202], synthetic interaction networks [5,144], Bayesian integration for functional associations networks [30,106,203-205] or predicting regulatory networks from specific primary datasets [206,207]. However, most previous efforts for predicting regulatory interactions have been focused on unicellular model organisms (e.g. e. coli and yeast) [16,208], while genome-wide integrated analysis in mammalian organisms have been often focused on cross-species integration for inferring functional couplings [36]. No prior integrative methods to our knowledge utilize one of the most significant sources of biological variation in human datasets: tissue context. Our work extends the methodological advancements achieved studying regulatory networks in unicellular model organisms and provides a platform for applying such methods to human data by addressing the challenge of tissue heterogeneity.

Ultimately, we envision that our global mechanistic networks can also be leveraged to increase the interpretability of each respective primary data types (ChIP-chip and -seq, proteomics on double knockouts/knockdowns, disease samples, etc) that captures condition-specific cellular states. As a proof of concept, we demonstrate the utility of our networks by overlaying our interaction networks to identify direct regulatory targets of transcription factors on all ChIP-seq experimental datasets generated by the ENCODE project [209]. In addition, we generated the first *in vivo* derived binding motifs for cancer-associated TANK-binding kinase 1 (TBK1) by overlaying our phosphorylation network onto a recent TBK1 knockdown phospho-proteomics

study [210]. Finally, we provide a web-based interface for exploring all our interaction networks. This allows investigators a unique resource to interactively survey and generates hypotheses from the diverse interaction types comprising the mammalian cellular interactome.

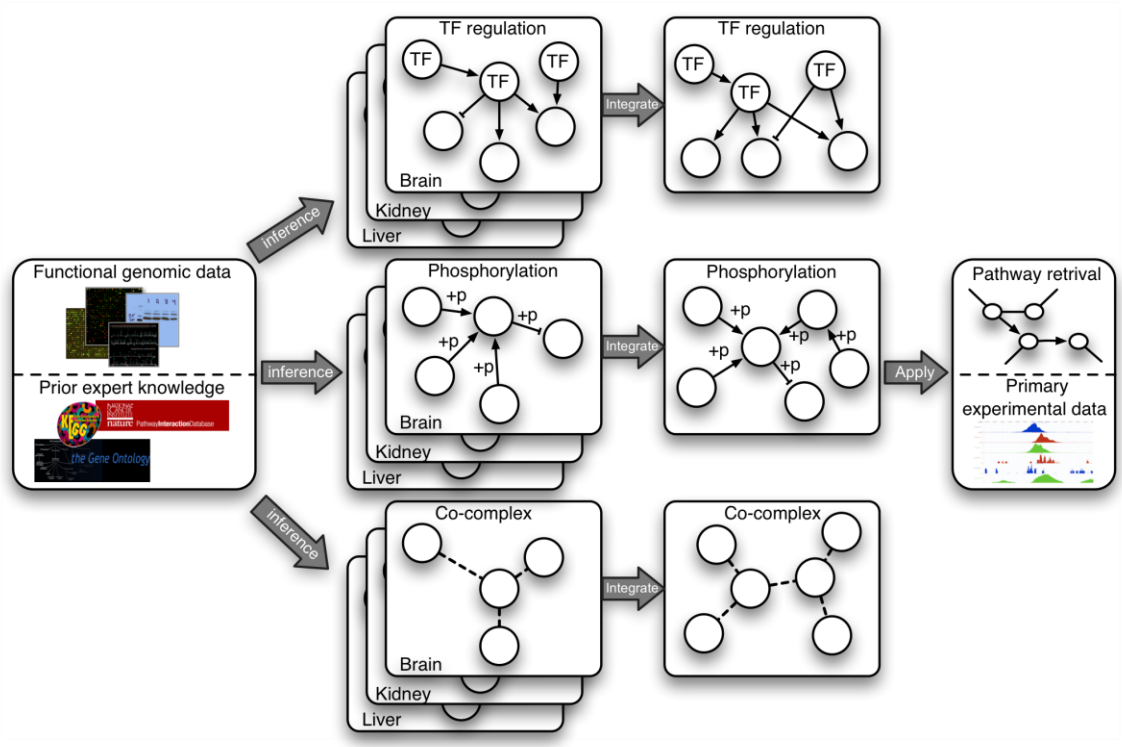

Figure 12. Schematic of our tissue-aware integrative pipeline for inferring metazoan biomolecular interactions.

We collect tissue and interaction-type specific gold standards, restricted to protein pairs that are both expressed in the tissue, based on curated databases and expression profiling of total 77 tissues. This focused gold standard allows us to separate the heterogeneity originated from tissues while predicting for each individual interaction types. Next, we infer each interaction type network by integrating the genomic data compendium independently in the context of each tissue (i.e. with tissue-specific learning examples), resulting in 77 intermediate networks per interaction-type. Following, we integrate the tissue-context based predictions to obtain the most probable set of labels for each gene pair across the biomolecular interaction types. Finally, our predicted networks can be used in multiple applications such as pathway retrieval or aiding the interpretation of condition-specific primary datasets.

4.2 Results

We demonstrate the importance of considering tissue context for predicting multiple pathway-level bio-molecular interaction types (transcriptional regulation, phosphorylation, co-complex and post-translational regulation). Specifically, we compare our tissue-aware learning approach with a simpler version that does not use information about tissue heterogeneity among human protein coding genes (i.e. tissue-unaware learning). In total, we apply our tissue-aware learning methodology integrating ~50,000 genome-scale experiments to generate whole-genome networks describing predicted gene-gene interactions across multiple pathway-level interaction types (prediction schematic shown in figure 12). We demonstrate the utility of our interaction networks in accurately retrieving pathway members across 447 expert-curated pathways. In addition, we show our interaction networks can be used to accurately identify false-positive regulatory targets in primary user data-sets generated by ChIP-Seq and Mass spectrometry based phospho-proteomics.

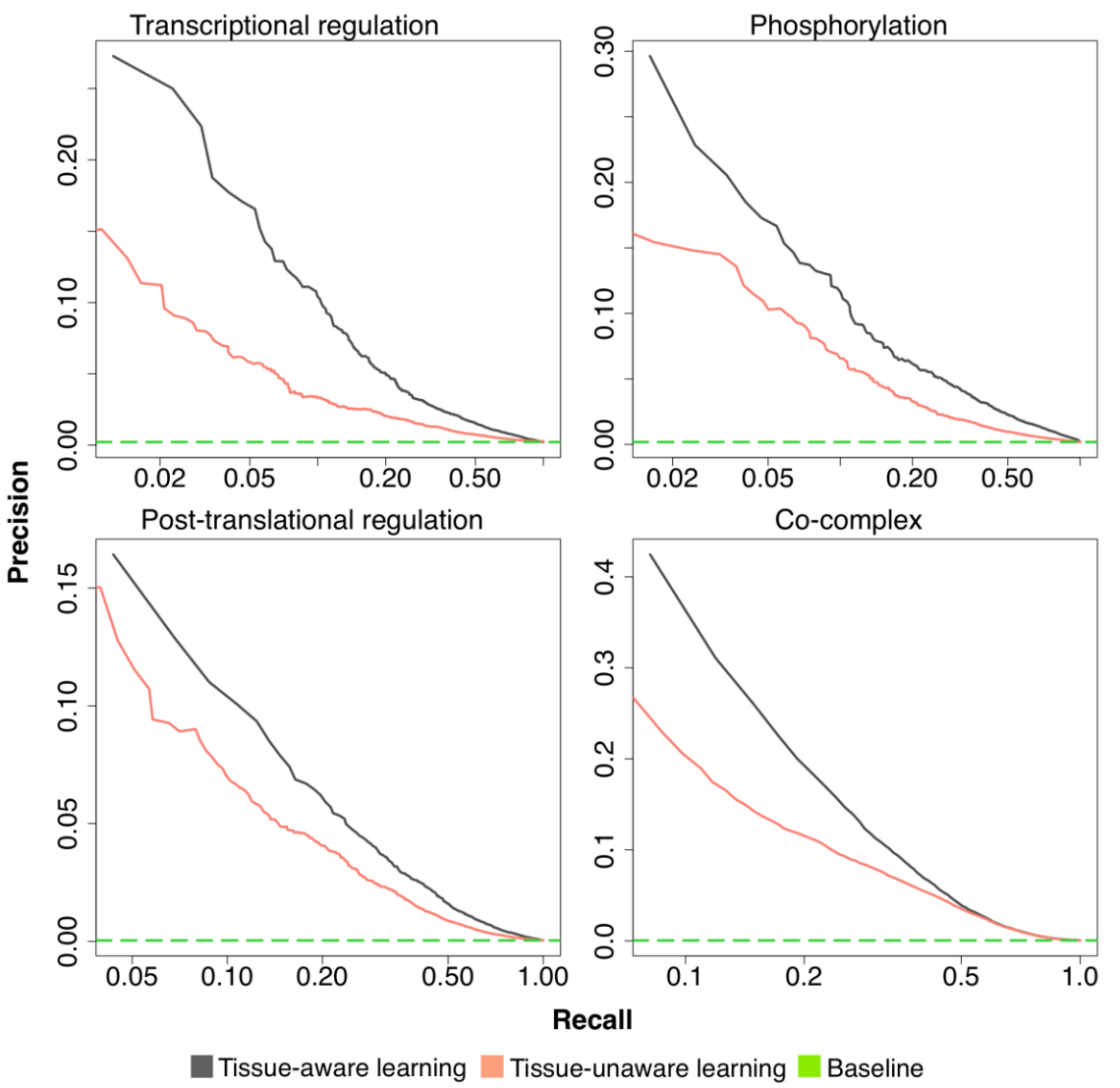

Figure 13. Tissue-aware learning allows accurate recovery of biomolecular interactions.

Biomolecular interactions are constructed for human biology based on tissue and interaction-type specific integration of diverse genomic datasets. The resulting interaction network consists of genes as nodes and connections between them representing the probability of two genes participating in the biomolecular interaction. Here, we show our threefold cross-validation prediction performance for predicting transcriptional regulation, phosphorylation, post-translational and co-complex interactions. Our tissue-aware learning methodology (black-line) significantly outperforms (p-value < 0.01 for all) compare to a simpler method that ignores tissue heterogeneity (labeled tissue-unaware learning and represented as salmon-line, x-axis in log-scale).

4.2.1 Tissue-aware learning improves human bio-molecular interaction predictions

To address the challenge of predicting pathway-level biomolecular interactions in metazoans, we ask the question if incorporating gene-level tissue contextual information can improve network prediction. Specifically, our tissue-aware learning method segments the training process into 77 diverse human tissue contexts allowing us to systematically separate the heterogeneity originated from tissue and interaction types in high-throughput genomic data. To measure the benefits of incorporating gene-level tissue context, we conducted a three-fold cross-validation experiment on both our tissue-aware learning method and a simpler non-tissue-aware learning method for predicting four interaction types (i.e. transcriptional regulation, phosphorylation, co-complex and post-translational regulation). For each of interaction type, we conduct a strict gene-wise holdout evaluation where at each fold the evaluation gold standard pairs consist of no genes observed during the training stage. In addition, identical human data compendium was used for tissue-aware and non-tissue-aware learning predictions.

For all four pathway-level interaction types, there was a substantial performance gain when using gene-tissue contextual information (figure 13). In total, there was an average of 71% increase in area under the precision-recall curve with transcriptional regulation showing the largest boost in performance >100%. In addition, our tissue-aware learning method outperforms random bagging [191] (random assignment of genes to tissue with matching number of contexts), except in co-complex that showed comparable performance gains mainly due to the significant portion of ubiquitously expressed gene pairs represented in the gold standard ~84%. In combination, these evaluations suggest that tissue-context prediction can significantly improve the accuracy when

predicting biomolecular interaction, especially when the interaction gold standard is tissue-heterogeneous.

4.2.2 Accurate retrieval of cellular pathway components

The concurrent inference of human biomolecular networks for multiple interaction types allows the generation of specific pathway level hypothesis. For example, often biologists are left with a set of genes that are believed to be functionally associated in a biological process resulting from a high-throughput assay (e.g. differential expression analysis of multi-condition RNA-Seq experiment). However, understanding the mechanistic connection between a set of functionally related genes have been challenging and often extremely time consuming due to the need of multiple laborious experiments. Thus, it would be of great value if direct mechanistic predictions inferred from the existing genomic data compendium can be made on any set of genes of interest to an investigator, allowing a systematic prioritization of hypothesis to be experimentally validated.

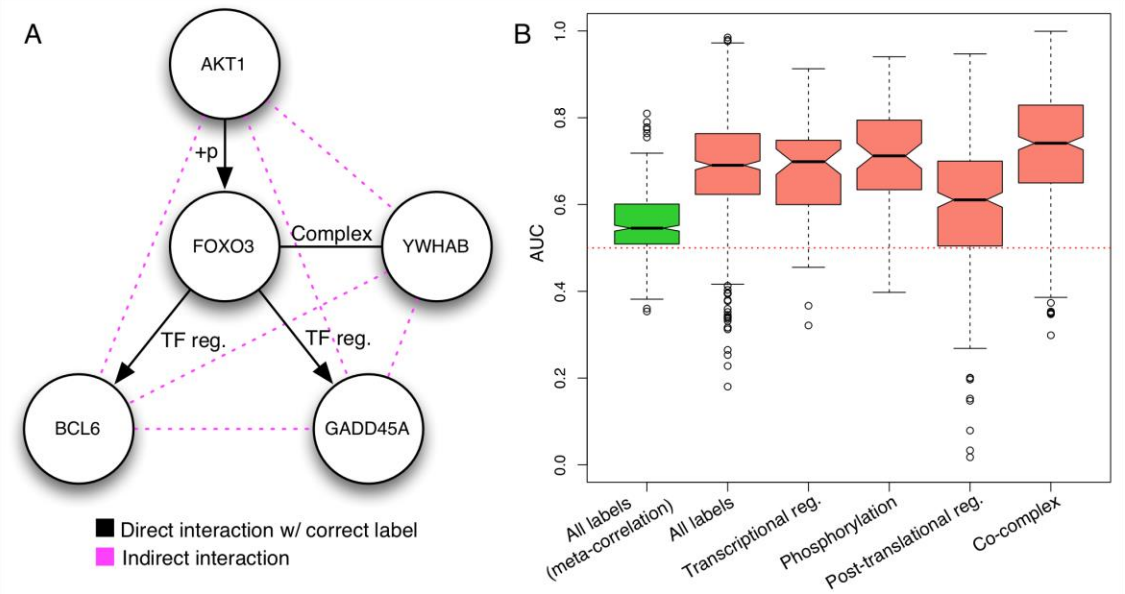

Figure 14. Our biomolecular interaction networks allow accurate human pathway component retrieval.

Pathway interaction networks provide investigators the means to generate pathway level hypotheses. In panel A, we detail a combination of transcriptional, post-translational, and physical interactions that we accurately prioritize surrounding the human tumor suppressor FOXO3 above indirect interactions and incorrect interaction-type labels. In panel B, we expand our pathway retrieval evaluation analysis to total 447 human curated pathways. Each dot in the box plot represents our accuracy of ranking the direct pathway interactions (with correct interaction label) above all incorrect interactions (i.e. direct interaction but with wrong labels or indirect interactions) between the constitute genes in a pathway. Overall, our networks show accurate performance in retrieval of direct pathway interactions when evaluated for both all interactions types combined ("all labels") and also for each individual interaction type. In addition, our performance is significantly better compared to a mega-clustering approach of calculating the Pearson correlation over all expression data concatenated together ("all labels (meta-correlation)").

To address the challenge of pathway recovery from functionally related genes, we test our ability of prioritizing the direct mechanistic interactions with known curated human pathways. For example, human FOXO3 is an important transcription factor involved in cell cycle regulation and oxidative stress response along with multiple tumorigenesis [211]. FOXO3 is known to be

functionally associated to human kinase Akt1 along with important regulatory genes such as BCL6, GADD45a and YWHAB [212-214]. If a researcher were to investigate the biomolecular interactions among these four clinically important genes FOXO3, Akt1, BCL6, GADD45a and YWHAB, our system can accurately prioritize the direct gene pairs with the confirmed mechanistic interaction type (figure 14A). In addition, such overlay provides a mechanistic hypothesis of the information flow among this set of genes. Starting with kinase AKT1 activating the transcriptional complex FOXO3-YWHAB through phosphorylation, next, the activated FOXO3-YWHAB complex to regulate the expression of targets BCL6 and GADD45A, connecting Akt1 to many downstream cellular processes such as DNA damage and apoptosis.

Next, to systematically evaluate for a broader set of pathways, we assessed our ability to recapitulate the mechanistic gene interactions for 447 expert curated human pathways by Pathway Interaction database [124]. For each pathway, we ask how accurate can our network predictions prioritize the direct gene pairs with the correct interaction type label compared to all indirect interactions and direct pairs but with incorrect interaction type labels. Our evaluation results are uniformly well above random (0.5), shown in figure 14B, with a median AUC performance of 0.69 across all pathways with minimal variations in performance across different interaction types. Specifically, we observe co-complex interactions showing the greatest accuracy at median AUC of 0.74 and post-translational regulation performing at a median AUC of 0.61. This indicates that our network interactions can accurately recover not only pathway structure from random gene pairs, but also can be used to resolve the direct mechanistic interactions among the more challenging closely related functional gene sets.

4.2.3 Interaction networks help interpret primary experimental datasets

In addition to pathway recovery, our predicted interactomes can be used to uncover high-accuracy regulatory targets in individual investigator generated high-throughput datasets.

Recovery of transcriptional regulatory targets from ChIP-Seq datasets

Chromatin immunopercipitation (ChIP) followed by high-throughput sequencing (ChIP-Seq) has been a prevalent experimental assay when applied to transcription factors to measure potential regulatory binding regions across the genome in *in vivo* settings [12]. Recently, ChIP-Seq data has been used to identify potential regulatory targets for an immuno-precipitating transcription factor (TF), mainly by identifying genes that have a TF ChIP-Seq binding profile proximally to its transcription start site (TSS) [215]. However, like many biological assays, ChIP-Seq is known to have a high rate of nonspecific cross-linking due to the usage of formaldehyde (compared to precise U.V. cross-linking that is specific to <1 angstrom proximity) that can lead to many false-positive regulatory candidate targets [9,216]. In addition, even when given a true biding incident of a TF, often there can be multiple TSS windows proximal to a binding locus thus obscuring the process of identifying the true regulatory target gene.

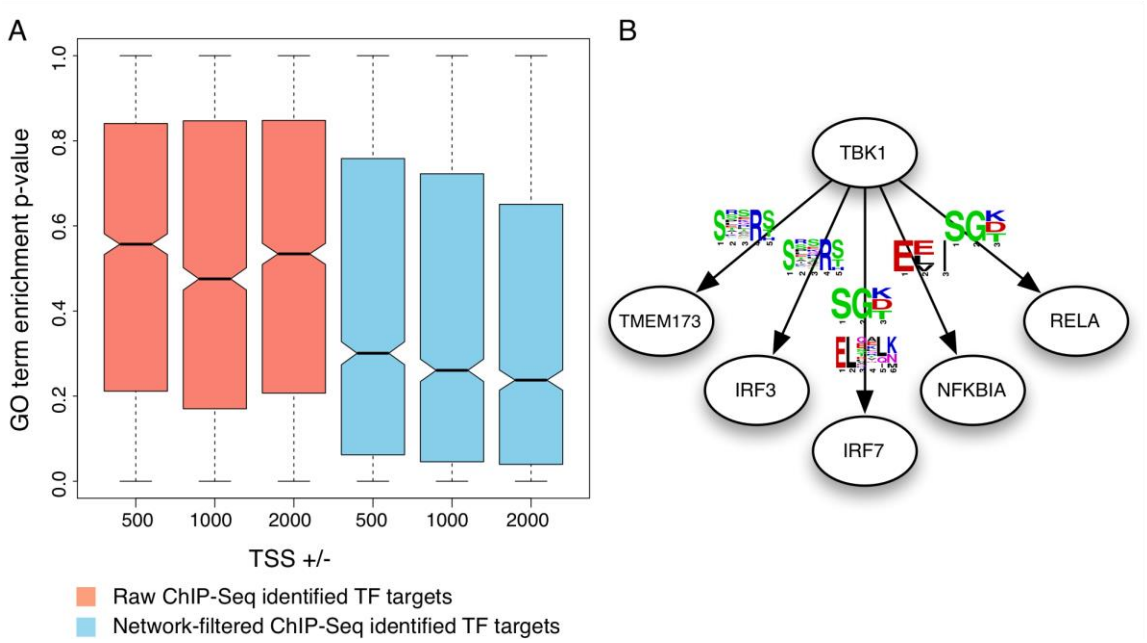

Figure 15. Improving the interpretability of primary experimental data.

Our biomolecular interaction networks can be used to help investigators increase the interpretability of condition-specific primary experimental data. In panel A, ChIP-seq datasets generated for the ENCODE project (109 TFs) were processed to identify transcription factor and potential regulatory target genes (identified by the existence of a TF ChIP-seq peak within a window surrounding the TSS of the putative target gene). We observe a significant increase in enrichment of TF associated GO terms among its putative regulator targets, when removing TF-targets pairs that have low predicted probability in our TF regulatory network (skyblue). Next, overlaying multiple data sources through our integrated phosphorylation network identifies unlikely direct targets of TBK1, which allows subsequent motif analysis to discover potential kinase binding or recognition sites. In panel B, represents our predicted phospho-binding motifs that are present in known TBK1 substrates not identified in the phospho-proteomics study. Such results support the biological relevance of these motifs and the accuracy of our phosphorylation network that enabled the analysis.

With our predicted transcriptional regulatory interaction network constructed from over 50,000 genome-scale experiments, we could discriminate false positive regulatory targets identified by a ChIP-Seq experiment either by the result of a non-specific binding or multiple TSS windows proximal to a binding profile. Thus, allowing the researchers to capture a more accurate binding landscape at the given experimental condition or tissue of the ChIP-Seq experiment. To test this

hypothesis, we overlay our transcriptional regulatory network to identify false-positive regulatory targets of transcription factors identified by ChIP-Seq experimental datasets generated for the ENCODE project (total 109 TF, 609 ChIP-Seq experiments) [209]. Specifically, for each 106 TF we identify potential regulatory target genes that have binding-profile peaks located within a window surrounding the gene's TSS. Next, we filter ChIP-Seq identified regulatory target genes that have a predicted probability of less than < 0.5 in our transcriptional regulatory network. In other words, genes with little experimental evidence in the human genomic data compendium of being transcriptionally regulated by the TF are flagged as false-positives.

To evaluate the accuracy of our ability to identify false-positive regulatory targets from ChIP-Seq experiments, we test if the regulatory target genes identified for each TF are enriched with genes known to be involved in the TF's functionally associated biological processes (limited to only experimentally annotations from Gene Ontology). We hypothesized that if false-positive TF-target gene pairs are accurately filtered the enrichment of biological processes with the TF annotation should improve. Indeed, the evaluation results shown in figure 15A demonstrate a significant increase in enrichment when filtering based on our transcriptional regulatory network compared to the original regulatory targets identified only from ChIP-Seq data.

In addition, the improvement in enrichment is consistent across varying windows (+/- 500, 1,000 and 2,000 bps) surrounding TSS for identifying regulatory target genes. Interesting, TSS window of +/- 1,000 bps have been used heuristically in recent studies [217], such window size showed the best performance in our evaluations due to a window of +/- 2,000 bps being too promiscuous (i.e. larger increase in false-positive regulatory target genes versus true-positives). However, our methodology permits a much larger window of +/- 2,000 bps while improving the overall

enrichment. Consequently, this allows investigators to identify larger number of functional regulatory profiles from the same ChIP-Seq dataset.

Phosphorylation network identifies TBK1 targets from phosphoproteomics data

With the recent advancement of stable isotope labeling techniques (e.g. SILAC), mass spectrometry has been the method of choice for high-throughput proteomics studies for many investigators. Especially, post-translational modifications (PTMs) such as phosphorylated proteins can be identified from mass shifts in the fragmented peptide ions scanned from the MS/MS readouts. In addition, coupling RNAi technology with quantitative Mass spectrometry has allowed the monitoring of the global alterations of the knock-down or knock-out phenotype of a specific gene at the proteome level (e.g. RNAi targeting a kinase) [218].

Although such global readout of the cellular state provides valuable information, RNAi/MS studies cannot distinguish the direct regulatory effect of the knock-down gene compared to the indirect effect. For example, the collection of differentially phosphorylated proteins after knocking-down a protein kinase will be a mix of direct substrates of the knocked-down kinase and also substrates of other de-activated kinases. This is because often signaling pathways consists of multiple kinase cascades [219], thus any knock-down of a kinase can lead to many de-activated kinases downstream in the pathway. Subsequently, the mixture of direct and indirect regulatory target genes complicates the generation of hypothesis generation and any follow-up analysis.

Thus, the investigator would benefit greatly if genes resulting from the differential analysis (e.g. differential phosphorylation) can be separated into direct regulated genes and indirect genes,

probabilistically classified by summarizing and overlaying the human data compendium. Specifically, our phosphorylation interaction network can provide a valuable resource for identifying direct regulated genes from an investigator's phospho-proteomics study. To test the applicability of our proposed approach, we applied our methodology to a recent phospho-proteomics study [210]. In this study, loss of phosphorylated proteins was measured using mass spectrometry following the RNAi mediated knock-down of TANK-binding kinase 1 (TBK1). TBK1 is an important kinase involved in innate immune response and implicated in multiple human cancers including lung cancer [220]. Therefore, the identification of regulatory targets of TBK1 and potential binding motifs can provide a great resource for future therapeutic studies.

In the published study, the researchers were not able to report any potential binding motifs of TBK1, most likely due to the mixture of direct and indirect targets among the differentially phosphorylated genes. This is especially unfortunate because no known *in vivo* binding motif has been identified for kinase TBK1. In fact, when we ran the state-of-art motif discovery tool FIRE [221] on the 2,150 differentially phosphorylated protein sequences, we retrieved many known binding motifs (total 4 motifs in the top 10 significant motifs) of other kinases (i.e. not the knock-down kinase TBK1). Interestingly, many of the kinases that bind to the identified motifs were among the differentially phosphorylated genes. This is consistent with the expectation that many of the substrates of these kinases contributing to the collection of differentially phosphorylated genes.

Next, we hypothesized that differentially phosphorylated genes (total 2,150) that have a low probability of being regulated by TBK1 in our phosphorylation network should be enriched with indirect regulatory targets of TBK1 (i.e. substrates of other kinases). In other words, by

overlaying our phosphorylation network, we could filter differentially phosphorylated genes that have little experimental evidence of being regulated by TBK1 in the human data compendium, and improve subsequent downstream analyses. Thus, we repeated the motif discovery analysis as conducted previously, however removing differentially phosphorylated genes with a TBK1 association probability of <0.5 in our phosphorylation network, resulting in 258 genomic data-supported substrate protein sequences. As predicted, we no longer retrieved any significant binding motifs of other kinases, compared to our previous analysis that resulted in multiple known motifs of TBK1 downstream kinase targets.

Furthermore, due to the lack of proinflammatory stimuli in the experiment [210], seven known phosphorylation substrates of TBK1 were not identified to be differentially phosphorylated in this study. Interestingly, shown in figure 15B, 4 motifs among the top 10 significant motifs (only identified through our integrated approach, phosphorylation network + FIRE) were present in these known TBK1 substrates that were not included in the motif discovery analysis (such occurrence happening by random chance, p-value < 0.05). Thus, supporting the possibility of these motifs being the first *in vivo* derived motifs identified to be biologically relevant for TBK1 recognizing its phosphorylation substrates.

4.3 Discussion

Genomic approaches have provided us with great opportunities in unearthing the complexity of human biology and diseases. While increasing amount of human datasets measuring the molecular changes at the expression, epigenomic and proteomic level has provided us with an

invaluable public recourse, the efficient integration and identification of regulatory pathways and processes has been challenging. In this work, integrating ~1,600 human genomic-scale datasets, we provide the means of studying human biomolecular pathway-level interactions at the whole-genome scale. For each pair of genes in the human genome, thousands of experimental data points measuring the behavior of these genes were probabilistically summarized to infer both the presence of a functional association and also the most likely biomolecular interaction type. By applying our biomolecular interaction networks to primary investigator's datasets, we were able to improve the accuracy of identifying transcription factor regulatory targets from ChIP-Seq data compared to traditional methods and also identify novel kinase binding motifs from phospho-proteomics data.

In addition, this study demonstrates that directly incorporating tissue contextual information in the data integration and inference process of biomoclecular interactions for metazoan mammalian organisms can significantly improve the prediction accuracy. Although, we have implemented our system utilizing a maximum-margin hyperplane based SVM algorithm, we anticipate the overall approach of directly exploiting tissue and cell-lineage heterogeneity in human datasets can be readily incorporated into many future and existing methods. We also anticipate that the continued effort of curating pathway-level interactions and the wide use of next-generation sequencing on a variety of tissue-samples will allow us to extend our approach to other metazoan model organisms. Finally, we now have all our predicted whole-genome networks publically available at an interactive web-portal for researchers to conduct exploratory analysis for future hypothesis generation.

4.4 Method

4.4.1 Tissue-aware integration

The final output to our new prediction pipeline consists of predicted probabilities of gene pair associations for multiple pathway level bimolecular interaction types (e.g. transcriptional regulation, phosphorylation). Each gene interaction type network is derived from per-tissue based SVM classifiers (total 77 tissues) that capture the tissue-specific reliability variation while integrating across ~50,000 genome-scale experiments. Details of construction of gold standard interaction pairs, tissue context, SVM classifier and input datasets are described in the following.

Gold standard construction

Unfortunately, there exists no comprehensive curated gold standard repository for all human pathway level interactions. For each interaction type evaluated here, we assembled a gold standard from various sources that have collected experimentally validated interactions. Curated transcriptional regulation gene pairs from KEGG [185] and TFactS [222] were collected. Phosphorylation interactions curated from Human Protein Reference Database [223] were included and Co-complex annotations were obtained from manually curated human protein complex database MIPS CORUM [17]. In addition, Nature Pathway Interaction Database [124] was used to collect expert curated transcriptional regulation, phosphorylation and other types of post-translational regulation gene pairs. This resulted in 51,525 unique experimentally validated positive interaction labels across four interaction types.

For supervised learning negative gene interactions are required. To obtain high-confidence set of negative examples, we repeated the procedure from our previous study [106] that collected gene pairs not co-annotated to any terms in a set of 433 expert selected Gene Ontology [39] biological processes. Thus this formed a set of global unrelated gene pairs. To obtain interaction type specific negative examples, we filtered the set of global unrelated gene pairs by gene property. Specifically, transcriptional regulation negative examples were the subset of unrelated gene pairs that contained at least one human transcription factor (total 1,321 TFs from annotation study [224]). Identically, phosphorylation were the subset that contained at least one of 514 human kinases [225] and post-transcriptional regulation the subset that contained at least one of 1,881 protein modifying enzymes [39].

Data sources and preprocessing

We collected total 1,564 mRNA human expression datasets from NCBI Gene Expression Omnibus (GEO) [226]. Expression data was normalized according to the procedure described in [42] and final features were generated using the Fisher's z-transformed Pearson correlation for each gene-pair as in our previous study [106].

For non-expression data, data types that closely related to the output interaction type were excluded to avoid any circularity (e.g. no physical interaction data was used in predicting co-complex or phosphorylation). To measure shared TF binding site profiles, motifs were obtained from JASPAR [112] and DNase I profiling [227]. For each motif, we searched for possible transcription factor binding sites by scanning each TF profile in 1 kb upstream sequence of all protein coding human genes using FIMO [113]. Motif matches were treated as a binary score (present if p-value<.001 and not-present otherwise) and the final gene pair score was obtained by

calculating the pearson correlation between the two genes' binary score vectors. Shared miRNA motif profile were obtained from MSigDB mir database [27] and EBI MicroCosm database [228] and was converted and scored as done with TF motifs. In addition, CISBP-TF database an extension of CISBP-RNA [229] has binding preference motifs for 568 human transcription factors. CISBP-TF was used to create a binary feature by connecting each TF and gene that contain a significant match to the binding motif in its upstream sequence using FIMO as described above. Human kinases phosphomotifs were collected from ELM [230] and PhosphoSite [231] and a binary feature was created by scanning across all human protein sequences with FIMO. Protein domain-domain and motif interactions were obtained from PrePPI [232] and DOMINE [233], converted into gene pairwise binary scores based on the existence of a protein domain/motif in each of the two genes. Post-translational modifications (PTMs) on a wide varied of proteins have been cataloged by UniProt [234] and PTMcode [235]. We created a binary feature that captures the potential of a protein to be post-translationally modified by the assigning a one to any protein that have been observed of a modification and zero for no observed modification (e.g. 7,583 human genes have been observed to have been phosphorylated). Also, we created a feature that calculates the correlation of observed PTMs between any given protein sequence pair catalogued by PTMcode [235]. To capture the cellular component profile similarity between genes, we took the Pearson correlation of Gene Ontology [39] cellular component annotation profile for terms that had maximum 100 genes annotations between all gene pairs. Finally, chemical and genetic perturbation studies curated by the MSigDB [27] were summarized into gene pairwise similarity phenotype profile scores as described in [42].

Human tissue context construction

In order to capture a wide variety of tissues in our study, we cataloged genes that are probabilistically identified to be expressed across 77 diverse set of human tissues utilizing the Gene Expression Barcode methodology [236,237]. Specifically, tissue terms were selected to create a slim cut through the BRENDA Tissue Ontology (BTO) [238]. Next, text-mining of sample descriptions and other textual information available in GEO [226] was utilized to annotate expression samples to BRENDA Tissue Ontology terms. Next the Barcode methodology was applied to each expression sample with a tissue BTO term annotation (total 14,092 samples) and genes that had an average Barcode probability above 0.7 across tissue annotated expression samples were flagged as transcriptionally active in the tissue (resulting in transcriptomes of average 7,543 genes)

Tissue-aware data integration

The goal of our integration is to harness the information from the genomic data compendium to predict accurate pathway level interactions. Specifically, the integration is designed to model and exploit the tissue-specific reliability variation across genomic datasets for robust integration in metazoan interactome prediction. For each interaction type, one start-of-art Support Vector Machine (SVM) [40] classifier was trained per tissue context. The training gold standard for each interaction type *i* and tissue *t* was define as the following:

$$GS_{i,t} = \left\{ gs^{i}_{g_n,g_m} \middle| g_n, g_m \in Tissue_t \wedge g_n, g_m \notin Ubiq \right\}$$

where *gs* is an interaction example for interaction type *i*, and gene *n*, *m*. Tissue contexts *t* are all genes identified by our Barcode analysis to be transcribed in tissue *t* and *Ubiq*(ubiquitous)

identifies all genes that are transcribed across all tissues. Thus, a gene pair was considered a tissue-specific interaction example if both genes were expressed in the tissue, while ubiquitous gene interactions were treated separately as an independent context to accurately capture tissue-specific variation. Predicting for co-complex, we were unable to separate out ubiquitous gene pair examples due to the high percentage of such pairs ~84% (transcriptional regulation is ~35%) in the gold standard.

For each training gene interaction example a feature vector was constructed from a total of 1,590 datasets as listed above. Continuous expression features were binned into 0.2 z-score intervals and missing values were set to 0. The set of feature vectors for positive and negative training examples were used to train a linear Support Vector Machine (SVM) according to the following formulation:

$$\min_{w,\xi_i \geq 0} \frac{1}{2} w^T w + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

$$\forall i : y_i \left(w^T x_i \right) \geq 1 - \xi_i$$

where n is the training example gene interactions, w is the weight vector for each dataset, y_i is the training label of interaction example i and x_i is the data vector for all the features for gene pair i. All classifiers were trained using the identical gene-wise 3-fold cross validation split and tissue-contexts with fewer than 30 gene pair cross-splits were dropped. Finally, we merged the tissue-context based intermediate-predictions to obtain the most probable set of labels for each gene pair by assigning the mean predicted score of the top quartile of prediction values across tissue-contexts for each interaction type (performed best compared to other aggregation methods

based on our cross-validation results). In addition, one global tissue-unaware classifier was trained for each interaction type using the entire gold standard (i.e. non-tissue segmented).

4.4.2 Transcriptional regulation network applied to ChIP-seq data

690 ChIP-Seq [12] datasets generated by the ENCODE project was used to identify potential transcription factor regulatory targets for total 109 TFs. All ChIP-Seq data was handled as "Uniform Peaks" identified based on the ENCODE analysis and normalization pipeline [209]. Potential regulatory targets for each transcription factor was determined if the TF's ChIP-Seq peak overlapped within a window surrounding the gene's transcription start site (TSS). Windows of +/- 500, 1,000 and 2,000 bps were used to identify potential regulatory targets. Next, GO biological process terms for each TF that have been experimentally annotated and had gene annotations of total 5~200 were identified. For each TF, its ChIP-Seq based regulatory targets were tested for enrichment of GO biological process term gene list using a hypergeometric test. Next our transcriptional regulation network probability scores, predicted from our tissue-aware learning method, were overlaid on the ChIP-Seq based identified targets. ChIP-Seq identified targets that had a probability of being transcriptionally regulated less than 0.5 were filtered and removed. Enrichment analysis of GO biological process terms were repeated for each TF as performed before on the network-based high-scoring ChIP-Seq targets.

4.4.3 Phosphorylation network applied to phospho-proteomics data

In addition to ChIP-Seq data, we demonstrated the utility of our networks to identify novel phospho-binding motifs from mass-spectrometry proteomics data. Altered phosphoproteins were measured using stable isotope labeling mass spectrometry (SILAC [239]) following TANK-binding kinase 1 (TBK1) RNAi-mediated knockdown experiment in a recent publication [210]. In this study, total of 1,154 genes were identified for a loss of phosphopeptide (PEP score < 0.5 and Mass error < 5 ppm). Protein motif discovery tool FIRE [221,240] was applied to these differentially phosphorylated protein sequences to find enriched motifs that could potentially be binding targets of TBK1. Next, similar to ChIP-Seq data, we filtered and removed differentially phosphorylated proteins that had a probability of less than 0.5 association score to TBK1 in our predicted phosphorylation network. FIRE was applied on this filtered set of differentially phosphorylated proteins for discovery of enriched motifs.

4.4.4 Implementation

All software was implemented using the open-source Sleipnir library [131], which interfaces with SVM^{perf} package [132] for linear kernel SVM classifiers (error parameter C was set to 250 and error-rate loss function was used). All network predictions and evaluations were conducted for the 17,939 genes that were available on the Affymetrix U133A and U133 Plus 2.0 platforms.

5 Conclusion and future work

In this thesis, I have focused on developing scalable and robust computational methods that can summarize the large compendium of heterogeneous genomic datasets into a variety of hypotheses to direct biological discoveries. In chapter 2, I developed a novel statistical method for transferring new experimental discoveries (i.e. gene functional annotations) back to human and other model organisms between genes that share functional behavior derived from the comprehensive compendium of genomic data. In chapter 3, I developed a machine learning methodology for simultaneously predicting diverse types of biomolecular interactions using high-throughput genomic data in the model organism *Saccharomyces cerevisiae*. In chapter 4, I extend our methodology in chapter 3 to metazoan organisms, by directing incorporating the tissue contextual information during the learning procedure of biomolecular interaction networks.

Despite our contribution and the tremendous growth in bioinformatics, still many future research opportunities exist in developing computational methods that can guide the investigation of tissue and cell type specific functional information (e.g. tissue-specific gene regulation). Especially, tissue-specific gene regulation has increasingly been appreciated as an important aspect of many human complex diseases (e.g. Alzheimer's disease). Specifically, our future research objective builds on our previous research experience with functional genomic data integration and analysis, and takes advantage of new computational and statistical methods and the continuously increasing genomic and clinical data compendium to address several specific challenges:

Computational inference of tissue-specific pathway components in multicellular organisms

In human biology, unlike in the unicellular organism *Saccharomyces cerevisiae*, many biomolecular pathway components are specific to tissue-types or dependent on the developmental stage cell types. I hope to extend our methodology from chapter 3 and 4 for predicting biomolecular pathway components, to multicellular organisms with a focus on inferring tissue specificity of regulatory pathways and human disease related pathways. The establishment of these tissue-specific pathway components in mammalian organisms will be especially useful to model the effect of genetic diseases since both diseases and drugs often have their specific target tissue.

Classification of tissue-specific post-translational regulation activity

Elucidating the relationship between post-translational regulation and the target biological components is critical for many diseases, because post-translational modification (PTM) events alter the property of many proteins (e.g. acitivity, localization) and often proteins with unintended modifications can lead to various cancers. Discovery through conventional approaches has been an extremely lengthy and expensive process, but recently there has been an increase in proteomic data measuring the PTM state of proteins in various cell types and conditions. A future research goal is to develop statistical machine learning algorithms to leverage the newly available data to address the tissue and cell type specific PTMs unique to each E3 ligases and protein kinase that can lead to new tissue specific drug targets.

Prioritizing experiments from genomic data and electronic medical records

Most experimental designs are based on the biomedical researcher's subject knowledge or laborious experimental trials. However, the coupling of a large genomic data compendium with

newly available electronic medical records can provide a holistic view of data coverage across many tissue, biological processes and diseases. A future research goal is to develop novel statistical models that can highlight experiments testing biological processes, tissues or diseases orthogonal to the existing experimental knowledge base. This research will build on our prior work of quantifying the landscape of experimental knowledge in each model organism for functional knowledge transfer and will aim to reduce the laborious human effort to characterize disease-associated genes.

6 References

1. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74: 5463-5467.

2. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotech 26: 1135-1145.

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

4. Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, et al. (2001) Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. Science 294: 2364-2368.

5. Bassik Michael C, Kampmann M, Lebbink Robert J, Wang S, Hein Marco Y, et al. (2013) A Systematic Mammalian Genetic Interaction Map Reveals Pathways Underlying Ricin Susceptibility. Cell 152: 909-922.

6. Bassett Andrew R, Tibbit C, Ponting Chris P, Liu J-L (2013) Highly Efficient Targeted Mutagenesis of Drosophila with the CRISPR/Cas9 System. Cell Reports 4: 220-228.

7. Fields S, Sternglanz R (1994) The two-hybrid system: an assay for protein-protein interactions. Trends in Genetics 10: 286-292.

8. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422: 198-207.

9. Darnell RB (2010) HITS-CLIP: panoramic views of protein–RNA regulation in living cells. Wiley Interdisciplinary Reviews - RNA 1: 266-286.

10. Watson JD, Crick FHC (1953) THE STRUCTURE OF DNA. Cold Spring Harbor Symposia on Quantitative Biology 18: 123-131.

11. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

12. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669-680.

13. Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437: 1173-1178.

14. Garraway Levi A, Lander Eric S (2013) Lessons from the Cancer Genome. Cell 153: 17-37.

15. Jacob HJ, Abrams K, Bick DP, Brodie K, Dimmock DP, et al. (2013) Genomics in Clinical Practice: Lessons from the Front Lines. Science Translational Medicine 5: 194cm195.

16. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, et al. (2012) Wisdom of crowds for robust gene network inference. Nat Meth 9: 796-804.

17. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, et al. (2010) CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Research 38: D497-D501.

18. Bloom J, Cross FR (2007) Multiple levels of cyclin specificity in cell-cycle control. Nat Rev Mol Cell Biol 8: 149-160.

19. Nasmyth K, Dirick L (1991) The role of SWI4 and SWI6 in the activity of G1 cyclins in yeast. Cell 66: 995-1013.

20. Skowyra D, Craig KL, Tyers M, Elledge SJ, Harper JW (1997) F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. Cell 91: 209-219.

21. Kaletta T, Hengartner MO (2006) Finding function in novel targets: C. elegans as a model organism. Nat Rev Drug Discov 5: 387-399.

22. Botstein D, Fink GR (2011) Yeast: An Experimental Organism for 21st Century Biology. Genetics 189: 695-704.

23. Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT, et al. (2011) The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Research 39: D842-D848.

24. Chikina MD, Troyanskaya OG (2011) Accurate quantification of functional analogy among close homologs. PLoS Comput Biol 7: e1001074.

25. Pena-Castillo L, Tasan M, Myers C, Lee H, Joshi T, et al. (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biology 9: S2.

26. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, et al. (2013) A large-scale evaluation of computational protein function prediction. Nat Meth 10: 221-227.

27. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102: 15545-15550.

28. Myers C, Robson D, Wible A, Hibbs M, Chiriac C, et al. (2005) Discovery of biological networks from diverse functional genomic data. Genome Biology 6: R114.

29. Efron B, Tibshirani R (2007) On testing the significance of sets of genes. 107-129.

30. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biology 9: S4.

31. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols 4: 44-57.

32. Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, et al. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. Bioinformatics 23: 2692-2699.

33. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A (2012) EnrichNet: network-based gene set enrichment analysis. Bioinformatics 28: i451-i457.

34. Khatri P, Drăghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21: 3587-3595.

35. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, et al. (2007) STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Research 35: D358-362.

36. Alexeyenko A, Sonnhammer ELL (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Research 19: 1107-1116.

37. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Research 41: D808-D815.

38. Schmitt T, Ogris C, Sonnhammer ELL (2014) FunCoup 3.0: database of genome-wide functional coupling networks. Nucleic Acids Research 42: D380-D388.

39. Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32: D258-261.

40. Noble WS (2006) What is a support vector machine? Nat Biotech 24: 1565-1567.

41. Breiman L (2001) Random Forests. Machine Learning 45: 5-32.

42. Park CY, Wong AK, Greene CS, Rowland J, Guan Y, et al. (2013) Functional Knowledge Transfer for High-accuracy Prediction of Under-studied Biological Processes. PLoS Comput Biol 9: e1002957.

43. Wong AK, Park CY, Greene CS, Bongo LA, Guan Y, et al. (2012) IMP: A multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. Nucleic Acids Research: In press.

44. Park CY, Hess DC, Huttenhower C, Troyanskaya OG (2010) Simultaneous Genome-Wide Inference of Physical, Genetic, Regulatory, and Functional Pathway Components. PLoS Comput Biol 6: e1001009.

45. Guan Y, Myers C, Hess D, Barutcuoglu Z, Caudy A, et al. (2008) Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biology 9: S3.

46. Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T (1999) Prediction of Gene Function by Genome-Scale Expression Analysis: Prostate Cancer-Associated Genes. Genome Research 9: 1198-1203.

47. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, et al. (2002) Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. Nat Genet 31: 255-265.

48. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, et al. (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. Mol Syst Biol 1.

49. Kim W, Krumpelman C, Marcotte E (2008) Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. Genome Biology 9: S5.

50. Pavlidis P, Weston J, Cai J, Grundy WN (2001) Gene functional classification from heterogeneous data. Proceedings of the fifth annual international conference on Computational biology. Montreal, Quebec, Canada: ACM. pp. 249-255.

51. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, et al. (2005) Discovery of biological networks from diverse functional genomic data. Genome Biol 6: R114.

52. Mostafavi S, Morris Q (2010) Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics 26: 1759-1765.

53. Greene CS, Troyanskaya OG (2011) PILGRM: an interactive data-driven discovery platform for expert biologists. Nucleic Acids Research 39: W368-W374.

54. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional Discovery via a Compendium of Expression Profiles. Cell 102: 109-126.

55. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402: 83-86.

56. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, et al. (2002) Prediction of Human Protein Function from Post-translational Modifications and Localization Features. Journal of Molecular Biology 319: 1257-1265.

57. Barutcuoglu Z, Schapire RE, Troyanskaya OG (2006) Hierarchical multi-label prediction of gene function. Bioinformatics 22: 830-836.

58. Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Global protein function prediction from protein-protein interaction networks. Nat Biotech 21: 697-700.

59. Hess DC, Myers CL, Huttenhower C, Hibbs MA, Hayes AP, et al. (2009) Computationally Driven, Quantitative Experiments Discover Genes Required for Mitochondrial Biogenesis. PLoS Genet 5: e1000407.

60. Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY (2010) Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nat Biotech 28: 149-156.

61. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95: 14863-14868.

62. Eisen JA (1998) Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. Genome Research 8: 163-167.

63. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, et al. (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34: D572-580.

64. O'Brien KP, Remm M, Sonnhammer ELL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Research 33: D476-D480.

65. Valenzuela DM, Griffiths JA, Rojas J, Aldrich TH, Jones PF, et al. (1999) Angiopoietins 3 and 4: Diverging gene counterparts in mice and humans. Proceedings of the National Academy of Sciences 96: 1904-1909.

66. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) LIBLINEAR: A Library for Large Linear Classification. J Mach Learn Res 9: 1871-1874.

67. Hwang S, Rhee SY, Marcotte EM, Lee I (2011) Systematic prediction of gene function in Arabidopsis thaliana using a probabilistic functional gene network. Nat Protocols 6: 1429-1442.

68. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

69. Hofken T, Schiebel E (2002) A role for cell polarity proteins in mitotic exit. EMBO J 21: 4851-4862.

70. Matei V, Pauley S, Kaing S, Rowitch D, Beisel KW, et al. (2005) Smaller inner ear sensory epithelia in Neurog1 null mice are related to earlier hair cell cycle exit. Developmental Dynamics 234: 633-650.

71. Garner M, van Kreeveld S, Su TT (2001) mei-41 and bub1 block mitosis at two distinct steps in response to incomplete DNA replication in Drosophila embryos. Current Biology 11: 1595-1599.

72. Yamaguchi M, Imai F, Tonou-Fujimori N, Masai I (2010) Mutations in N-cadherin and a Stardust homolog, Nagie oko, affect cell-cycle exit in zebrafish retina. Mechanisms of Development 127: 247-264.

73. Hartwell L, Weinert T (1989) Checkpoints: controls that ensure the order of cell cycle events. Science 246: 629-634.

74. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, et al. (2000) Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. Mol Biol Cell 11: 4241-4257.

75. Kuhar SG, Feng L, Vidan S, Ross ME, Hatten ME, et al. (1993) Changing patterns of gene expression define four stages of cerebellar granule neuron differentiation. Development 117: 97-104.

76. Furlong EEM, Andersen EC, Null B, White KP, Scott MP (2001) Patterns of Gene Expression During Drosophila Mesoderm Development. Science 293: 1629-1633.

77. Arlotta P, Molyneaux BJ, Chen J, Inoue J, Kominami R, et al. (2005) Neuronal Subtype-Specific Genes that Control Corticospinal Motor Neuron Development In Vivo. Neuron 45: 207-221.

78. Liu L, Ji C, Chen J, Li Y, Fu X, et al. (2008) A global genomic view of MIF knockdown-mediated cell cycle arrest. Cell Cycle 7: 1678-1692.

79. Mackley JR, Ando J, Herzyk P, Winder SJ (2006) Phenotypic responses to mechanical stress in fibroblasts from tendon, cornea and skin. Biochemical Journal 396: 307-316.

80. Hunt-Newbury R, Viveiros R, Johnsen R, Mah A, Anastas D, et al. (2007) High-Throughput In Vivo Analysis of Gene Expression in *Caenorhabditis elegans*. PLoS Biol 5: e237.

81. Hovatta I, Tennant RS, Helton R, Marr RA, Singer O, et al. (2005) Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. Nature 438: 662-666.

82. Carter T, Greenhall J, Yoshida S, Fuchs S, Helton R, et al. (2005) Mechanisms of aging in senescence-accelerated mice. Genome Biology 6: R48.

83. Ramsdell AF (2005) Left–right asymmetry and congenital cardiac defects: Getting to the heart of the matter in vertebrate left–right axis determination. Developmental Biology 288: 1-20.

84. van der Linde D, Konings EEM, Slager MA, Witsenburg M, Helbing WA, et al. (2011) Birth Prevalence of Congenital Heart Disease Worldwide: A Systematic Review and Meta-Analysis. Journal of the American College of Cardiology 58: 2241-2247.

85. Baker K, Holtzman NG, Burdine RD (2008) Direct and indirect roles for Nodal signaling in two axis conversions during asymmetric morphogenesis of the zebrafish heart. Proceedings of the National Academy of Sciences 105: 13924-13929.

86. Smith KA, Chocron S, von der Hardt S, de Pater E, Soufan A, et al. (2008) Rotation and Asymmetric Development of the Zebrafish Heart Requires Directed Migration of Cardiac Progenitor Cells. Developmental Cell 14: 287-297.

87. Rohr S, Otten C, Abdelilah-Seyfried S (2008) Asymmetric Involution of the Myocardial Field Drives Heart Tube Formation in Zebrafish. Circulation Research 102: e12-e19.

88. de Campos-Baptista MIM, Holtzman NG, Yelon D, Schier AF (2008) Nodal signaling promotes the speed and directional movement of cardiomyocytes in zebrafish. Developmental Dynamics 237: 3624-3633.

89. Wang X, Yost HJ (2008) Initiation and propagation of posterior to anterior (PA) waves in zebrafish left–right development. Developmental Dynamics 237: 3640-3647.

90. Liang JO, Etheridge A, Hantsoo L, Rubinstein AL, Nowak SJ, et al. (2000) Asymmetric nodal signaling in the zebrafish diencephalon positions the pineal organ. Development 127: 5101-5112.

91. Essner JJ, Amack JD, Nyholm MK, Harris EB, Yost HJ (2005) Kupffer's vesicle is a ciliated organ of asymmetry in the zebrafish embryo that initiates left-right development of the brain, heart and gut. Development 132: 1247-1260.

92. Alexander J, Rothenberg M, Henry GL, Stainier DYR (1999) casanova Plays an Early and Essential Role in Endoderm Formation in Zebrafish. Developmental Biology 215: 343-357.

93. Rebagliati MR, Toyama R, Fricke C, Haffter P, Dawid IB (1998) Zebrafish Nodal-Related Genes Are Implicated in Axial Patterning and Establishing Left–Right Asymmetry. Developmental Biology 199: 261-272.

94. Hami D, Grimes AC, Tsai H-J, Kirby ML (2011) Zebrafish cardiac development requires a conserved secondary heart field. Development 138: 2389-2398.

95. Feldman B, Concha ML, Saúde L, Parsons MJ, Adams RJ, et al. (2002) Lefty Antagonism of Squint Is Essential for Normal Gastrulation. Current Biology 12: 2129-2135.

96. Lenhart KF, Lin S-Y, Titus TA, Postlethwait JH, Burdine RD (2011) Two additional midline barriers function with midline lefty1 expression to maintain asymmetric Nodal signaling during left-right axis specification in zebrafish. Development 138: 4405-4410.

97. Smith KA, Noël E, Thurlings I, Rehmann H, Chocron S, et al. (2011) Bmp and Nodal Independently Regulate *lefty1* Expression to Maintain Unilateral Nodal Activity during Left-Right Axis Specification in Zebrafish. PLoS Genet 7: e1002289.

98. Goudevenou K, Martin P, Yeh Y-J, Jones P, Sablitzky F (2011) Def6 Is Required for Convergent Extension Movements during Zebrafish Gastrulation Downstream of Wnt5b Signaling. PLoS ONE 6: e26548.

99. Corey D, Abrams J (2001) Morpholino antisense oligonucleotides: tools for investigating vertebrate development. Genome Biology 2: reviews1015.1011 - reviews1015.1013.

100. Lopes CAM, Prosser SL, Romio L, Hirst RA, O'Callaghan C, et al. (2011) Centriolar satellites are assembly points for proteins implicated in human ciliopathies, including oral-facial-digital syndrome 1. Journal of Cell Science 124: 600-612.

101. Glazer AM, Wilkinson AW, Backer CB, Lapan SW, Gutzman JH, et al. (2010) The Zn Finger protein Iguana impacts Hedgehog signaling by promoting ciliogenesis. Developmental Biology 337: 148-156.

102. Aamar E, Dawid IB (2010) Sox17 and chordin are required for formation of Kupffer's vesicle and left-right asymmetry determination in zebrafish. Developmental Dynamics 239: 2980-2988.

103. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proceedings of the National Academy of Sciences 100: 8348-8353.

104. Lee I, Date SV, Adai AT, Marcotte EM (2004) A Probabilistic Functional Network of Yeast Genes. Science 306: 1555-1558.

105. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, et al. (2005) Probabilistic model of the human protein-protein interaction network. Nat Biotech 23: 951-959.

106. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, et al. (2009) Exploring the human genome with functional maps. Genome Research 19: 1093-1106.

107. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39: D698-704.

108. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, et al. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40: D841-846.

109. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40: D857-861.

110. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, et al. (2000) MIPS: a database for genomes and protein sequences. Nucleic Acids Research 28: 37-40.

111. Abdulrehman D, Monteiro PT, Teixeira MC, Mira NP, Lourenço AB, et al. (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. Nucleic Acids Research 39: D136-D140.

112. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) JASPAR: an open access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research 32: D91-D94.

113. Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif. Bioinformatics 27: 1017-1018.

114. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, et al. (1998) SGD: Saccharomyces Genome Database. Nucleic Acids Research 26: 73-79.

115. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA, et al. (2008) The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Research 36: D724-D728.

116. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J (2001) WormBase: network access to the genome and biology of Caenorhabditis elegans. Nucleic Acids Research 29: 82-86.

117. Drysdale RA, Crosby MA, Consortium TF (2005) FlyBase: genes and gene models. Nucleic Acids Research 33: D390-D395.

118. Sprague J, Clements D, Conlin T, Edwards P, Frazer K, et al. (2003) The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Research 31: 241-243.

119. Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. Database 2011.

120. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Research 32: D138-D141.

121. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, et al. The PROSITE database. Nucleic Acids Research 34: D227-D230.

122. Myers C, Barrett D, Hibbs M, Huttenhower C, Troyanskaya O (2006) Finding function: evaluation methods for functional genomic data. BMC Genomics 7: 187.

123. Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M (2012) The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. Methods Mol Biol 802: 19-39.

124. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, et al. (2009) PID: the Pathway Interaction Database. Nucleic Acids Res 37: D674-679.

125. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 40: D742-753.

126. Guan Y, Ackert-Bicknell CL, Kell B, Troyanskaya OG, Hibbs MA (2010) Functional genomics complements quantitative genetics in identifying disease-gene associations. PLoS Comput Biol 6: e1000991.

127. John CP (1999) Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. MIT Press.

128. Hoerl AE, Kennard RW (2000) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 42: 80-86.

129. Tibshirani R (1994) Regression shrinkage and selection via the lasso.

130. Diaz-Uriarte R, Alvarez de Andres S (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7: 3.

131. Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG (2008) The Sleipnir library for computational functional genomics. Bioinformatics 24: 1559-1561.

132. Joachims T (2006) Training linear SVMs in linear time. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA: ACM. pp. 217-226.

133. Huang C-J, Tu C-T, Hsiao C-D, Hsieh F-J, Tsai H-J (2003) Germ-line transmission of a myocardium-specific GFP transgene reveals critical regulatory elements in the cardiac myosin light chain 2 promoter of zebrafish. Developmental Dynamics 228: 30-40.

134. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415: 180-183.

135. Mann M, Jensen ON (2003) Proteomic analysis of post-translational modifications. Nature Biotechnology 21: 255-261.

136. Hershko A, Ciechanover A (1998) The Ubiquitin system. Annual Review of Biochemistry 67: 425-479.

137. Cowell IG (1994) Repression versus activation in the control of gene transcription. Trends in Biochemical Sciences 19: 38-42.

138. Barutcuoglu Z, Schapire R, Troyanskaya O (2006) Hierarchical multi-label prediction of gene function. Bioinformatics 22: 830 - 836.

139. Friedman N (2004) Inferring Cellular Networks Using Probabilistic Graphical Models. Science 303: 799-805.

140. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science 308: 523-529.

141. Pe'er D, Tanay A, Regev A (2006) MinReg: A Scalable Algorithm for Learning Parsimonious Regulatory Networks in Yeast and Mammals. Journal of Machine Learning Research 7: 167-189.

142. Hartemink A, Gifford D, Jaakkola T, Young R (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pacific Symposium on Biocomputing: 422 - 433.

143. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, et al. (2005) Probabilistic model of the human protein-protein interaction network. Nature Biotechnology 23: 951-959.

144. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, et al. (2004) Global Mapping of the Yeast Genetic Interaction Network. Science 303: 808-813.

145. Wong SL, Zhang LV, Tong AHY, Li Z, Goldberg DS, et al. (2004) Combining biological networks to predict genetic interactions. PNAS 101: 15682-15687.

146. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD (2004) Flux Coupling Analysis of Genome-Scale Metabolic Network Reconstructions. Genome Research 14: 301-312.

147. Lee I, Date S, Adai A, Marcotte E (2004) A probabilistic functional network of yeast genes. Science 306: 1555 - 1558.

148. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan N, et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302: 449 - 453.

149. Qiu J, Noble WS (2008) Predicting Co-Complexed Protein Pairs from Heterogeneous Data. PLoS Comput Biol 4: e1000054.

150. Qi Y, Bar-Joseph Z, Klein-Seetharaman J (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins: Structure, Function, and Bioinformatics 63: 490-500.

151. Ben-Hur A, Noble WS (2005) Kernel methods for predicting protein-protein interactions. Bioinformatics 21: i38-46.

152. Hess DC, Myers CL, Huttenhower C, Hibbs MA, Hayes AP, et al. (2009) Computationally Driven, Quantitative Experiments Discover Genes Required for Mitochondrial Biogenesis. PLoS Genetics 5: e1000407.

153. Ratnakumar S, Kacherovsky N, Arms E, Young ET (2009) Snf1 Controls the Activity of Adr1 Through Dephosphorylation of Ser230. Genetics 182: 735-745.

154. Schneper L, Düvel K, Broach JR (2004) Sense and sensibility: nutritional response and signal integration in yeast. Current Opinion in Microbiology 7: 624-630.

155. Pausch MH, Kaim D, Kunisawa R, Admon A, Thorner J (1991) Multiple Ca2+/calmodulin-dependent protein-kinase genes in a unicellular eukaryote. Embo Journal 10: 1511-1522.

156. Thon VJ, Vigneronlesens C, Mariannepepin T, Montreuil J, Decq A, et al. (1992) Coordinate regulation of glycogen-metabolism in the yeast Saccharomyces-cerevisiae - induction of glycogen branching enzyme. Journal of Biological Chemistry 267: 15224-15228.

157. Jahn R, Scheller RH (2006) SNAREs - engines for membrane fusion. Nature Reviews Molecular Cell Biology 7: 631-643.

158. Gaynor EC, Chen C-Y, Emr SD, Graham TR (1998) ARF Is Required for Maintenance of Yeast Golgi and Endosome Structure and Function. Molecular Biology of the Cell 9: 653-670.

159. Sapperstein SK, Lupashin VV, Schmitt HD, Waters MG (1996) Assembly of the ER to Golgi SNARE complex requires Uso1p. The Journal of Cell Biology 132: 755-767.

160. Newman AP, Ferronovick S (1990) Defining components required for transport from the ER to the Golgi-complex in yeast. Bioessays 12: 485-491.

161. Gabriely G, Kama R, Gerst JE (2007) Involvement of Specific COPI Subunits in Protein Sorting from the Late Endosome to the Vacuole in Yeast. Molecular Biology of the Cell 27: 526-540.

162. Wilsbach K, Payne GS (1993) Vps1p, a member of the dynamin GTPase family, is necessary for Golgi membrane-protein retention in Saccharomyces-cerevisiae. Embo Journal 12: 3049-3059.

163. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29.

164. Yoshikawa K, Tanaka T, Furusawa C, Nagahisa K, Hirasawa T, et al. (2009) Comprehensive phenotypic analysis for identification of genes affecting growth under ethanol stress in Saccharomyces cerevisiae. FEMS Yeast Research 9: 32-44.

165. Watts DJ, Strogatz SH (1998) Collective dynamics of `small-world' networks. Nature 393: 440-442.

166. Barabasi A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5: 101-113.

167. Przulj N (2007) Biological network comparison using graphlet degree distribution. Bioinformatics 23: e177-183.

168. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network Motifs: Simple Building Blocks of Complex Networks. Science 298: 824-827.

169. Eichenberger P (2004) The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis. PLoS Biology 2: e328.

170. Lee TI (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298: 799-804.

171. Boyer LA (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947-956.

172. Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. Proceedings of the National Academy of Sciences of the United States of America 100: 11980-11985.

173. Alon U (2007) Network motifs: theory and experimental approaches. Nature Reviews Genetics 8: 450-461.

174. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. (2002) Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science 298: 799-804.

175. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics 31: 64-68.

176. Artzy-Randrup Y, Fleishman SJ, Ben-Tal N, Stone L (2004) Comment on "Network Motifs: Simple Building Blocks of Complex Networks" and "Superfamilies of Evolved and Designed Networks". Science 305: 1107c-.

177. Holz MK, Ballif BA, Gygi SP, Blenis J (2005) mTOR and S6K1 Mediate Assembly of the Translation Preinitiation Complex through Dynamic Protein Interchange and Ordered Phosphorylation Events. Cell 123: 569-580.

178. Saunders NFW, Kobe B (2008) The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. Nucleic Acids Research: gkn279.

179. Liu Y, Tozeren A (2010) Modular composition predicts kinase/substrate interactions. BMC Bioinformatics 11: 349.

180. Ratsch E, Schultz Jo, Saric J, Lavin PC, Wittig U, et al. (2003) Developing a Protein Interactions Ontology. Comparative and Functional Genomics 4: 85–89.

181. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, et al. (2004) The HUPO PSI's Molecular Interaction format - a community standard for the representation of protein interaction data. Nature Biotechnology 22: 177-183.

182. Cochrane G, Akhtar R, Bonfield J, Bower L, Demiralp F, et al. (2009) Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Research 37: D19-25.

183. Cherry J, Adler C, Ball C, Chervitz S, Dwight S, et al. (1998) SGD: Saccharomyces Genome Database. Nucleic Acids Research 26: 73-79.

184. MacIsaac K, Wang T, Gordon DB, Gifford D, Stormo G, et al. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 7: 113.

185. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28: 27 - 30.

186. Ben-Hur A, Noble W (2005) Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics 7: S2.

187. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520-525.

188. Huh W, Falvo J, Gerke L, Carroll A, Howson R, et al. (2003) Global analysis of protein localization in budding yeast. Nature 425: 686 - 691.

189. Finn R, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: clans, web tools and services. Nucleic Acids Research 34: D247 - 251.

190. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC(R): transcriptional regulation, from patterns to profiles. Nucleic Acids Research 31: 374-378.

191. Breiman L (1996) Bagging predictors. Machine Learning. pp. 123-140.

192. Wernicke S, Rasche F (2006) FANMOD: a tool for fast network motif detection. Bioinformatics 22: 1152-1153.

193. Milenkovic T, Lai J, Przulj N (2008) GraphCrunch: A tool for large network analyses. BMC Bioinformatics 9: 70.

194. Joachims T (2006) Training linear SVMs in linear time. Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 20 to 23 August; Philadelphia, PA: 217 - 226.

195. Lauritzen SL, Wermuth N (1989) Graphical Models for Associations between Variables, some of which are Qualitative and some Quantitative. The Annals of Statistics 17: 31-57.

196. Druzdzel M (1999) SMILE: structural modeling, inference, and learning engine and genie: a development environment for graphical decision-theoretic models. Proceedings of the Sixteenth National Conference on Artificial Intelligence: 18 to 22 July 1999; Orlando, FL: 902 - 903.

197. Hand SC, Hardewig I (1996) Downregulation of Cellular Metabolism During Environmental Stress: Mechanisms and Implications. Annual Review of Physiology 58: 539-563.

198. de Lange T (2005) Shelterin: the protein complex that shapes and safeguards human telomeres. Genes & Development 19: 2100-2110.

199. Wistow GJ, Piatigorsky J (1988) Lens Crystallins: The Evolution and Expression of Proteins for a Highly Specialized Tissue. Annual Review of Biochemistry 57: 479-504.

200. Britten RJ, Davidson EH (1969) Gene Regulation for Higher Cells: A Theory. Science 165: 349-357.

201. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, et al. (2007) STRING 7--recent developments in the integration and prediction of protein interactions. Nucl Acids Res 35: D358-362.

202. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. Cell 122: 957-968.

203. Lee I, Date S, Adai A, Marcotte E (2004) A probabilistic functional network of yeast genes. Science 306: 1555 - 1558.

204. Myers C, Robson D, Wible A, Hibbs M, Chiriac C, et al. (2005) Discovery of biological networks from diverse functional genomic data. Genome Biol 6: R114.

205. Date SV, Stoeckert CJ (2006) Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. Genome Research 16: 542-549.

206. Feizi S, Marbach D, Medard M, Kellis M (2013) Network deconvolution as a general method to distinguish direct dependencies in networks. Nat Biotech 31: 726-733.

207. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics 7: S7.

208. Haynes BC, Maier EJ, Kramer MH, Wang PI, Brown H, et al. (2013) Mapping functional transcription factor networks from gene expression data. Genome Research 23: 1319-1328.

209. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research 22: 1813-1831.

210. Kim J-Y, Welsh EA, Oguz U, Fang B, Bai Y, et al. (2013) Dissection of TBK1 signaling via phosphoproteomics in lung cancer cells. Proceedings of the National Academy of Sciences 110: 12414-12419.

211. Myatt SS, Lam EWF (2007) The emerging roles of forkhead box (Fox) proteins in cancer. Nat Rev Cancer 7: 847-859.

212. Brunet A, Park J, Tran H, Hu LS, Hemmings BA, et al. (2001) Protein Kinase SGK Mediates Survival Signals by Phosphorylating the Forkhead Transcription Factor FKHRL1 (FOXO3a). Molecular and Cellular Biology 21: 952-965.

213. Fernández de Mattos S, Essafi A, Soeiro I, Pietersen AM, Birkenkamp KU, et al. (2004) FoxO3a and BCR-ABL Regulate cyclin D2 Transcription through a STAT5/BCL6-Dependent Mechanism. Molecular and Cellular Biology 24: 10058-10071.

214. Lehtinen MK, Yuan Z, Boag PR, Yang Y, Villén J, et al. (2006) A Conserved MST-FOXO Signaling Pathway Mediates Oxidative-Stress Responses and Extends Life Span. Cell 125: 987-1001.

215. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489: 91-100.

216. Mercer TR, Mattick JS (2013) Understanding the regulatory and transcriptional complexity of the genome through structure. Genome Research 23: 1081-1088.

217. Tiwari VK, Stadler MB, Wirbelauer C, Paro R, Schubeler D, et al. (2012) A chromatin-modifying function of JNK during stem cell differentiation. Nat Genet 44: 94-100.

218. Nesvizhskii AI, Vitek O, Aebersold R (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Meth 4: 787-797.

219. Nishida E, Gotoh Y (1993) The MAP kinase cascade is essential for diverse signal transduction pathways. Trends in Biochemical Sciences 18: 128-131.

220. Guo J, Kim D, Gao J, Kurtyka C, Chen H, et al. (2013) IKBKE is induced by STAT3 and tobacco carcinogen and determines chemosensitivity in non-small cell lung cancer. Oncogene 32: 151-159.

221. Elemento O, Slonim N, Tavazoie S (2007) A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. Molecular Cell 28: 337-350.

222. Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, et al. (2010) Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. Nucleic Acids Research 38: e120.

223. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database—2009 update. Nucleic Acids Research 37: D767-D772.

224. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10: 252-263.

225. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The Protein Kinase Complement of the Human Genome. Science 298: 1912-1934.

226. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 30: 207-210.

227. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, et al. (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489: 83-90.

228. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Research 36: D154-D158.

229. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, et al. (2013) A compendium of RNA-binding motifs for decoding gene regulation. Nature 499: 172-177.

230. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, et al. (2010) Phospho.ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Research.

231. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, et al. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Research 40: D261-D270.

232. Zhang QC, Petrey D, Garzón JI, Deng L, Honig B (2013) PrePPI: a structure-informed database of protein–protein interactions. Nucleic Acids Research 41: D828-D833.

233. Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R (2011) DOMINE: a comprehensive collection of known and predicted domain-domain interactions. Nucleic Acids Research 39: D730-D735.

234. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32: D115-D119.

235. Minguez P, Letunic I, Parca L, Bork P (2012) PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. Nucleic Acids Research.

236. Zilliox MJ, Irizarry RA (2007) A gene expression bar code for microarray data. Nat Meth 4: 911-913.

237. McCall MN, Uppal K, Jaffee HA, Zilliox MJ, Irizarry RA (2011) The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Research 39: D1011-D1015.

238. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, et al. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Research 39: D507-D513.

239. Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, et al. (2002) Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. Molecular & Cellular Proteomics 1: 376-386.

240. Lieber DS, Elemento O, Tavazoie S (2010) Large-Scale Discovery and Characterization of Protein Regulatory Motifs in Eukaryotes. PLoS ONE 5: e14444.