Statistical methods to infer biological interactions

by

George Jay Tucker

B.S., Harvey Mudd College (2008)

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Applied Mathematics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Author . . . .
Department of Mathematics
May 1, 2014

Certified by . . . .
Bonnie Berger
Professor of Applied Mathematics
Thesis Supervisor

Accepted by . . . .
Michel X. Goemans
Chairman, Applied Mathematics Committee

Statistical methods to infer biological interactions

by

George Jay Tucker

Submitted to the Department of Mathematics
on May 1, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Applied Mathematics

Abstract

Biological systems are extremely complex, and our ability to experimentally measure interactions in these systems is limited by inherent noise. Technological advances have allowed us to collect unprecedented amounts of raw data, increasing the need for computational methods to disentangle true interactions from noise. In this thesis, we focus on statistical methods to infer two classes of important biological interactions: protein-protein interactions and the link between genotypes and phenotypes. In the first part of the thesis, we introduce methods to infer protein-protein interactions from affinity purification mass spectrometry (AP-MS) and from luminescence-based mammalian interactome mapping (LUMIER). Our work reveals novel context dependent interactions in the MAPK signaling pathway and insights into the protein homeostasis machinery. In the second part, we focus on methods to understand the link between genotypes and phenotypes. First, we characterize the effects of related individuals on standard association statistics for genome-wide association studies (GWAS) and introduce a new statistic that corrects for relatedness. Then, we introduce a statistically powerful association testing framework that corrects for confounding from population structure in large scale GWAS. Lastly, we investigate regularized regression for phenotype prediction from genetic data.

Thesis Supervisor: Bonnie Berger
Title: Professor of Applied Mathematics

Acknowledgments

This journey would not have been possible without the many people that have supported me throughout my time at MIT. First, I would like to thank my advisor, Bonnie Berger, for her guidance and her unwavering support. I would also like to thank Alkes Price for welcoming me into his group meetings and journal clubs, for countless interesting discussions about medical genomics, and for mentorship as I learned about statistical genetics. My family and friends have supported me throughout my time here. In particular, my heartfelt thanks go to:

> Po-Ru Loh, for being an inspiration and exceptional friend. I will always be grateful that I had the chance to work with someone so hard-working, careful, considerate, and humble.
>
> Mark Lipson, for our discussions about research and everyday life over our weekly lunches.
>
> Jian Peng, for teaching me that research is difficult even if it feels like it shouldn't be and that's okay.
>
> The rest of the members of the Berger lab, in particular Alex Levin, Irene Kaplow, Leonid Chindelevitch, Deniz Yorukoglu, Fulton Wang and Sean Simmons, for teaching me how to do research.
>
> Patrice Macaluso, for keeping us sane and staving off chaos.
>
> Mikko Taipale, for being an amazing collaborator and general academic badass.
>
> Polina Golland, for welcoming me into her reading group and teaching me to ask questions about the "trivial" things that usually turn out to confuse everyone.
>
> Mark Behrens, for mentoring me through my short trip in Algebraic Topology and for being understanding.

Lastly, I would like to thank my love, Holly Johnsen, who has been with me through the best and the darkest times of this journey.

Contents

Chapter 1

Introduction

In this thesis, we focus on inferring two classes of important biological interactions: protein-protein interactions (PPI) and the link between genotypes and phenotypes. Biological systems are extremely complex and our ability to experimentally measure interactions is limited by inherent noise. Computational methods can identify patterns that are invisible to the human eye and disentangle true interactions from noise. Throughout this thesis we draw on a wide range of statistical methods to achieve this goal. In this chapter, we set the context for and summarize the main contributions of this thesis.

1.1 Inferring protein-protein interactions

Proteins are the building blocks of cells, constituting most of the cell's dry mass and executing nearly all cell functions. However, proteins do not act alone; all proteins interact with other molecules, from enzymes catalyzing chemical reactions to proteins transmitting extracellular signals to change gene expression and protein levels. The biological properties of a protein depend on its physical interactions with other proteins. As such, an important way to begin characterizing the biological role of a protein is to identify its binding partners. In the past two decades, significant effort has been devoted to generating comprehensive PPI networks (e.g., [141, 61, 49, 51, 53]) to uncover the molecular basis of genetic interactions and provide functional roles for proteins.

These networks have been used as scaffolds to transfer known annotations to uncharacterized proteins in our lab and others. For example, IsoRank [119] and IsoRankN [80] predict functional orthologs across species by aligning PPI networks. In signaling network reconstruction, perturbation studies are used to reveal the critical components of the pathway. However, in many cases, these studies identify proteins that are not directly part of the core pathway. Huang *et al.* [59] and Yeger Lotem *et al.* [154] developed methods that use network flows and minimal trees in the PPI network to organize these disparate proteins into functionally coherent pathways.

Before we can realize the benefits of a comprehensive PPI network, we first have to generate the interaction network. Mapping protein-protein interactions is extremely time and labor intensive because of the sheer number of potential interactions. Mass spectrometry or affinity purification mass spectrometry (AP-MS) and yeast two-hybrid (Y2H) are two widely used high-throughput techniques for identifying protein interactions. The first large-scale PPI networks were generated for the model organism *Saccharomyces cerevisiae*, initially using yeast two-hybrid screens (Y2H) [141, 61] and subsequently by AP-MS [49, 56]. Similarly, high throughput approaches have been applied to comprehensively map the *Drosophila melanogaster* interactome, initially using Y2H [51] and more recently by AP-MS [53].

Both approaches have advantages and disadvantages. Y2H tests pairs of proteins by introducing them into a yeast cell with a reporter that detects an interaction. Specifically, one protein is fused with a DNA binding domain and the other protein is fused with a transcriptional activation domain and both proteins are expressed in a yeast cell. If the two proteins interact, a reporter gene is transcribed. Y2H is a binary interaction assay, so it may not detect interactions that rely on more than two proteins (e.g. interactions between protein complexes) or other endogenous factors. A typical AP-MS study consists of performing a set of experiments on several proteins of interest, called bait proteins. In each experiment, the bait protein is epitope tagged so that it can be easily purified. Any prey protein that interacts with the bait protein is also pulled down with the bait protein. Finally, the resulting mixture of bait and bound prey proteins is analyzed by mass spectrometry to determine the identity of the interacting prey proteins. AP-MS is done *in vivo*, so interactions that involve endogenous factors or multiple proteins can be detected. However, because it simultaneously pulls down all interactors, interactors that are expressed at low levels may not be detected reliably.

A third recently developed technique, luminescence-based mammalian interactome mapping (LUMIER) [6] and its extension LUMIER with bait control (BACON) [133] detect protein interactions that can be missed by standard interaction assays. LUMIER is a co-affinity purification assay that uses luminescence to measure interaction strength. Renilla luciferase, an enzyme that emits light, is fused to prey proteins. In each interaction test, the prey protein is coexpressed with a tagged bait protein. As in AP-MS, the bait protein is affinity purified and any bound prey protein is pulled down as well. Then, we can measure the luminescence to quantify the abundance of the bound prey protein and determine if an interaction occurred. Because interactions are interrogated *in-vivo*, we can detect interactions that involve additional protein partners and interactions that are contingent on post-translation modifications.

However, in all cases, the raw data include many false positive and false negative interactions, which are serious confounding factors in their interpretation. To address these issues, we introduce computational methods to distinguish true interactions from noise.

Combining a perturbation screen with PPIs to understand the MAPK signaling pathway

As part of joint work with the Perrimon lab, we investigate the canonical MAPK pathway by combining parallel genome-wide RNAi screens with PPI mapping. The PPI mapping was done at baseline and following stimulation with insulin or epidermal growth factor (EGF) to identify interactions that depended on the stimulus. We post-processed the raw AP-MS data and identified context dependent interactions.

Inferring protein interactions from noisy AP-MS data

We introduce a general method that can be used to extend existing PPI inference methods to take advantage of semi-quantitative spectral count information that has recently become widely available in affinity purification for mass spectrometry (AP-MS) data sets. Our approach introduces a probabilistic framework that models the statistical noise inherent in observations of co-purifications. We validate our approach on three MS data sets and demonstrate improvement over state-of-the-art methods.

Our two key contributions are:

- A sampling framework for incorporating quantitative information into existing PPI inference methods. We focus on matrix models for PPI inference, a class of methods that has recently attracted significant research interest because of the ability of matrix models to leverage the rich co-occurrence information in newer, large-scale AP-MS experiments. With few exceptions, existing methods in this class only analyze binary experimental data (in which each potential interaction tested is deemed either observed or unobserved), neglecting quantitative information available from AP-MS such as spectral counts. The framework we propose represents quantitative data sets as ensembles of binary data sets, allowing analysis of each member of the ensemble by direct application of a previous method. The ensemble predictions can then be aggregated to produce a robust prediction that we demonstrate improves performance.

- An in-depth discussion comparing the theoretical bases of existing approaches. We further identify common aspects of established PPI inference methods that may be key to their performance and suggest a common framework for future investigation.

Inferring interactors from LUMIER using mixture models

We describe a novel method for determining significant protein interactions from raw LUMIER data that corrects for spatial biases that occur in high-throughput LUMIER screens. We apply this method to a large LUMIER screen with 60 preys and 800 baits to characterize chaperone, co-chaperone, and client interactions (Taipale, Tucker, Peng, *et al.* "A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways" Cell, in press). We show that our method is able to recover significantly more true interactions than previous methods. From this data, we assemble a comprehensive network of chaperone, co-chaperone, and client interactions that reveals new insights into co-chaperone specificity.

1.2 Statistical genetics

Recent technological advances in the past two decades have revolutionized our understanding of human diseases. The genetic architecture of diseases in humans ranges from diseases that are caused by just a single genetic variants in a single gene to multiple variants in multiple genetic loci contributing to disease risk and often interacting with environmental factors. Most common diseases fall in the latter category of complex genetic disease, necessitating large studies and sophisticated computational methods. The Human Genome Project and the HapMap project have paved the way for genome-wide association studies (GWAS) that have identified hundreds of loci associated with complex diseases.

Genome-wide association studies scan through hundreds of thousands or even millions of genetic variants, called genetic markers, to look for associations between markers and a disease in a large sample of individuals. GWAS test all regions of the genome in an approximately unbiased fashion, including non-coding regions. Results from genome wide association studies have revealed that complex diseases are influenced by genetic variants in non-coding regions as well as coding regions. Previous study designs leveraged related individuals (e.g., sibling pairs), whereas GWAS can be done with unrelated individuals allowing for large samples of tens or hundreds of thousands individuals. This has allowed researchers to discover small effects and to begin to disentangle the genetic factors driving complex diseases. In the past decade, GWAS have had great success, including new findings for many complex diseases: cancer, diabetes, obesity, inflammatory bowel disease (IBD), multiple sclerosis

(MS), and others (National Human Genome Research Institute catalog of published genome-wide association studies: http://www.genome.gov/26525384).

In GWAS, single nucleotide polymorphisms (SNP) are the most commonly used genetic marker. SNPs are single DNA base pair changes, and we typically focus on biallelic SNPs, those with two possible variants in the population. SNPs occur throughout the genome and genotyping hundreds of thousands of SNPs in thousands of individuals is cost effective, making them suitable for use in large-scale GWAS. Fundamentally, GWAS search for correlations between SNP markers and the phenotype. The idea is not that the SNP markers themselves are causal, but rather that the SNP markers are correlated with the causal variants. Nearby genetic variants are inherently correlated (this correlation is called linkage disequilibrium) due to the block inheritance of genetic material. GWAS attempt to cover the genome with a sufficiently dense set of markers so that any causal variant is in reasonably high linkage disequilibrium with a marker SNP. As such, the choice of the marker SNPs in a GWAS is crucial to the success of the study. Ideally, SNPs are chosen such that they have a high correlation with causal variants of the phenotype of interest, however we may not *a priori* know which regions are likely to contain causal variants. The HapMap project has provided a comprehensive linkage disequilibrium map of the entire human genome for multiple ethnicities. This information has been used to judiciously select SNPs to provide coverage of the entire genome for most ethnicities.

Collecting a large number of individuals for a GWAS study is at odds with ensuring the sample is genetically homogeneous, however systematic differences in ancestry between samples, called population stratification, can cause spurious associations or can obscure gene-disease relations. Genetic differences between populations (e.g., northern and souther European) occurs due to random genetic drift and although the differences may be small and random, they can lead to confounding. For example, if the cases and controls have different proportions of two populations, then spurious associations may be identified. This arises because markers that are informative about ancestry then contain information about case-control status.

Although GWAS have presented new opportunities, they have created numerous statistical and computational challenges. When testing hundreds of thousands of markers for association with a phenotype on a sample of only tens of thousands of individuals, many markers will appear to be highly correlated by chance and slight model deviations can cause false positive associations. Statistical methods are necessary to adjust for confounding effects such as population stratification without losing power and to find ways to intelligently apply domain knowledge to maximize power.

Mixed models with related individuals

As GWAS sample sizes get larger, they inevitably contain increasing numbers of related individuals. Mixed models are the state-of-the-art method for calculating association statistics in GWAS and are generally thought to correct for relatedness. Through extensive simulation and application to real genotypes and phenotypes, we clarify when mixed models are properly calibrated and propose a solution when standard mixed models for GWAS fail.

Powerful methods to detect associations in human genome-wide association studies

In recent years, there has been extensive research on mixed models to calculate GWAS association statistics (e.g., [67, 66, 117, 168, 131, 101]). We introduce PC-Select, a novel mixed model approach that addresses a serious concern in a recent state-of-the-art method for computing mixed model association statistics (FaST-LMM Select [84, 83]). While FaST-LMM Select significantly improves power over standard mixed model and linear regression association statistics, a recent Perspective paper [153] shows that FaST-LMM Select can significantly inflate statistics in the presence of population stratification, leading to false positive associations. As population stratification is a serious concern in many large-scale GWAS, this limitation precludes the use of FaST-LMM Select in such studies. Our approach PC-Select overcomes this limitation by including principal components as fixed effects in multiple steps of the algorithm. As a result, we achieve comparable or superior power gains as FaST-LMM

Select, both in the context of population stratification and in its absence, without inflating statistics in the presence of population stratification.

Risk prediction from genotype and gene expression data

Association testing measures the predictive quality of a test marker, so is intimately connected with phenotype prediction. Phenotype prediction from genomic data is a burgeoning field in computational biology with great practical significance due to its medical applications, e.g., predicting susceptibility to disease or response to treatment. Along these lines, the 2010 Dialogue for Reverse Engineering Assessments and Methods (DREAM) Systems Genetics B Challenge asked contestants to predict disease susceptibility of soybean plants to the plant pathogen Phytophthora sojae. We describe a computational method for predicting phenotype from genotype or gene expression data that won a best-performer award in the challenge. We provide a detailed analysis of the applicability of regularized regression techniques to this problem, finding that optimal regularized models pick out fewer than ten predictors (among thousands or tens of thousands available) that achieve small but positive predictive power.

Beyond bioinformatics, our work should also be of interest to the broader community of scientists and researchers that seek to provide objective evaluation of algorithmic methods by establishing benchmarks and running contests. Many examples of such initiatives exist; DREAM for instance was inspired in part by the Critical Assessment of protein Structure Prediction (CASP) competition. On a larger scale, the Netflix Prize contest spurred a great deal of scientific research (and public interest) in machine learning. Careful contest design and choice of performance metrics is essential to the success of such initiatives, however, and as such, a secondary focus of our paper is an analysis of the DREAM5 SysGen B contest methodology, finding in particular that the test set used created high variance in submission results.

Part I

Inferring protein-protein interactions

Chapter 2

Proteomic and Functional Genomic Landscape of Receptor Tyrosine Kinase and Ras to Extracellular Signal-Regulated Kinase Signaling

Abstract

[1]Characterizing the extent and logic of signaling networks is essential to understanding specificity in such physiological and pathophysiological contexts as cell fate decisions and mechanisms of oncogenesis and resistance to chemotherapy. Cell-based RNA interference (RNAi) screens enable the inference of large numbers of genes that regulate signaling pathways but these screens cannot provide network structure directly. We describe an integrated network around the canonical receptor tyrosine kinase (RTK)-Ras-extracellular signal-regulated kinase (ERK) signaling pathway, generated by combining parallel genome-wide RNAi screens with protein-protein interaction (PPI) mapping by tandem affinity purification/mass spectrometry. We found that only a small fraction of the total number of PPI or RNAi screen hits was isolated under all conditions tested and that most of these represented the known canonical pathway components, suggesting that much of the core canonical ERK pathway is

[1] This chapter previously appeared in *Science Signaling* (2011) as "Proteomic and Functional Genomic Landscape of Receptor Tyrosine Kinase and Ras to Extracellular Signal-Regulated Kinase Signaling " by Adam A. Friedman, George Tucker, Rohit Singh, Dong Yan, Arunachalam Vinayagam, Yanhui Hu, Richard Binari, Pengyu Hong, Xiaoyun Sun, Maura Porto, Svetlana Pacifico, Thilakam Murali, Russell L. Finley, Jr., John M. Asara, Bonnie Berger, and Norbert Perrimon [44]. I analyzed the mass spectrometry data to generate the protein interaction network and identified interactions that were context dependent.

known. Because most of the newly identified regulators are likely cell-type and RTK-specific, our analysis provides a resource for understanding how output through this clinically relevant pathway is regulated in different contexts. We report in vivo roles for several of the previously unknown regulators, including CG10289 and PpV, the Drosophila orthologs of two components of the serine/threonine-protein phosphatase 6 complex; the Drosophila ortholog of TepIV, a glycophosphatidylinositol-linked protein mutated in human cancers; CG6453, a noncatalytic subunit of glucosidase II; and Rtf1, a histone methyltransferase.

2.1 Introduction

Intracellular signaling mediated by growth factor-stimulated receptor tyrosine kinases (RTKs), such as those activated by insulin or epidermal growth factor (EGF), acting through Ras to extracellular signal-regulated kinases (ERKs) is required for metazoan development and physiology. Mutations in genes encoding components of this conserved signaling network, the RTK-Ras-ERK pathway, have been repeatedly identified as drivers in multiple malignancies. Understanding the hierarchical relationships among pathway regulators can have profound clinical significance, as exemplified by Kras genotype in determining responsiveness to inhibitors of the epidermal growth factor receptor (EGFR) [68].

A complete understanding of cell signaling through this pathway requires identification of (i) all components of the system, (ii) the quantitative contribution of these components to various signaling outputs, and (iii) the hierarchical relationships, including physical connections, between these components. Systematic functional genetic approaches, such as genome-wide RNA interference (RNAi) screening used to identify previously unknown signaling genes, are inferential in that they do not distinguish between direct and indirect effects. Large-scale protein-protein interaction (PPI) mapping complements genetic studies by revealing physical associations, but fails to reveal the function of interacting proteins or the functional consequences of the interactions. Separate such systems-level functional genomic and interactome studies in the past few years have revealed that signaling is likely propagated within large networks of hundreds of proteins, and thus have challenged linear cascade models derived from traditional reductive approaches [43]. However, each systematic screening approach performed separately suffers from inherent technical limitations of the methods used, leading to false negatives and positives, restricting the comprehensiveness of pathway regulator discovery.

We have previously described an antibody-based, genome-wide RNAi screen assay for ERK activity in Drosophila cells following insulin stimulus [41]. This assay relies on an antibody that recognizes phosphorylated Drosophila ERK (dpERK). We showed specific examples from secondary screens of a small subset of genes that were required downstream of insulin receptor, but not of the EGFR, for activation of ERK in particular cell types, suggesting that many potential components of this pathway may have been missed by a single primary screen [41]. Although multiple RTKs can signal through Ras to ERK, their output is context-dependent despite the apparent similarity in signal propagation through the core pathway [16, 74, 151].

A combined systematic approach using complementary functional genomic and interactome technologies would be more likely to uncover direct regulators and more completely describe the landscape of a signaling pathway [86]. We performed multiple genome-wide RNAi screens in parallel to generating a tandem affinity purification/mass spectrometry (TAP/MS)-based PPI network surrounding the canonical pathway components of the RTK-Ras-ERK signaling pathway, using data from cells responding to insulin or EGF. Although we identified several previously unknown pathway regulators, the functional genomic and interactome data sets suggest that much of the core canonical pathway is complete.

2.2 Results

A functional genomic compendium of RTK-Ras-ERK signaling To comprehensively discover genes that regulate ERK signaling output and to identify other specificity-generating proteins, we conducted four systematic, cell-based RNAi screens for regulators of EGF-stimulated ERK activation in two stable Drosophila cell lines expressing EGFR, S2R+mtDER and Kc167mtDER ([44]: Figure S1A, B). These four screens combined with our two previously published screens performed with S2R+ cells that were unstimulated (baseline) or stimulated with insulin [16] interrogated >20,000 dsRNAs targeting roughly 14,000 Drosophila genes. We compared all six primary screens, divided into three groups by stimulus (insulin, EGF) and cell line (S2R+, Kc) (Figure 2-1A). These screens uncovered 2,677 annotated genes, in addition to 756 unannotated predicted genes (Figure2-1A, [44]: Table S1). As expected, these genes include most of the known canonical pathway-associated genes ([44]: Table S5). We identified both EIF4AIII and mago ([44]: Table S1) as positive regulators in our RNAi screen in Kc cells and these two genes were also found in an RNAi screen for regulators of the mitogen-activated protein kinase (MAPK) pathway in Drosophila S2 cells [4].

Gene Ontology (GO) annotation of the hits from the RNAi screens showed expected enrichment for processes controlled by RTK-Ras-ERK signaling, including tracheal development, photoreceptor differentiation, imaginal disc morphogenesis, and hematopoesis; genes controlling mitosis, neuronal differentiation, cell motility, female gamete generation, and SUMO binding were also enriched in the hits from the RNAi screens ([44]: Table S2). The hits from the RNAi screens were also significantly enriched for proteins conserved in humans and implicated in a human disease ($p < 3.5 \times 10^{-9}$ and 9.8×10^{-4}, respectively), implying that many of the newly identified regulators are also involved in mammalian MAPK signaling. Human orthologs had stronger RNAi scores on average ($p < 0.001$), suggesting that genes with more central roles in the pathway have been conserved.

We observed distinct subsets of genes isolated in the primary RNAi screens under specific cell or RTK-stimulus contexts ([44]: S1C). We were also able to identify genes that were common to both cell types under both stimulus conditions (Figure 2-1B). These genes were quantitatively stronger regulators than the remaining hits ([44]: S1D). Our systematic screens permitted global observation of the processes regulating specificity; compared to all hits from the RNAi screens, those identified in the insulin screen were enriched for cytoskeletal genes and cell cycle processes ($p < 1.3 \times 10^{-6}$ and 0.03, respectively), whereas transcriptional and peptidase activities were enriched in the EGF screen in Kc cells ($p < 4 \times 10^{-4}$ and 0.02, respectively).

Distinct subsets of genes were specific to insulin or EGF signaling in either cell type or were regulated by insulin or EGF in both cell types ([44]: Table S3). Signaling downstream of the insulin receptor (InR) activates both ERK and Akt signaling pathways; we confirmed that genes encoding components of the Akt-Tor pathway, including InR itself, PTEN, Akt, Tor, and gig (Tsc2) were insulin-specific regulators of ERK. This insulin-specific regulation of ERK and the Akt-Tor pathway is likely mediated through feedback from S6 kinase to InR [72]. (Note throughout the text where different from the Drosophila gene or protein names, mammalian common names or abbreviations of the proteins are shown after the names or abbreviations for these components in Drosophila.) Other genes specifically associated with InR signaling included PRL-1, encoding a phosphatase that can transform cells [162], the kinase-encoding gene Tak1, and CG9468 and CG5346, which are genes predicted to encode proteins with alpha-mannosidase and iron oxygenase activities, respectively. Genes specifically associated with EGF signaling included EGFR itself, and those encoding several components potentially involved in receptor localization, or downregulation, or both, including Snap, encoding a protein required for vesicular transport, CG7324, encoding a Rab guanosine triphosphatase (GTPase) activating protein, and RSG7, encoding a putative heterotrimeric G protein subunit that also interacts with Snapin, a component of the SNARE complex [60]. Because these genes were associated with EGF signaling but not insulin signaling, this suggests that these are required for EGFR but not InR localization.

2.2.1 An RTK-Ras-ERK interaction network

Many of the previously unknown regulators identified in the RNAi screens may act indirectly through general cellular processes or through multiple levels of transcriptional feedback. Furthermore, RNAi screens suffer from off-target effects even after computational filtering and use of multiple RNAi reagents for each gene [39]. PPI mapping provides an orthogonal representation of network regulators compared to functional genomic approaches because it reveals physical associations. Although large-scale yeast two-hybrid (Y2H) screening can reveal potential PPIs with high accuracy [157], and has been performed on a large scale for MAPK-related proteins [5], Y2H cannot detect interactions that may rely on regulatory posttranslational modifications that occur in endogenous signaling contexts. Large-scale TAP/MS has been used to discover PPIs, most comprehensively in yeast [12, 47, 75] and in human cells in pathway-oriented mapping of tumor necrosis factor (TNF) signaling [11], Wnt signaling [41], and autophagy [7].

We used TAP/MS to capture the dynamic mini-interactome surrounding 15 well-recognized, conserved canonical components of the RTK-Ras-ERK pathway: InR, PDGF (platelet-derived growth factor)- and VEGF (vascular endothelial cell growth factor)-receptor related (PVR), EGFR, the adaptors Drk (Grb2) and Dos (Gab), the GTPase Ras85D, the Ras GTP exchange factor Sos, the cytoplasmic tyrosine kinase Src42A, the GTPase-activating protein Gap1, the phosphatase Csw (Shp2), the MAPK kinase kinase Phl (Raf), the MAPK kinase Dsor1 (MEK), the scaffolds Ksr and Cnk, and the MAPK Rl (ERK). These 15 proteins served as the baits in the affinity purification assay. The proteins and a control were expressed in S2R+ cells using TAP vectors [143] and lysates prepared at baseline (unstimulated cells) or following stimulation with insulin or EGF. Two or more biological replicates were performed for each bait and condition. Interacting proteins were determined by tandem affinity purification and microcapillary liquid chromatography/tandem mass spectrometry (LC/MS/MS). 54,339 peptides were identified representing 12,208 proteins, encompassing an unfiltered network of 5,009 interactions among 1,188 individual proteins ([44]: Table S4). Among the most abundant proteins identified in replicate pull-downs and absent in control preparations were other known RTK-Ras-ERK canonical proteins. A network based on the observed interactions among these canonical proteins recapitulates many of the known RTK-Ras-ERK signaling pathway interactions (Figure 2-1C), validating the sensitivity of our TAP/MS approach in robustly identifying pathway interactors.

Raw TAP/MS data often contain sticky proteins found in control preparations. To provide a ranked list of the most specific pathway interactors by filtering out these sticky proteins, we applied the Significance Analysis of Interactome (SAINT) method to our PPI dataset [12]. Using a SAINT cutoff of 0.83 and false discovery rate (FDR) of 7.2%, we generated a filtered PPI network of 386 interactions among 249 proteins surrounding the canonical components of the RTK-Ras-ERK signaling pathway (Figure 2-2 and [44]: Table S4). We evaluated our PPI network by comparing it with various literature-derived physical interaction networks ([44]: Figure S2A, S2B). For this network comparison, we generated a master physical interaction network (MasterNet) composed of five different types of networks (see Materials and Methods). Our filtered network is significantly overrepresented in the MasterNet, with 29% overlap, compared to 17% for the excluded proteins; the canonical network has a 97% overlap with MasterNet. SAINT scores were highly correlated with appearance in literature datasets, implying that the PPI network as filtered by SAINT represents high-confidence pathway interactors ([44]: Figure S2C). Of the literature-derived networks, appearance in the Drosophila binary PPI dataset most closely correlated with higher SAINT scores ([44]: Figure S2D).

We corroborated selected previously unknown interactions using traditional co-immunoprecipitation techniques and quantitative Western blotting ([44]: Figure S3). Among these, we verified an ERK interaction with the cyclin-dependent kinase cdc2c (CDK2), as reported for mammalian cells [10], implying that ERK can directly regulate the cell cycle through this interaction. Many of the proteins that interacted with multiple RTKs were adaptors ([44]: Table S4). A notable exception was CG10916, which was one of the few common interactors of multiple RTKs (InR, PVR, and EGFR) that was not an adaptor ([44]: Figure S3A, B). Thus, individual RTKs likely recruit distinct complexes during signaling and may compete for a common set of canonical interactors. As a negative regulator of ERK activation and a predicted RING-domain-containing protein, CG10916 may be involved in receptor degradation or downregulation of RTKs. We also found that some interactions below our conservative SAINT threshold of 0.83 could be verified by coimmunoprecipitation ([44]: Figure S3C), suggesting the true size of the network may be larger than the cutoff we chose.

On the basis of GO classifications, we found that the filtered PPI network was enriched in genes encoding regulators of Ras signaling, signaling by the RTKs Sevenless and Torso, and R7 photoreceptor differentiation, all processes known to involve ERK activation, and also those encoding proteins associated with mitosis, the cytoskeleton, axis specification, oogenesis, kinase activity, and SUMO binding ([44]: Table S2). Compared to the total filtered network, proteins interacting with Drk (Grb2) were enriched for GO terms associated with epithelium development and cell fate ($p < 0.02$ for both), but otherwise individual bait networks were representative of the entire network. As with the RNAi hits, our filtered PPI network was enriched for genes conserved in humans and in human diseases ($p < 5.4 \times 10^{-16}$ and $p < 4.6 \times 10^{-3}$, respectively).

Feedback regulation is a mechanism of ensuring pathway robustness [127]. Several studies have examined the transcriptional responses to RTK-Ras-ERK signaling stimulation or perturbation in vivo [3, 38, 64]. We culled genes in these studies responsive to pathway modulation and overlaid them with our PPI dataset. We found that the expression of 25% of the genes for these interactors was changed in response to pathway modulation, a significantly enriched proportion ($p < 2.4 \times 10^{-9}$; [44]: Table S4 and Figure 2-3A). These genes are strong candidates encoding mediators of feedback regulation of RTK-Ras-ERK signaling. Among these were several ribosomal genes (e.g., RpL6 RpL23A, RpL27, RpS18, RpS30) that exhibited reduced expression in response to pathway activation (Figure 2-3A) and that were isolated as negative regulators in the RNAi screens, implying feedback amplification through inhibition of translational repression. These genes also had negatively correlated gene expression with their canonical pathway interactors in published gene expression studies (Figure 2-3B).

During assembly of the RTK-Ras-ERK interactome, we identified complexes under baseline, insulin-, and EGF-stimulated conditions to find pathway interactors and to study the dynamics of complex assembly and disassembly using quantitative label-free proteomics [73]. Previous systematic evaluation of dynamics in interactomes has been limited to individual proteins; for example one study identified dynamic interactors of ERK [144]. Using the SAINT scores at baseline and stimulated conditions, we assembled interactomes of proteins with a high probability of a dynamic interaction with the canonical baits in response to insulin (Figure 2.4.6A) or EGF stimulation (Figure 2.4.6B). We observed several expected interaction dynamics, including the association of subunits of phosphatidylinositol 3-kinase (PI3K) with InR following insulin stimulus, which likely occurs through the adaptor Chico (IRS) and association of the adaptor Drk (Grb2) with EGFR following EGF stimulus ([44]: Table S4). Our global analysis showed that proteins that interacted with the adaptor Dos were more likely to associate than dissociate under insulin stimulus; whereas those that interacted with Drk (Grb2) did not significantly change based on SAINT probabilities. EGFR interactors dissociated when cells were stimulated with insulin. Upon EGF stimulus, interactors with Cnk, Dsor1, Gap1, and Ksr all preferentially dissociated, whereas Phl (Raf) interactors associated (Figure 2.4.6B).

An integrated map of RTK-Ras-ERK signaling We overlaid the functional genomic data from our six systematic RNAi screens for ERK activation with the TAP/MS network structural data (Figure 2-2). Nearly half of the proteins [119] of the filtered PPI network were encoded by genes that scored in the RNAi screens, which represented a significant enrichment over the genome for regulators of this pathway (19%, $p < 7 \times 10^{-25}$) and was an overlap higher than achieved with a more directed RNAi screening of TNFα pathway interactors [11]. Strikingly, 32% (38/119) of the interacting proteins were isolated from RNAi screens in both cell types and following both stimuli (Figure 2.4.6C), whereas if all of the hits from all of the RNAi screens were counted, then only 8% were isolated from both cell types and stimuli.

Together, our RNAi and PPI experiments identified hundreds of previously unknown RTK-Ras-ERK regulators, as well as a core network of genes that were identified with both methods. Because visualization, navigation, and comprehension of complex networks of interacting proteins with functional data can be challenging, we provide our resource of RTK-Ras-ERK interactome and functional genomic data as browseable data files and in Cytoscape format, a graph layout and querying tool [27]. However, given the widespread importance of this pathway and to make the integrated network interactive and widely accessible, we also provide access to the data using the Interaction Map (IM) Browser, an online network visualization tool for interactive, dynamic visualization of PPIs [94]. Because integration of multiple data sources improves the specificity and reliability of individual high-throughput data, we merged our data with the Drosophila Interactions Database (www.droidb.org), which contains previously determined PPIs from Y2H and other studies, a wealth of Drosophila genetic interactions, and predicted conserved interactions, or interologs, from yeast, worms, and humans [90]. Using these tools, the RTK-Ras-ERK network can be searched, filtered, and overlaid with multiple genomic datasets.

Rtf1, TepIV, PPP6 complex, and CG6453 as regulators of ERK activation in vivo Receptor tyrosine kinase signaling to ERKs regulates diverse processes during Drosophila development. Among these, phenotypic alterations in the Drosophila eye and wing are the most easily scored, because Ras activity promotes cell growth, cell proliferation, cell survival, and differentiation into vein tissue downstream of EGFR activity. Because most of our newly identified pathway-associated genes do not have known alleles, we tested for phenotypes by expressing RNAi hairpins in Drosophila, which can faithfully recapitulate known phenotypes [92, 32]. We tested for phenotypes of multiple genes isolated in our screens by expressing hairpins from a library created for transgenic RNAi, or in a few cases by cDNA overexpression, in the developing wing disc (Figure 2-5, [44]: Figure S4, Table 2.1, and [44]: Table S5). Of the 84 genes tested, 48 (57%) had a phenotype in the wing. Consistent with systematic PPI analyses in yeast [157], we found that proteins with a high degree (hubs) in MasterNet were no more likely than proteins with a lower degree to result in a wing phenotype.

Surprisingly, we found that even genes that were identified both in RNAi screens and in the PPI interaction network were no more likely than genes isolated from each individually to score in wing phenotypes. One of the genes that was positive in both the functional genomic screen and the interaction screen was CG6453, which encodes a noncatalytic subunit of glucosidase II. The interaction between the CG6453 protein with Raf had a high SAINT score and coimmunoprecipitation experiments confirmed this interaction ([44]: Figure S3A). In the S2R+ EGFR RNAi screen, this gene was a negative regulator and we demonstrated that its depletion by RNAi resulted in a growth and patterning defect (ectopic wing vein material) in the wing, which is consistent with negative regulation of the pathway (Figure 2-5A). Although genes encoding TepIV, the Drosophila homolog of a glycophosphatidylinositol-linked protein that is mutated in human cancers, and components of the protein phosphatase PPP6 complex, its catalytic subunit PpV and regulatory subunit CG10289, were not found in the RNAi screens, these proteins were positive in the interaction screen. We confirmed their interactions with pathway components by coimmunoprecipitation ([44]: Figure S3A) and demonstrated that their knock down produced in vivo phenotypes (Figure 2-5B and C). TepIV interacted with Ksr and, despite not scoring in our RNAi screens and having a weak RNAi phenotype in cells, nevertheless modified the RasN17 phenotype, consistent with a role as a positive regulator (Figure 2-5B). PpV and CG10289 interacted with each other and Raf, and PpV depletion resulted in a growth defect in the wing (Figure 2-5C). Finally, Rtf1, a histone methyltransferase, was a weak interactor with multiple pathway components and was filtered out of the final PPI network because of its SAINT score. However, the gene encoding this protein was identified as a negative regulator in our RNAi screens and we confirmed an in vivo phenotype associated with increased dpERK (indicating increased activity) in the wing (Figure 2-5D), showing that Rtf1 is a bona fide regulator of ERK activation.

2.3 Discussion

Dissection of oncogenic signaling pathways using functional genomics and proteomics approaches facilitates understanding dynamic information processing and how these pathways may be disrupted by mutations or targeted therapeutically [73]. By combining multiple, parallel genome-wide RNAi screens and TAP/MS interactome screens, we have assembled an integrated network of RTK-Ras-ERK signaling with both PPI interactions and functional information obtained in the same signaling environment. This network provides a resource for subsequent hypothesis-driven, mechanistic investigation of hundreds of conserved regulators. Because high-throughput datasets are individually susceptible to multiple sources of technical and biological noise, confidence in subsets of any given omics dataset can be increased by overlapping contrasting experimental approaches. Most integrative efforts up to now have queried datasets generated under disparate conditions and even different organisms. We found that only a small fraction of the hits from interactome or functional screening were isolated under all conditions tested, and most of these represented known canonical pathway components. Many of the hits that were identified from each method individually also showed evidence of activity in vivo. Comparing our studies to other studies of MAPK regulators suggests that the complete landscape of proteins regulating RTK-Ras-ERK signaling under specific conditions is likely to be larger than the conservative overlapping network that we describe. In comparison to a Y2H screen for MAPK pathway interactors, where > 600 interactions were identified [5], only 54 proteins overlapped with our network, 30 (56%) of which also were positive in our RNAi screen, including the proton transporter ATPsyn-beta (ATP5B), which was a negative regulator in our RNAi screens. Of the 31 proteins from a study of dynamic ERK interactors that overlapped with our filtered dataset [144], 22 were encoded by genes positive in our RNAi screens, but only one, heat shock protein 60 (HspD1), was pulled down by ERK itself in our study. However, another 16 proteins interacted with Raf and 8 interacted with Dsor (MEK). By considering the Raf-MEK-ERK cassette as a whole, the number of overlapping interactions increased. Although these comparisons are limited by the differences in Y2H and TAP/MS techniques, the population of regulators that can be identified is probably highly technique- and condition-specific, and this work should be seen as a first pass at identifying the universe of proteins regulating the output of this pathway. We used PPI mapping and functional genomic methods to identify several previously unknown regulators that also exhibited in vivo roles in RTK-Ras-ERK signaling. Translation of cell culture regulators to in vivo phenotypes is challenging due to lack of knowledge of the correct tissue in which to test for activity. Because many of the newly identified regulators are likely cell-type and RTK-specific, we were unable to identify phenotypes in the wing disc for many of these regulators. A large number of genes positive in the RNAi screens was not identified in the PPI network, either due to false negatives, or because the encoding proteins modulate activity of the pathway indirectly. A prime example of this latter category is Rtf1, a histone methyltransferase knock down of which enhanced ERK activation in vivo. Rtf1 enhances Notch pathway activity [136] and the Notch pathway can inhibit ERK activity [156], and thus Rtf1 may be a key mediator of Notch-ERK crosstalk. In contrast, we identified another protein phosphatase 2A (PP2A) family member, the PPP6 ortholog PpV and its regulatory subunit CG10289, as interacting with Raf, but did not identify the genes encoding these proteins in our RNAi screens. In mammals, PPP6 components can interact with the inhibitor of nuclear factor B IB [11, 126] and regulate the cell cycle in normal and pathological contexts. The role of the Ser/Thr phosphatase PP2A in the Ras pathway has been principally described as a positive regulator through dephosphorylation of Ser259 on Raf and Ser392 on Ksr (numbering is based on human proteins), inducing 14-3-3 protein dissociation [93]; PPP6 may play a similar role in Raf activation in specific in vivo contexts. Interestingly, CG6453, a noncatalytic subunit of glucosidase II, was identified in the interaction screen and was identified in the RNAi screens, indicating a high-confidence interactor. Although its mechanism of regulating MAPK output remains unknown, it is consistent with the growing recognition that metabolic and other genes previously thought to have housekeeping roles, in fact, can have specific functions in signaling [145, 26]. Finally, despite its interaction with intracellular Ksr, TepIV has homology with CD109, a GPI-linked cell surface marker of T cells, endothelial cells, and activated platelets that contains a protease inhibitor 2 macroglobulin domain [81]; CD109 is mutated in 7% of colorectal cancers [121] and may thus affect ERK output in these cancers. As more human cancers are characterized through ongoing large-scale next-generation sequencing, our dataset of regulators of RTK-Ras-ERK signaling will provide a resource for understanding the potential mechanistic contribution of somatic mutations to cancer development.

2.4 Materials and Methods

2.4.1 RNAi screening

Primary screening procedures were performed as published previously [16, 42]. We derived a S2R+ cell line expressing DER (EGFR) from a metallothionein promoter (S2R+mtDER) also expressing cyan fluorescent protein (CFP)-tagged Dsor1 (MEK) and yellow fluorescent protein (YFP)-tagged Rl (ERK) [16]. We confirmed ERK activation following secreted Spitz (sSpitz) (EGF in mammals) stimulus of both endogenous and tagged ERK by Western blotting and high-throughput format, and confirmed assay sensitivity using dsRNAs targeting canonical components of the RTK-Ras-ERK pathway. For primary screening in Kc167 cells, we used our previously described cell line Kc167 expressing DER (EGFR) from a metallothionein promoter (Kc mtDER) [16] and modified the high-throughput assay by using our Alexa647-conjugated dpERK antibody normalized to DAPI staining of nuclei to quantify ERK activity. Cells were stimulated with conditioned media containing sSpitz for 10 min or 30 min. Secondary screens were performed as described [42] using S2R+ and Kc cell lines with 25g/mL insulin or sSpitz-containing conditioned media. Briefly, cell lines were seeded in plates pre-populated with resynthesized dsRNA amplicons identified from the primary screen as InR- or EGFR-specific. Following stimulation, cells were fixed and stained for dpERK as previously described. Primary screen hits were pre-filtered for computationally-predicted off-target effects, which is generally sufficient to reduce off-target noise to below assay noise [16]; however, any individual dsRNA should be treated with caution until validated with multiple amplicons [33]. A Z-score threshold of +/-1.5 was used as the primary screen cutoff, and is an average of replicate screens under each condition. Full datasets and dsRNA sequence information are available at the DRSC website (www.flyrnai.org).

2.4.2 TAP and mass spectrometry

TAP expression vectors permitting low-level expression of tagged components in stable Drosophila cell lines using the metallothionein promoter have been previously described [143]. For the bait proteins, we cloned InR, PVR, EGFR, Drk (Grb2), Dos (Gab), Sos, Src42A, Gap1, Csw (Shp2), Ras85D, Phl (Raf), Dsor1 (MEK), Ksr, Cnk, and Rl (ERK) into the C-terminal tag TAP vector and created stable cell lines for each, as well as a control cell line for subtracting nonspecific interactors or contaminants. All cell lines except InR-TAP also expressed EGFR from an uninduced metallothionein promoter (resulting in minimal low level expression) for induction with sSpitz (EGF). 1 to 2 x 109 cells induced with 140 M CuSO4 overnight were used for each lysis at the given condition. Cells were lysed as described [143] and in-solution TAP was performed essentially as described [120], with the exception of final washes and elution, which was performed in ammonium bicarbonate buffer without detergent for LC-MS/MS analysis. At least two biological replicates were performed for each bait and condition. Several micrograms of TAP immunoprecipitation from each bait condition were reduced with 10 mM DTT at 55C, alkylated with 55 mM iodoacetamide at room temperature, and then digested overnight with 2.5 g of modified trypsin (Promega) at pH 8.3 (50 mM ammonium bicarbonate) in a total of 200 L. The digest was stopped with 5% trifluoroacetic acid (TFA) and cleaned of buffer and debris using a C18 ZipTip (Millipore). 35 L of aqueous HPLC A buffer was added to the C18 Ziptip elution (50% acetonitrile/0.1% TFA) was dried to 10 L to concentrate the sample and remove organic content. A 5 L aliquot was injected onto the microcapillary LC/MS/MS system for sequencing. The microcapillary LC/MS/MS setup consisted of a 75 m id x 10 cm length microcapillary column (New Objective Inc., Woburn, MA) self-packed with Magic C18 (Michrom Bioresources, Auburn, CA) operated at a flow-rate of 300 nL/min using a splitless EASY-nLC (Thermo Fisher Scientific). The HPLC gradient was 3% B to 38% B over 60 minutes followed by a 7 minute wash at 95% B. The column was pre-equilibrated with A buffer for 15 minutes at 0% B prior to the runs (A: 99% water/0.9% acetonitrile/0.1% acetic acid;

B: 99% acetonitrile/0.9% water/0.1% acetic acid). The microcapillary LC system is coupled directly to a LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, San Jose, CA) operated in positive ion mode for data dependent acquisitions (DDA) (Top 5: 1 FT survey scan followed by five scans of peptide fragmentation (MS/MS) in the ion trap by collision-induced dissociation (CID) using helium gas. The spray tip voltage was 2.8 kV and capillary voltage was 35 V. A single micro scan with a maximum inject time of 400 msec was used for the FT-MS scan in the Orbitrap and 110 msec was used for the MS/MS scans in the ion trap. Typically, between 3000-6000 MS/MS spectra were collected per run. The total number of LC/MS/MS runs collected for this study was 94 and collected over a six-month period. All LC/MS/MS runs were separated by at least one blank run to prevent column carryover. Raw MS/MS spectra are available by request and are deposited in TRANCHE.

All collected MS/MS fragmentation spectra were searched against the reversed dmel-all-translation protein database (FlyBase Consortium,) version 5.4 (41,644 protein entries, Jan, 2008) using the Sequest search engine in Proteomics Browser Software (Thermo Scientific, San Jose, CA). Differential posttranslational modifications including deamidation of QN (glutamine and asparagine) (+0.989 Da) and oxidation of methionine (+15.9949 Da), common in vitro modifications that occur during sample processing were included in the database searches. From Sequest, protein groups containing at least two unique identified peptides were initially accepted if they were top ranked matches against the forward (target) dmel-all-translation protein database and with a consensus score of greater or equal to 1.0. Individual peptides that were not part of protein groups were accepted if they matched the target database and passed the following stringent Sequest scoring thresholds: 1+ ions, Xcorr $\geq$ 1.9, Sf $\geq$ 0.75, P $\geq$ 1; 2+ ions, Xcorr $\geq$ 2.0, Sf $\geq$ 0.75, P $\geq$ 1; 3+ ions, Xcorr $\geq$ 2.55, Sf $\geq$ 0.75, P $\geq$ 1. After passing the initial scoring thresholds, all peptide hits not contained in protein groups were then manually inspected to be sure that all b- (fragment ions resulting from amide bond breaks from the peptides N-terminus) and y- ions (fragment ions resulting from amide bond breaks from the peptides C-terminus) aligned with the assigned sequence using tools (FuzzyIons and GraphMod) in Proteomics Browser

Software (Thermo Fisher Scientific). A FDR rate of 1.84% for peptide hits and FDR of 0.6% for protein hits was calculated based on the number of reversed database hits above the scoring thresholds.

2.4.3 Computational analysis of TAP-MS data

We used the “significance analysis of interactome” (SAINT) algorithm to calculate probability scores for interactions observed by MS. SAINT uses spectral count data and constructs separate distributions for true and false interactions to derive the probability of a bona fide protein-protein interaction. Because SAINT models spectral counts with a unimodal distribution, we ran the algorithm separately for each condition and combined the scores. Specifically, we assumed that each condition was conditionally independent given the spectral count data and computed the probability that the interaction was true in any condition. For proteins A and B in conditions 1 to n the combined score is computed as:

$$\begin{aligned} P(A \leftrightarrow B \text{ any cond}) &= 1 - P(A \leftrightarrow \text{ no cond}) \\ &= 1 - (1 - P(A \leftrightarrow B \text{ cond } 1)) \cdots \\ &\quad (1 - P(A \leftrightarrow B \text{ cond } n)), \end{aligned}$$

where $P(A \leftrightarrow B \text{ cond } i)$ is the SAINT score for condition i. Some proteins were not used as baits in all conditions, hence some interactions that were observed in one condition could not be observed in another. In this case, we used the prior probability of an interaction occurring in that condition as computed by the SAINT algorithm. In the general setting, this would be the probability that a randomly chosen pair of proteins interact, in other words (#interacting pairs of proteins)/(#pairs of proteins). In our specific case, we are choosing a pair of proteins from proteins that are observable in mass spectrometry, so we adjust the ratio accordingly to our specific setting.

Additionally, we computed pairwise dynamic difference scores between conditions (the probability that an interaction is true in one condition but not the other) assuming the conditions were conditionally independent given the spectral count data. To determine a high-confidence threshold, we constructed a set of true positive interactions by overlapping our experimental interactions with BioGRID. This list contained 49 interactions between 114 proteins. We formed a true negative set by taking interactions that were more than 3 hops away in the BioGRID protein interaction network. A receiver operating characteristic (ROC) curve generated using this gold standard list and generated using Fly binary and fly complex data is shown in [44]: Figure S2A and B. We chose 0.83 as the cutoff to achieve a 7.2% FPR and 26.5% true positive rate, which is comparable to the results achieved in [12].

2.4.4 Additional statistical analysis

Filtered binary interactions were graphed using the Cytoscape environment [27]. For analysis of feedback regulation, three in vivo microarray studies were collated [3, 38, 64]. Microarray data from in vivo analysis of mesoderm [38] were reanalyzed to focus on subgroups for RTK-Ras-ERK pathway only, excluding other pathway datasets.

Human orthologs were predicted using DIOPT, an integrative ortholog prediction tool developed at Drosophila RNAi Screening Center [58] `http://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl`. The orthologs with the best prediction score, reflecting the number of methods from which the prediction was identified, were selected. Potential human disease-related fly homolog information was obtained from Homophila vs 2.1 [22]. Gene expression levels were obtained from DRSC (`http://www.flyrnai.org/cgi-bin/RNAi_expression_levels.pl` and cell line gene expression data was obtained from the modENCODE project [17]. The significance of conserved genes, expressed genes, or disease-related genes was tested by calculating cumulative hypergeometric probability. The enrichment of GO annotations for Molecular Function and Biological Process, as well as Panther pathway annotation, was performed using online DAVID tool (`http://david.abcc.ncifcrf.gov/` [30]. Hierarchical clustering and graphing was performed using the MultiExperiment Viewer, Cluster, and Java TreeView programs [108, 110, 36].

MasterNet is a compilation of databases. (i) Fly binary PPI network: This network was constructed by integrating experimentally identified binary PPIs (direct physical interactions) from major PPI databases, such as BioGrid [125], IntAct [2], MINT [18], DIP [111]), and DroID [90]. The fly binary PPI network consists of 29,325 interactions between 8,161 proteins. The PPIs were downloaded from the source databases in PSI-MI format [70] and the gene/protein identifiers were mapped to FlyBase gene identifiers. (ii) Interolog binary PPI network: PPIs were predicted on the basis of experimentally identified binary PPIs for human, mouse, worm, and yeast. (iii) Interolog protein complexes network: PPIs were predicted from experimentally identified protein complexes for human, mouse, worm, and yeast. Both the interolog networks were compiled from BioGrid, IntAct, MINT, DIP, and HPRD [99] databases. The PSI-MI files were downloaded from the source databases and the experimental identifier from interaction detection type field was used to sort the PPI as either binary or complex. Using ortholog annotation from DIOPT database 129,090 PPIs between 5,954 proteins were mapped to fly. (iv) Kinase-substrate network: For each experimentally verified phosphorylation site, the kinase that phosphorylates that site was predicted using the NetPhorest program [89, 135]. The program uses probabilistic sequence models of linear motifs to predict kinase-substrate relationship. The fly kinase-substrate network consists of 26,736 interactions between 55 kinases and 2,518 substrate proteins. (v) Domain-domain interaction network: Known and predicted protein domain-domain interactions (DDI) were extracted from DOMINE database [155], which includes 26,219 interactions inferred from Protein Databank (PDB, www.pdb.org) entries and those that are predicted by 13 different computational approaches using Pfam domain definitions. For network integration, we considered only high-confidence DDIs as defined by DOMINE and those derived from crystal structures.

2.4.5 Western blotting and coimmunoprecipitation

All Western blotting and co-immunoprecipitation procedures and antibodies used were previously described [16]. Quantification of dpERK and total ERK (used as normalization value) was performed using the LiCor detection system. Western blotting and coimmunoprecipitation experiments were performed a minimum of two times.

2.4.6 In vivo analysis

Stocks used for genetic analysis were obtained from Bloomington except where noted. All HA-tagged cDNA constructs were cloned by PCR cloning using Phusion Polymerase (New England Biolabs) into pUAST. cDNA clones or libraries used as templates were as follows: Dco (LD04938), CG31666 (SD04616), Rack1 (RE74715), CG1884 (cDNA library), and CG31302 (AT04807). Hairpins described in the text were cloned into pWiz as described previously [118] using the following primers:

- CG7282: CACGCCCAGCTGTCAG, TTCACGTTCTCCAGTTTCTC
- CG3878: CAGCTCCGCAGTGCTCGTGT, AGTTGTCGTCGTCGGAGCTC
- CG1884: TCGGCTTGGGCACAAAC, AAGGACTTCGCCCTGGAT
- CG17665: GCAGAAGCAATAGCCGAATC, ATTTTCTCATCTGCCGCATC.

Other RNAi hairpins were designed using the attP targeted transgenic system for an in vivo RNAi project ("TRiP" lines) as described [92], as well as RNAi lines from Vienna Drosophila RNAi Center and NIG-Fly Japan stock center. Other fly lines: y,w,hsFlp, MS1096-Gal4, UAS Ras1N17, ElpB1/CyO, apterous-Gal4, UAS-mCD8-GFP/CyO. For dpERK staining, wing discs from third instar larvae were dissected in cold PBS, fixed for 15 minutes in 4% formaldehyde, and washed in PBS+0.1%Triton. Discs were stained using a rabbit antibody that recognizes dpERK (Cell Signaling). Wings of the indicated genotype were mounted in a 1:1 mixture of Permount and xylenes. A complete list of the hairpin lines used in this study is given in table S6.

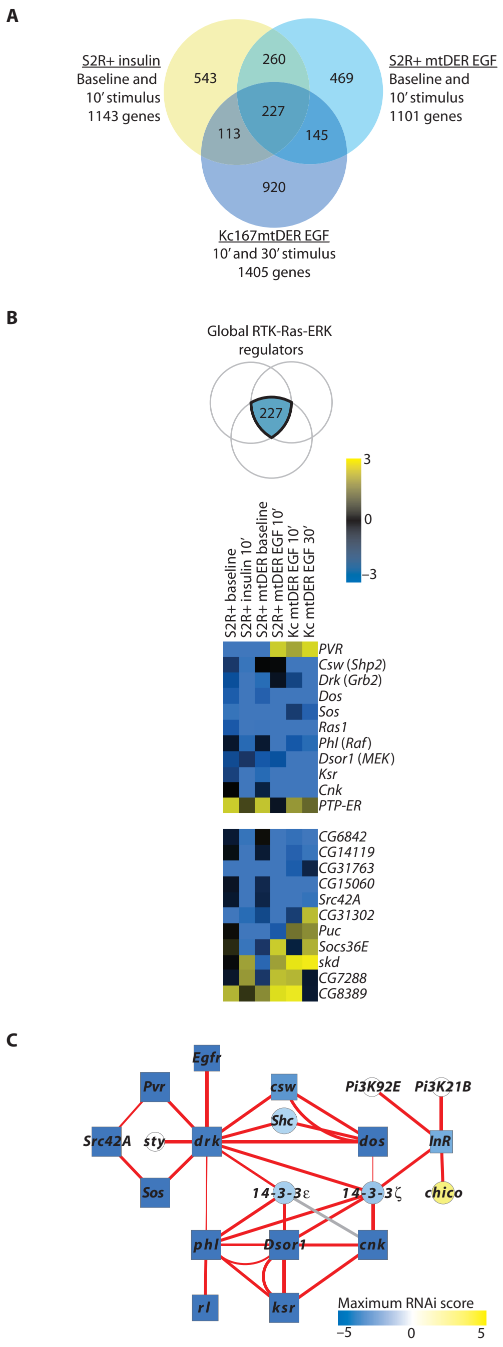


Figure 2-1: (Continued on the following page.)

Figure 2-1: Parallel RTK-Ras-ERK genome-wide RNAi screens in Drosophila. (A) Comparison of six RNAi screens grouped into three experimental categories based on ligand stimulus or cell type, with number of annotated genes in each category. Note S2R+ cells used for EGF stimulation express EGFR for robust ERK activation in response to EGF; thus a baseline RNAi screen was performed in these cells in the absence of EGF in addition to the baseline performed previously in the EGFR-negative S2R+ cells. The total gene set is enriched for genes with human orthologs and for known components of the canonical signaling cassette. (B) 227 genes appeared in all three groups, representing a common or global set of RTK-Ras-ERK regulators, which included those encoding proteins in the canonical cascade (top array graph). Examples of other global regulators are shown in the lower array graph. Genes are listed with by common abbreviation, with mammalian names listed in parentheses when different than Drosophila. Color represents average Z-score in each primary RNAi screen. (C) Many known canonical interactions are recapitulated by the TAP/MS analysis, including those involving adaptors (Drk-Sos), the Phl activation complex (Phl-Dsor1, Phl-Rl, Phl-Ksr, and Ksr-Cnk), and InR with the PI3K subunits p110 (PI3K92E) and p60 (p85 ortholog, PI3K21B). Red edges denote those found both in our study and in MasterNet, a literature-based compilation of previously known PPIs. The gray edge denotes those not found in MasterNet. Edge thickness represents SAINT score. Circles represent prey; rectangles represent baits.

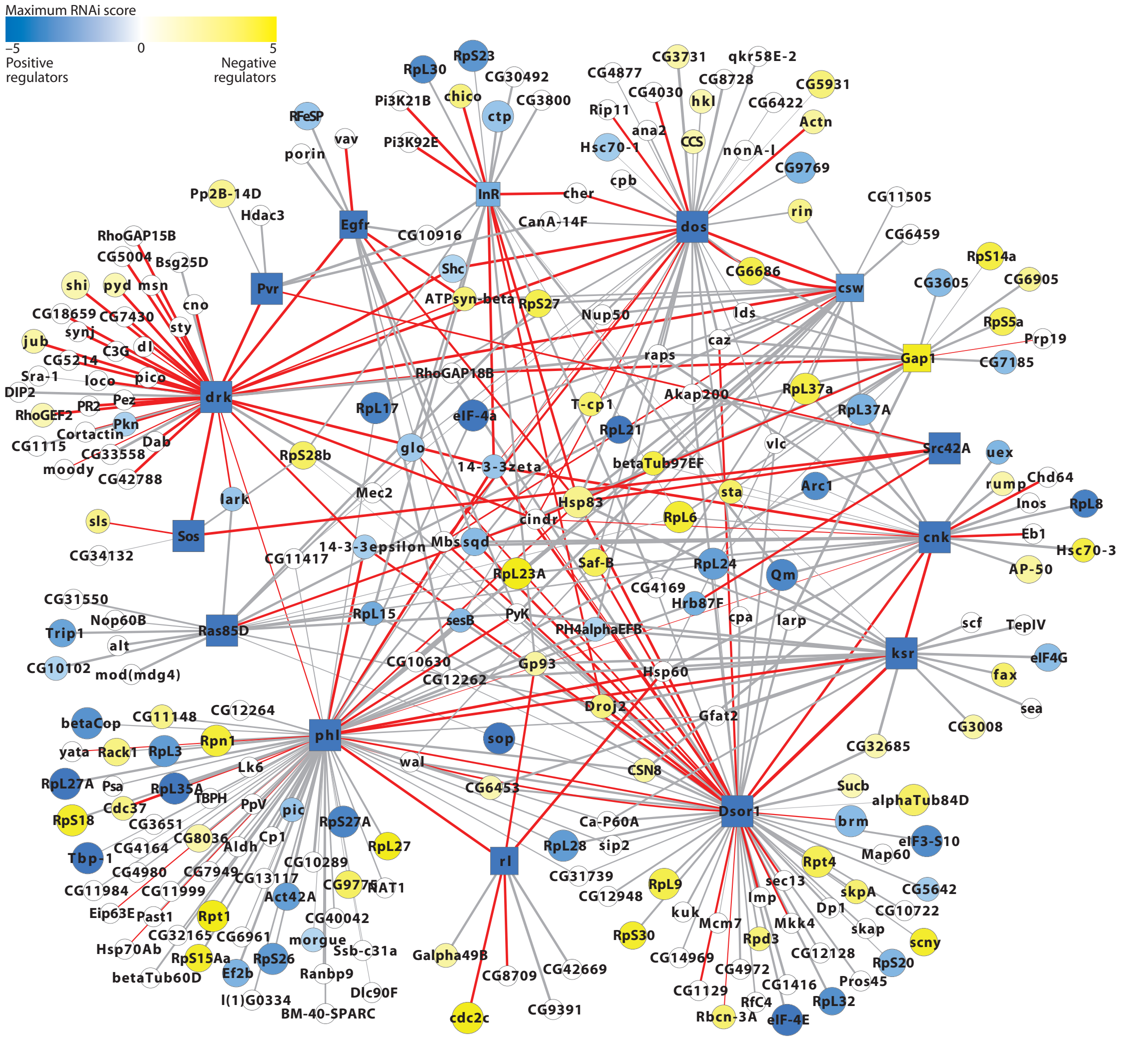


Figure 2-2: TAP/MS PPI RTK-Ras-ERK signaling network. Filtered PPI map of RTK-Ras-ERK signaling in Drosophila, including primary RNAi screen scores, if present. Z-score RNAi result describes negative regulators (yellow) and positive regulators (blue). Edge thickness denotes SAINT score, a measure of interaction confidence. Red edges denote those found both in our study and in MasterNet, a literature-based compilation of previously known PPIs. Edge thickness represents SAINT score. Circles represent prey; rectangles represent baits. The size of the node correlates with the number of RNAi screens from which the proteins were isolated. See [44]: Table S4 for details of all node and edge parameters and names of the proteins identified as pathway interactors.

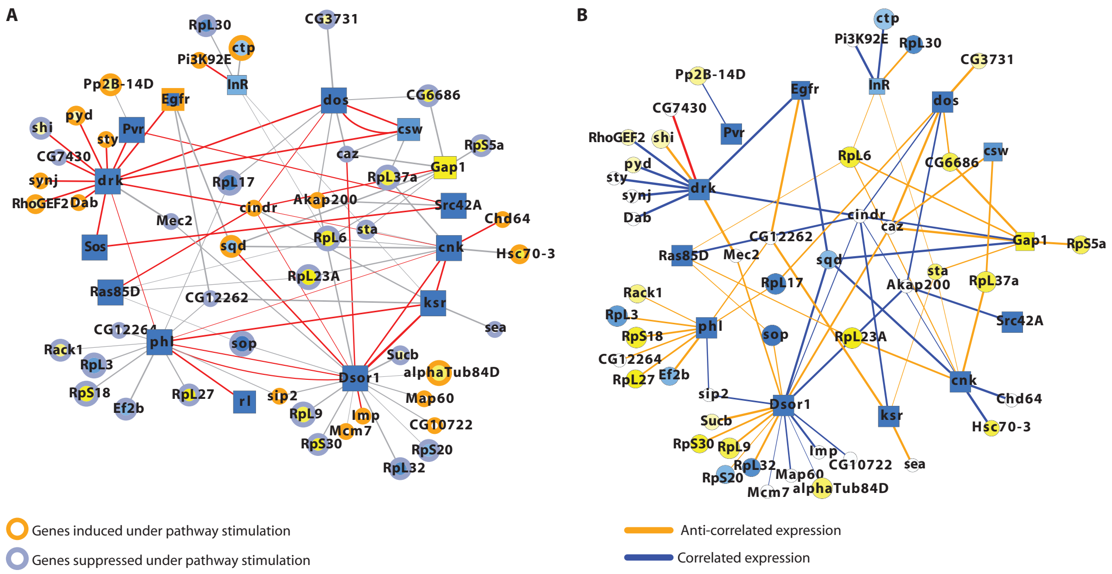


Figure 2-3: Additional analysis of the PPI network. (A) Nodes in the PPI network that were also regulated by pathway output, as mined from in vivo transcriptome analyses. Blue nodes were encoded by genes that were downregulated by pathway output; yellow nodes were upregulated. (B) Correlation between expression of the genes encoding the baits and preys. Orange edges denote interacting partners that exhibited an inverse correlation in expression; blue edges denote interacting partners that exhibited positive correlation in expression.

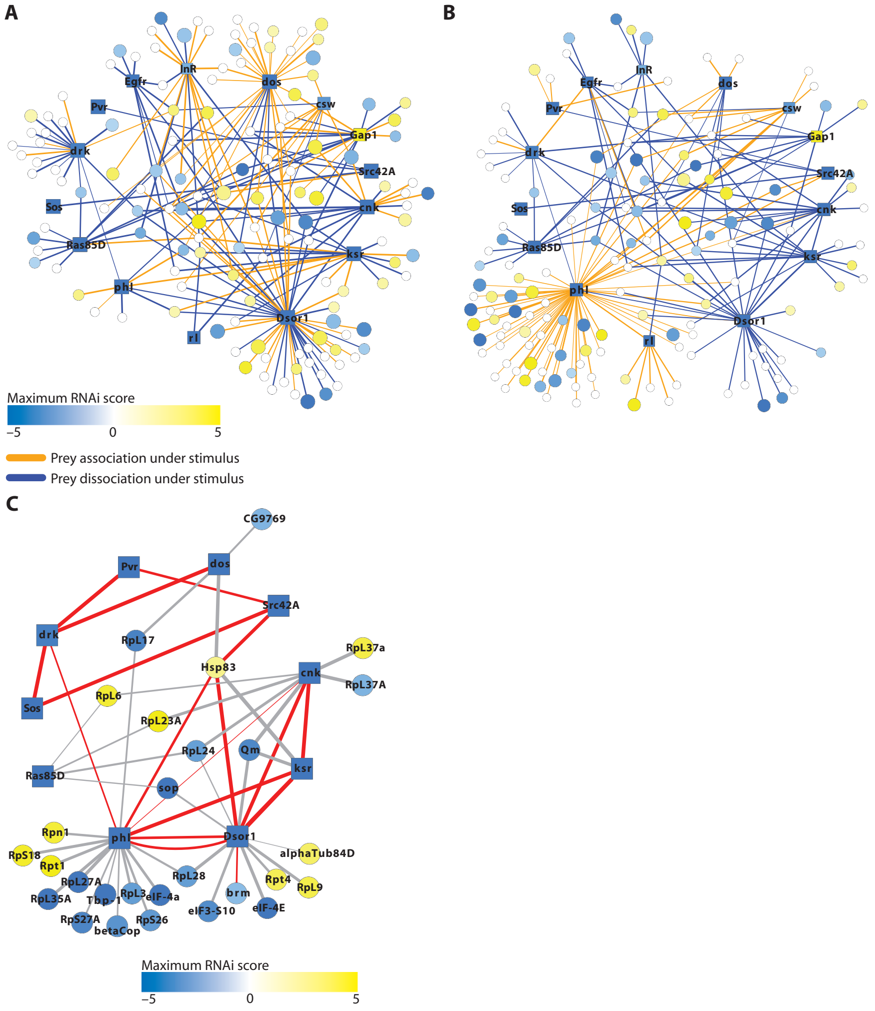


Figure 2-4: Dynamics in the RTK-Ras-ERK signaling network. (A) Subset of total PPI network with a SAINT-based probability of dynamic interaction > 0.8 by comparing baseline to insulin condition. (B) Subset of PPI network with dynamic interactions under EGF stimulus. In both panels (A and B), orange lines indicate protein association; blue lines denote dissociation. (C) "Core" network of PPIs that were identified in all three RNAi screen sets. Edge thickness denotes SAINT score, a measure of interaction confidence. Red edges denote those found both in our study and in MasterNet, a literature-based compilation of previously known PPIs. Edge thickness represents SAINT score. Circles represent prey; rectangles represent baits. The size of the node correlates with the number of RNAi screens from which the proteins were isolated. See [44]: Table S4 for details of all node and edge parameters and names of the proteins identified as pathway interactors.

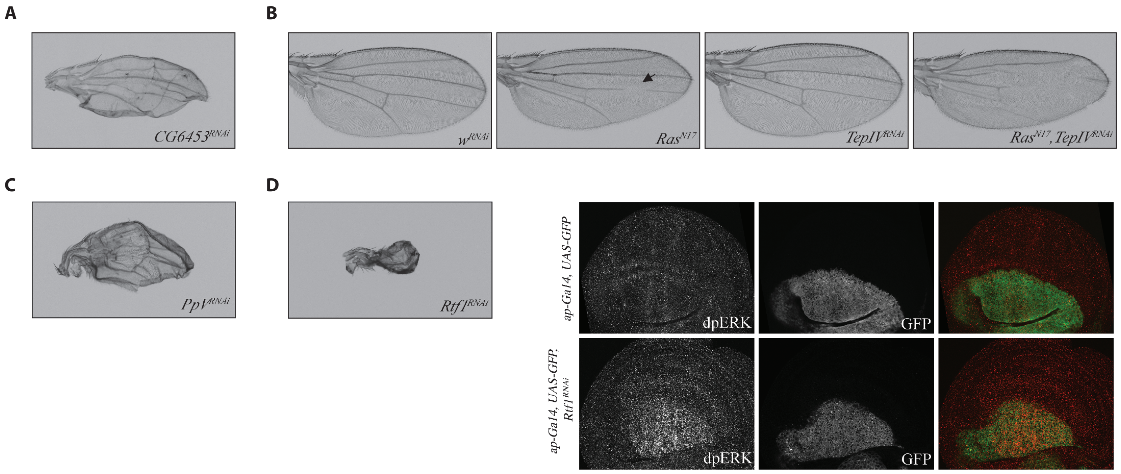


Figure 2-5: In vivo analysis of newly identified regulators of RTK-Ras-ERK signaling in Drosophila. (A) Knock down of CG6453 results in a growth defect in the wing. CG6453-Phl interaction was isolated with a SAINT score of 0.99, and CG6453 had a Z-score of 1.6 under EGF stimulus in S2R+ cells. (B) Knock down of TepIV enhances the RasN17 loss of wing vein phenotype consistent with a role as a positive regulator. TepIV had a SAINT probability of 1.0. (C) Knock down of the PPP6 subunit PpV results in a growth defect in the wing. PpV had a SAINT score of 0.88 with Phl. (D) Knock down of Rtf1 results in a growth defect in the wing and induces ectopic dpERK staining in the wing disc. Rtf had a Z-score of 2.25 under insulin stimulus in S2R+ cells.

Symbol	Name	Wing Phenotype	Maximum SAINT Probability	Baseline S2R+	10’ Insulin S2R+	Baseline S2R+mtDER	10’ EGF S2R+mtDER	10’ EGF KcmtDER	30’ EGF KcmtDER	Screen
uex	unextended	slight curling	1	0.32	-0.47	1.02	-2.7	0.1	0.06	E
CG6453	-	slight curling	0.99	-0.94	0.94	0.51	1.6	0.85	0.28	E
brm	brahma	wing size reduced; slight curling; wing shape abnormal	0.89	-0.97	-2.46	-2.09	0.36	0.61	2.42	IEK
betaCop	beta-coatomer protein	nearly complete loss of wing tissue	0.88	0.15	2.08	2.85	-3.86	0.27	-1.74	IEK
TepIV	Thiolester containing protein IV	Enhancement of Ras[N17] wing vein phenotype	1							
PpV	Protein phosphatase V	nearly complete loss of wing tissue	0.88							
Dref	DNA replication-related element factor	nearly complete loss of wing tissue	0.78	0.2	-0.38	-0.6	-4.34	-0.76	0.11	E
kis	kismet	ecoptic wing vein material	0.56	1.08	-1.97	0.71	0.46	0.32	2.56	IK
CG6907	-	Enhancement of Ras[N17] wing vein phenotype	0.56	0.64	-5.67	-1.37	-3.82	-1.26	0.94	IE
ACC	Acetyl-CoA carboxylase	”severely blistered, misshapen, and reduced wing size”	0.56	-1.45	-2.89	-1.81	2.02	0.42	-0.58	IE
Chro	Chromator	”severely blistered, misshapen, and reduced wing size”	0.56	-1.2	-1.34	-1.51	-1.26	2.16	0.18	EK
CG3523	-	slight curling	0.56	-0.28	2.1	1.07	-3.69	0.2	-0.96	IE

cact	cactus	wing size reduced; ecoptic and missing vein material; Null allele suppresses Elp[B1] phenotype in eye; Null allele suppresses tor[4021] phenotype in embryo	0.56	-1.5	-2.45	-0.73	-1.85	0.22	0.29	IE
Rtf1	Rtf1	”severely blistered, misshapen, and reduced wing size; enhanced dpERK staining”	0.56	-0.95	2.25	1.22	0.45	0.2	1.37	I
sl	small wing	Ectopic wing veins with MS1096	0.53	1.61	1.19	-0.43	0.72	0.76	-1.03	I
dgt2	dim gamma-tubulin 2	Enhancement of Ras[N17] wing vein phenotype	0.53	4.39	0.15	1.27	0.46	-0.52	-0.54	I
Caf1	Chromatin assembly factor 1 subunit	”blistered, misshapen, curled, and reduced wing size”	0	-2.29	-2.72	-0.34	-0.63	0.09	0.1	I
CG8963	-	ecoptic wing vein material	0	-0.18	-3.01	-0.67	-3.52	0.02	1.61	IEK
beta’Cop	beta’-coatomer protein	nearly complete loss of wing tissue	0	-0.54	2.2	5.49	-3.14	-0.29	-0.72	IE
hyx	hyrax	nearly complete loss of wing tissue	0	3.94	2.94	2.48	0.62	2.16	1.02	IEK
deltaCOP	delta-coatomer protein	”severely blistered, misshapen, and reduced wing size”	0	0.18	1.69	1.3	-2.91	-0.31	-1.57	IEK

alpha-Tub85E	alpha-Tubulin at 85E	”severely blistered, misshapen, and reduced wing size”	0	2.74	-2.83	1.88	-2	-0.21	0.14	IE
Arc42	Arc42	slight curling	0	1.82	-0.13	1.44	0.17	-0.59	-0.42	I
chinmo	Chronologically inappropriate morphogenesis	wing size reduced; blistering; upward curling; cDNA overexpression lethal with all drivers tested	0	-0.44	-3.46	-2.01	-1.77	0.49	-0.03	IE
CG34422	-	”wing size reduced; slight curling; pWiz second hairpin also results in smaller, rough wings”	0	-2.05	0.65	-1.99	0.19	0.77	1.44	IE
CG5844	-	wing size reduced; loss of wing vein material	0	-3.41	-4.65	-1.71	-6.92	-1.92	-0.54	IEK
CG6854	-	possible ectopic wing vein material; Enhancement of Ras[N17] wing vein phenotype	N/A	2.54	-0.81	-1.48	-2.84	0.74	1.7	IEK
Axn	Axin	possible ectopic wing vein material	N/A	2.75	-2.2	-0.29	0.41	2.56	1.81	IK
CG13298	-	”blistered, curled, and reduced wing size; ectopic wing vein material”	N/A	6.73	1.42	1.06	0.38	-1.19	-0.71	I

Spc105R	Spc105-related	”blistered, misshapen, curled, and reduced wing size”	N/A	-3.2	-1.71	-0.33	-3.1	0.83	0.21	IE
Spn42Dc	Serpin 42Dc	ecoptic wing vein material; Suppression of Ras[N17] wing vein phenotype	N/A	4.67	1.26	1.7	0.48	-0.43	0.02	IE
Cdc27	Cdc27	misshapen and reduced wing size	N/A	3.11	-4.07	6.12	-6.57	-1.6	-2.16	IEK
wah	waharan	”much smaller wing with loss of wing vein material; Hairpin wing phenotype enhanced by DER[DN] (dominant negative); Hairpin expression in eye results in rough, smaller eyes”	N/A	-0.88	1.42	-0.61	-7.11	-1.86	-0.12	EK
Cap-G	-	nearly complete loss of wing tissue	N/A	3.25	0.57	1.72	0.64	-0.18	1.07	IE
CG6984	-	Enhancement of Ras[N17] wing vein phenotype	N/A	6.42	-0.83	-0.34	0.69	-0.27	-0.3	I
CG43073	-	Overexpression in wing results in expanded A-P axis and ectopic wing veins	N/A	-3.79	-2.58	-1.65	-6.08	-1.39	2.23	IEK

dco	discs overgrown	”Overexpression in wing results in expanded A-P axis and thickened wing veins, and suppresses RasN17 (dominant negative) wing vein phenotype. Overexpression in the eye produces a rough eye phenotype.”	N/A	1.99	1.03	0.86	0.58	-0.19	0.22	I
Not1	Not1	”severely blistered, misshapen, and reduced wing size; pWiz second hairpin expression in wing also results in loss of most tissue; Overexpression in wing results in ectopic wing veins; Hairpin expression in eye results in rough, smaller eye”	N/A	1.61	-1.97	-3.28	-4.41	2.66	2.31	IEK
ago	archipelago	slight curling; Enhancement of Ras[N17] wing vein phenotype	N/A	0.71	4.15	1.47	1.47	0.13	-0.53	I
fzy	fizzy	slight curling; Suppression of Ras[N17] wing vein phenotype	N/A	-0.76	-3.97	-1.47	-4.98	-0.51	1.24	IE

Vha100-5	Vacuolar H[+] ATPase subunit 100-5	slight curling	N/A	1.64	-1.12	-1.29	-2.07	1.53	0.08	IEK
CG6608	-	slight curling	N/A	-0.25	-2.59	-0.35	-1.22	0.55	0.25	I
CoRest	CoRest	slight curling; ectopiv wing vein material; Suppression of Ras[N17] wing vein phenotype	N/A	2.82	-3.13	1.3	-1.98	1.52	0.88	IEK
Vps4	Vacuolar protein sorting 4	upward curling; Suppression of Ras[N17] wing vein phenotype	N/A	-0.64	-2.87	0.12	-5.01	-2.68	-3.6	IEK
aret	arrest	upward curling; possible wing vein defect	N/A	-2.8	-6.28	-2.76	-5.98	0.48	5.45	IEK
CG17665	-	wing size reduced; upward curling; pWiz and TRiP hairpin enhance Ras[N17] wing vein phenotype	N/A	1.99	0.65	0.77	-0.57	0.78	1	I
uri	unconventional prefoldin RPB5 interactor	wing size reduced; upward curling; Suppression of Ras[N17] phenotype	N/A	-0.11	-3.65	-0.27	-2.21	-1.13	-1.02	IE
CG3332	-	wing size reduced; upward curling	N/A	-0.02	-2.21	-0.53	-0.02	1.09	0.78	I

Table 2.1: In vivo analysis of PPI or RNAi screen hits. Shown are hits with wing phenotypes of any kind. Bolded genes encode proteins identified in the PPI network, as well as were positive hits in the RNAi screens. “N/A,” not identified in any TAP experiments; PPIs with SAINT < 0.83 were removed as nonspecific. Values in primary screen categories represent average Z-score for two replicates. Ras[N17], dominant-negative Ras; Elp[B1], gain-of-function EGFR allele; tor[4021], a gain of function torso allele; MS1096, promoter used to drive expression of the given transgene in the wing. All genes tested are listed in [44]: Table S5.

Chapter 3

Incorporating quantitative mass spectrometry data in protein interaction analysis

Abstract

[1]Comprehensive protein-protein interaction (PPI) maps are a powerful resource for uncovering the molecular basis of genetic interactions and providing mechanistic insights. Over the past decade, high-throughput experimental techniques have been developed to generate PPI maps at proteome scale, first using yeast two-hybrid approaches and more recently via affinity purification combined with mass spectrometry (AP-MS). Unfortunately, data from both protocols are prone to both high false positive and false negative rates. To address these issues, many methods have been developed to post-process raw PPI data. However, with few exceptions, these methods only analyze binary experimental data (in which each potential interaction tested is deemed either observed or unobserved), neglecting quantitative information available from AP-MS such as spectral counts.

We propose a novel method for incorporating quantitative information from AP-MS data into existing PPI inference methods that analyze binary interaction data. Our approach introduces a probabilistic framework that models the statistical noise inherent in observations of co-purifications. Using a sampling-based approach, we model the uncertainty of interactions with low spectral counts by generating an ensemble of possible alternative experimental outcomes. We then apply the existing method of choice to each alternative outcome and aggregate results over the ensem-

[1] The material in this chapter previously appeared in *BMC Bioinformatics* (2013) as “A sampling framework for incorporating quantitative mass spectrometry data in protein interaction analysis” by George Tucker, Po-Ru Loh, and Bonnie Berger [139].

ble. We validate our approach on three recent AP-MS data sets and demonstrate performance comparable to or better than state-of-the-art methods. Additionally, we provide an in-depth discussion comparing the theoretical bases of existing approaches and identify common aspects that may be key to their performance.

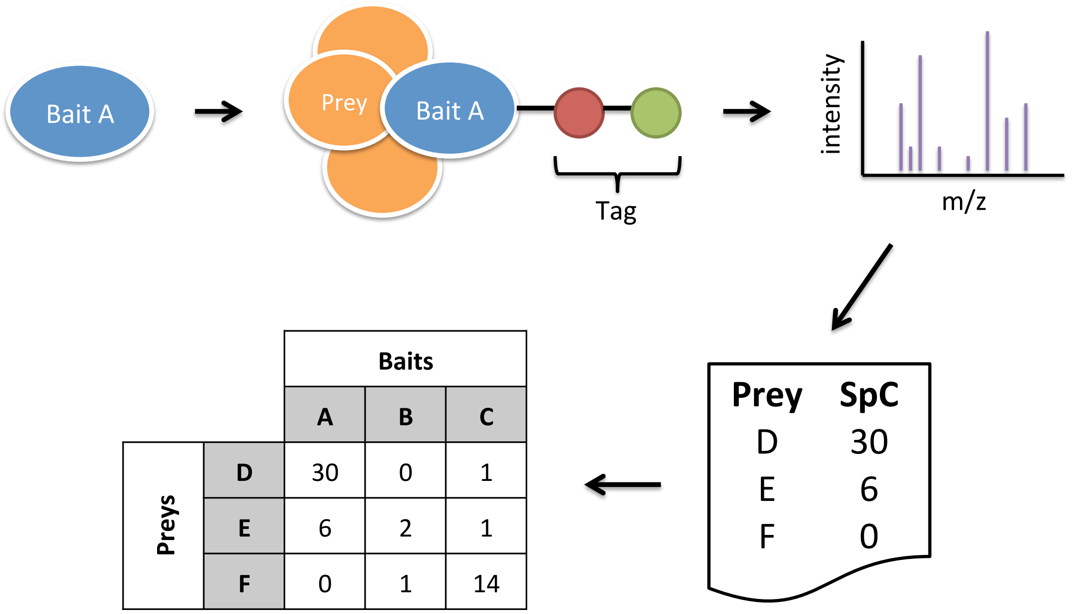
Our sampling framework extends the existing body of work on PPI analysis using binary interaction data to apply to the richer quantitative data now commonly available through AP-MS assays. This framework is quite general, and many enhancements are likely possible. Fruitful future directions may include investigating more sophisticated schemes for converting spectral counts to probabilities and applying the framework to direct protein complex prediction methods.

3.1 Introduction

The importance of protein interactions and protein complexes in understanding cellular functions has driven the generation of comprehensive protein-protein interaction (PPI) maps. The first large-scale PPI maps were generated for the model organism *Saccharomyces cerevisiae*, initially using yeast two-hybrid screens (Y2H) [141, 61] and subsequently by affinity purification combined with mass spectrometry (AP-MS, Figure 3-1) [49, 56]. Similarly, high throughput approaches have been applied to comprehensively map the *Drosophila melanogaster* interactome, initially using Y2H [51] and more recently by AP-MS [53]. With advances in experimental protocols and decreasing costs, medium-scale AP-MS studies have become ubiquitous in proteomics for targeted investigation of specific pathways or interactions. The PPI networks these analyses generate have provided exciting insights into biological pathways and protein complexes, e.g., with relevance to human disease [62]. However, raw AP-MS data includes many false positive and false negative interactions, which are serious confounding factors in their interpretation [15, 28].

To address these issues, numerous methods have been developed to post-process AP-MS data sets. These generally fall in two classes: spoke and matrix models (Figure 3-2). Spoke models [112, 124, 77, 25, 130, 23] produce confidence scores on bait-prey interactor pairs directly observed in the data (i.e., those with non-zero spectral counts), whereas matrix models [28, 54, 159, 149, 53] additionally infer prey-prey interactions that are not directly observed and hence have broader coverage at

Figure 3-1: A typical AP-MS workflow. A typical AP-MS study consists of performing a set of experiments on *bait proteins* of interest, with the goal of identifying their interaction partners. In each experiment, a bait protein is tagged (e.g., using a FLAG-tag or TAP-tag) and expressed in cells. The bait protein and interacting *prey proteins* are affinity purified. The resulting mixture of bait and bound prey proteins is trypsinized into peptide fragments, which are separated by liquid chromatography and passed to a mass spectrometer for analysis. The mass spectrometer produces intensity spectra, which are matched to peptides to deduce proteins present in the purification. Interacting preys thus identified are assigned semi-quantitative *spectral counts* (SpC) indicating the propensity of each prey to bind to the bait. Data is collated from across the experiments into a matrix of bait-prey spectral counts, which serves as the input to post-processing methods that filter contaminants and identify true interactions.

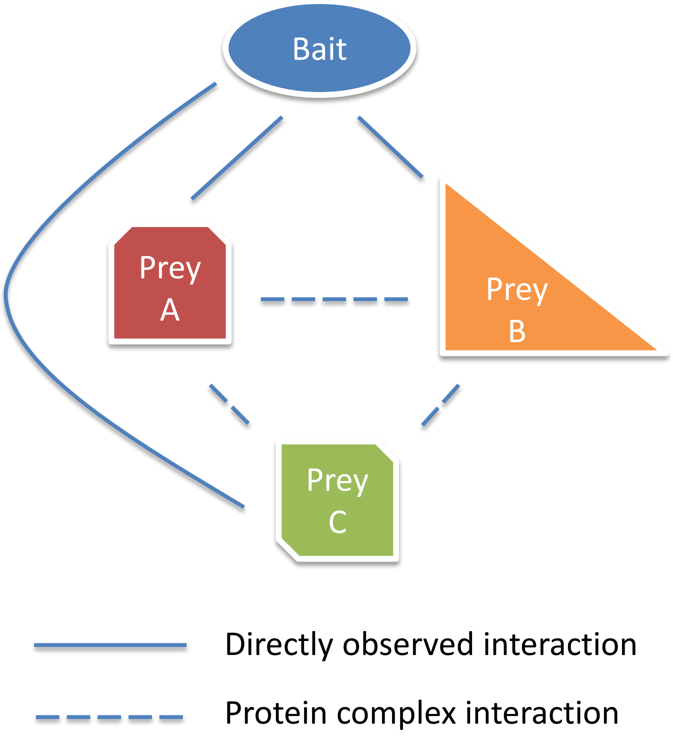


Prey	SpC
D	30
E	6
F	0

		Baits		
		A	B	C
Preys	D	30	0	1
	E	6	2	1
	F	0	1	14

the expense of increased false positives. Development of spoke models has been an intense area of research from the outset; see Nesvizhskii [91] for a thorough review. Matrix models rely on analyzing co-occurrences of pairs of proteins across many experiments and were thus less effective on the initial medium-scale AP-MS studies first performed. As larger AP-MS experiments have become more common, however, matrix models have become increasingly relevant because they can leverage the rich co-occurrence information in these data sets. For example, Guruharsha *et al.* [53] reported significantly improved inference on the *Drosophila melanogaster* interactome using a matrix model approach as compared to state-of-the-art spoke methods.

Figure 3-2: Direct and indirect interactions in AP-MS data sets. The diagram depicts a bait protein bound to a prey protein complex. Solid lines indicate bait-prey interactions that could be observed in an AP-MS experiment, while dashed lines indicate prey-prey protein complex interactions that are not directly observable. Spoke methods make predictions only on directly observed interactions (e.g., Bait with Prey A), whereas matrix models infer protein complex interactions (e.g., Prey A with Prey B). Because the prey proteins do not necessarily form a single complex that interacts with the bait, inferences of prey-prey interactions need to be based on the co-occurrence of pairs of preys across many purification experiments, which strengthens the evidence for interaction.



The existing literature on matrix approaches has almost exclusively considered only binary experimental data (i.e., data sets in which bait-prey interactions are deemed either observed or unobserved, with no additional information about propensity of proteins to interact). An exception is the HGSCore method [53], which to our knowledge is the first to use quantitative information from AP-MS experiments in the form of bait-prey spectral counts. In contrast, spoke models have successfully used quantitative information (e.g., spectral counts [124, 77, 130, 112, 25, 44] and MS1 intensity data [23]) to filter contaminants and assign confidence scores to interactions.

In this study, we propose a novel approach for incorporating quantitative interaction information into AP-MS PPI inference. Our approach aggregates scores over an ensemble of binary data sets that represents the quantitative data, capturing the uncertainty of interactions with low spectral counts. Importantly, the sampling-based framework we propose allows us to directly harness previous binary methods without modification, thus extending previous methods to use quantitative information. We validate our results on a large-scale PPI network and two medium-scale networks. Our approach improves all binary methods that we tested across a broad range of parameter values. In many cases, the improved performance is comparable to or better than state-of-the-art methods that have been developed to leverage spectral counts. Additionally, in the Discussion we characterize previous approaches and identify a common mathematical framework that several successful approaches have used, providing insights that may be valuable in continuing to refine PPI inference techniques.

3.2 Results

3.2.1 Sampling framework

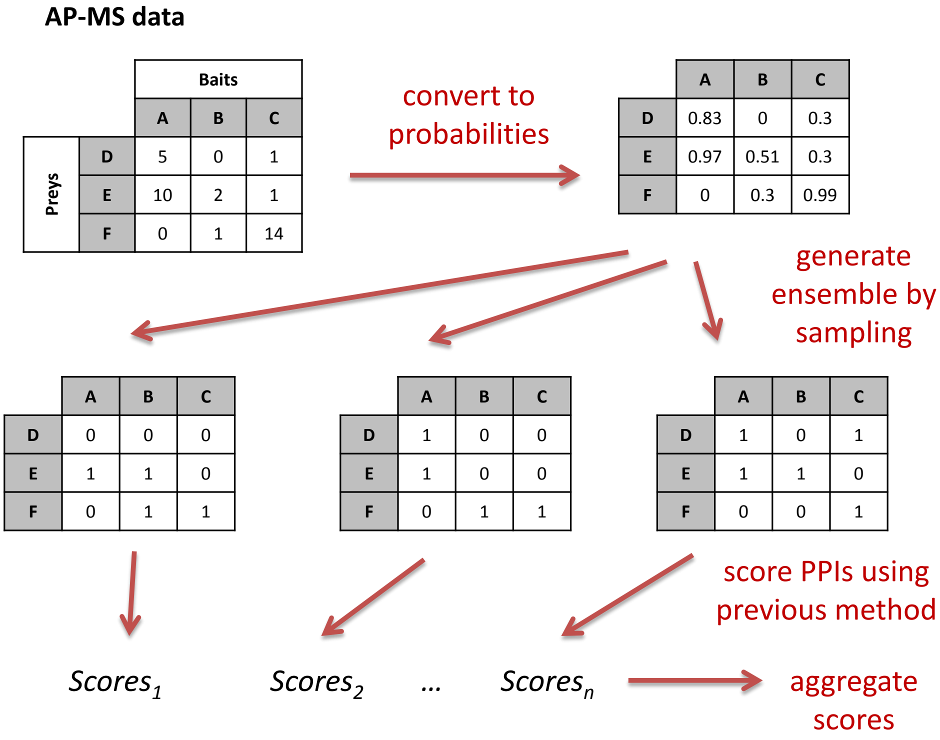
The motivation behind our approach is that spectral count values in AP-MS data sets span a very large dynamic range (from single-digit values to numbers in the thousands - Figure 3-3), and collapsing this range into binary values—as is necessary to apply several previous methods [48, 28, 54, 159, 149]—loses a great deal of potentially useful information. In particular, our intuition is that bait-prey interactions observed with high spectral counts are much more likely to be true interactions than those with spectral counts of only 1 or 2, which might arise through experimental noise. However, there are exceptions; lower abundance proteins can be true interactors if they are pulled-down reproducibly, and high abundance proteins can be sticky proteins that are not necessarily true interactors.

To model this uncertainty in the bait-prey interaction data in a way that allows us to harness existing methods that operate on binary data, we propose a sampling framework that represents the quantitative (spectral count) data set using an ensemble of binary data sets (Figure 3-4). We do so by first converting each positive spectral count into a probability that represents the confidence that the observed interactions were not experimental artifacts. Then, for each of a specified number of trials, we create a binary data set by sampling bait-prey interactions according to their probabilities, and we apply the existing method to the binary data set. Finally, we aggregate the results over the ensemble to produce an overall ranking of possible PPIs.

Explicitly, our framework takes as input a matrix of spectral counts (n_{ij}), where columns correspond to purification experiments and rows to prey proteins. We convert a spectral count of n to the probability $1-(1-p)^n$, where p is a user-defined parameter representing the probability that a single spectral count is the result of a true observation, and we view the n observed spectral counts as arising independently. Using these probabilities, we generate binary data sets of the same size as the original spectral count input matrix by putting a 1 in each matrix cell independently with probability $1-(1-p)^{n_{ij}}$. The resulting distribution of alternative binary realizations of the spectral count matrix thus reflects the range of confidences in different bait-prey interactions, in contrast to the common approach of converting the spectral count matrix to a single binary matrix simply by replacing all positive spectral counts with 1s.

Given an ensemble of alternative binary realizations and an existing PPI scoring algorithm that operates on binary data, we apply the PPI scoring algorithm to each

Figure 3-3: Histogram of spectral counts in the DPiM data set [53]. Of 438,557 positive spectral counts, 94% are less than 20 (shown) and nearly half are either 1 or 2. In contrast, the largest spectral count value is 753.



Preys		Baits		
		A	B	C
	D	5	0	1
	E	10	2	1
	F	0	1	14

	A	B	C
D	0.83	0	0.3
E	0.97	0.51	0.3
F	0	0.3	0.99

	A	B	C
D	0	0	0
E	1	1	0
F	0	1	1

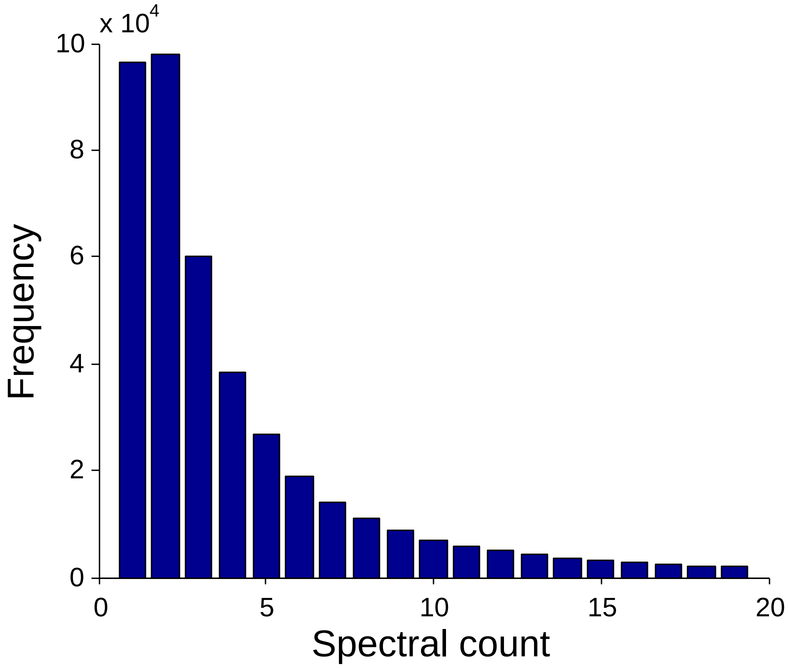
	A	B	C
D	1	0	0
E	1	0	0
F	0	1	1

	A	B	C
D	1	0	1
E	1	1	0
F	0	0	1

realization, in each case producing a score for every possible PPI. We then produce an aggregate score for every PPI by taking the mean of the ensemble of scores for that PPI, possibly after applying an appropriate transformation. (A slight subtlety can arise in aggregating scores because depending on the shape of the score distribution, taking the mean may not be robust. Among the algorithms we evaluated, we observed that the SAI score [48] could produce unbounded negative values, so we lower-bounded SAI scores at 0 before aggregation in order to prevent a single realization from having an extreme effect on the ensemble score.)

An additional consideration is the size of the ensemble required to produce stable results. In the tests we describe below, we ran 120 independent trials and found

Figure 3-4: Sampling approach: Representing spectral counts with ensembles of binary matrices. A summary of our sampling approach. First, each spectral count in the AP-MS data matrix is converted to a probability $1 - (1 - p)^n$, where n is the spectral count. Then, for each cell of the matrix, we sample an independent Bernoulli random variable according to its probability. We repeat this procedure independently for a desired number of trials, obtaining an ensemble of binary matrices representing the original quantitative AP-MS data. Each binary matrix is then used as input to a PPI inference method of choice that operates on binary data, and the results from each trial are aggregated to produce an ensemble score. Notably, the existing PPI inference method is directly applied to each binary matrix without modification.



reasonable score separation between low, medium and high confidence interactions (Figure A-7). Then we further verified that increasing the ensemble size by a factor of four had a negligible impact on the results, indicating that 120 trials was sufficient to average out the stochasticity of the method. Although the minimum number of trials required will vary with the specific data set, our experiments suggest that in general, such a number of trials should sufficiently explore the space of binary realizations without presenting a computational burden, especially because the ensemble computations can be easily parallelized.

3.2.2 Validation on three AP-MS data sets

We benchmarked our method by producing predictions from three AP-MS data sets: the recently published Drosophila Protein interaction Map (DPiM) [53] which includes over 3000 baits, a medium-scale human data set (TIP49) with 27 baits [112], and a *Drosophila* study focusing on the MAPK pathway with 21 baits [130]. On each evaluation data set, we applied our sampling framework to three previously published binary matrix methods for PPI inference: Hart et al. [54], PE [28], and SAI [48]. Each method produced a ranked list of interactions.

A standard approach to evaluating inferred interactions is to compare predictions with a high-confidence gold standard set. However, such a reference is challenging to construct. Few large-scale databases are available, and even the largest are understood to be incomplete and include false positive interactions. In light of these concerns, we follow the validation strategy used in Guruharsha et al. [53] of considering the overlaps between multiple curated data sets, obtaining subsets of PPIs with increasingly stringent thresholds on the number of supporting sources. The idea is that we can have high confidence in interactions supported by multiple lines of medium-confidence evidence, reducing the false positive rate in the gold standard data set (with the caveat that this approach may be biased toward well-studied pathways). We applied this procedure to create validation data sets from the Drosophila Interactions Database (DroID) [90] for *Drosophila* PPI predictions and BioGRID [125] for human PPI predictions (See **Methods** for details).

For each method, we compared the top 25,000 predicted interactions for the DPiM data set and the top 2,500 predicted interactions for the TIP49 and MAPK data sets to gold standard interactions supported by increasing numbers of sources, as in Guruharsha et al. [53]. Our sampling framework produced robust improvement to the binary methods across all levels of support and all data sets (Figure 3-5). Moreover, the improved methods perform better than or comparably to state-of-the-art methods that use spectral count data (HGSCore [53] and SAINT [25]). The choice of cutoff at the top 25,000 and 2,500 interactions was arbitrary, and the results are similar at different cutoffs (Figures A-1,A-2,A-3).

The sole parameter in our method is the probability p that represents the reliability of a single peptide observation. We suggest a default value of $p = 0.3$, but the performance improvements obtained using our sampling framework are robust across a wide range of values of p (Figure 3-6,A-4,A-5) and for different confidence cutoffs (Figure A-6).

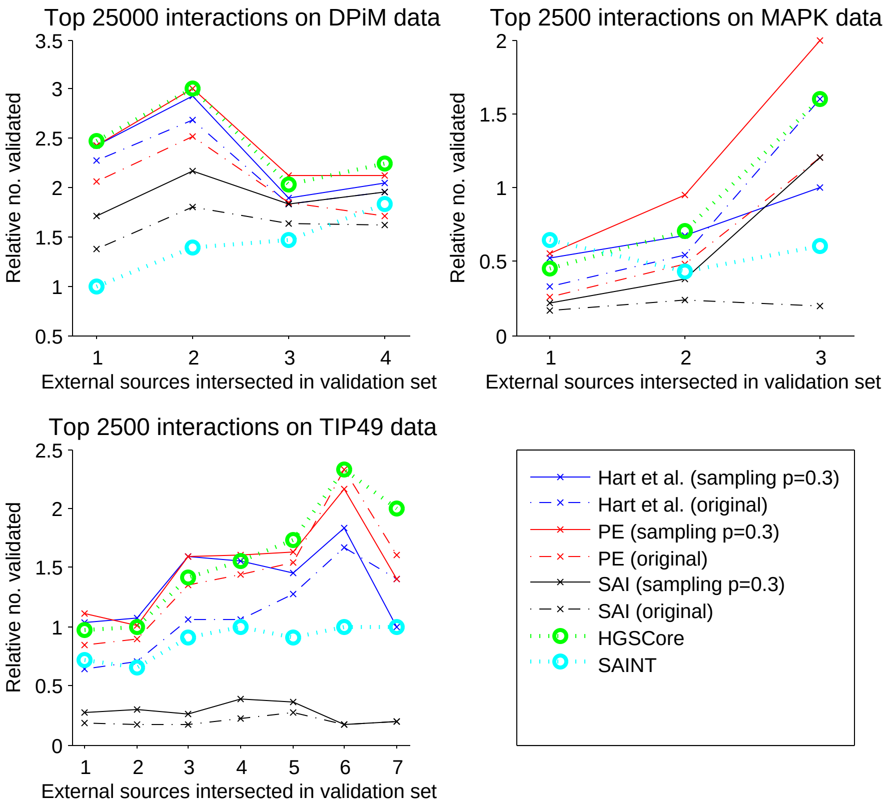
3.3 Discussion

The literature of published methods for PPI inference from AP-MS data is substantial, and in continuing to develop methodological improvements, it is valuable to understand the similarities and differences among existing approaches and identify key ideas.

3.3.1 Characterization of methods

Broadly speaking, methods can be broken down into two classes of models—spoke and matrix models—and by their scoring method. Spoke models make predictions solely on bait-prey interactions, while matrix models infer prey-prey interactions as well. Because prey-prey relationships are never directly observed, matrix models use the co-occurrence of pairs of proteins over multiple experiments to make inferences. Methods can also be characterized by their scoring functions, which generally fall into two classes: evidence-based scoring and null model-based scoring. In evidence-based scoring, models are built that estimate the likelihoods of observations under interacting and non-interacting pairs. Typically, a log likelihood ratio is then summed across experiments, implicitly assuming independence. Evidence-based scoring approaches, such as the PE [28] and C2S [149] scores, can easily combine direct bait-prey observations and prey-prey observations in the same model. However, because likelihood models for interacting and for non-interacting pairs must be constructed, these scores tend to have more tuning parameters that must be estimated from scarce gold standard validation data. In null model based approaches, such as Hart et al. [54],

Figure 3-5: Performance comparison of PPI inference methods. Performance of our sampling approach applied to PPI inference methods that operate on binary bait-prey interaction data (Hart et al. [54], PE [28], and SAI [48]), and compared to state-of-the-art methods that make use of spectral counts (HGSCore [53] and SAINT [25]). For each method that operates on binary data, two curves are plotted: (i) a dashed curve that shows the performance of the method when applied to a direct binarization of the spectral count data (i.e., converting all nonzero spectral counts to 1s)—a common approach—and (ii) a solid curve showing performance upon applying our sampling approach with $p = 0.3$. We evaluate performance according to the number of PPI inferences (out of the highest-confidence 25,000 or 2,500) validated on gold standard tests, as explained in the main text. The plot shows performance relative to a baseline method of simply ranking PPIs in decreasing order of observed spectral counts. All methods were run using default parameter settings.



HGSCore [53], and SAI [48], a model for non-interacting pairs is assumed and fit from the data. This forms an empirical null distribution under which observations can be scored. The advantage of such an approach is that only the null distribution has to be tuned, so in many cases tuning with gold standard validation sets is unnecessary.

An additional consideration for any method that combines spoke and matrix information is the balance between information from direct bait-prey observations and prey-prey co-occurrences. These sources of information are clearly distinct, so the weighting between the two must be carefully calibrated, potentially requiring gold standard validation data. Proper calibration is critical to performance and may explain why Hart et al. and HGSCore, which seemingly sub-optimally ignore spoke information, perform significantly better on our tests than SAI [48], which uses both spoke and matrix information.

For experiments with a handful of baits, we expect that methods relying on spoke information will have the best performance because matrix methods rely on analyzing co-occurrences of pairs of proteins across many experiments. However, even for the medium-scale experiments that we analyzed, methods that rely solely on matrix information performed competitively with methods that used spoke information. We foresee that as experiment sizes grow, matrix relationships will be increasingly informative, so it will be crucial to consider both spoke and matrix information. Although our approach is applicable to any binary method, in our experiments, we found that for nearly all experiments PE was the top performer amongst the binary methods. In addition, because PE uses spoke and matrix information, we recommend using it in our framework.

3.3.2 Low rank plus sparse matrix framework

Interestingly, several methods (e.g., Hart et al. [54], HGSCore [53], SAINT [25]) can be understood under a common "low rank plus sparse matrix" framework. Hart et al. [54] considered a null model in which interaction partners are chosen independently at random in proportion to the number of interactions each partner protein was observed in. Although Hart et al. [54] used a hypergeometric distribution, for large-scale studies, the score for interaction between proteins A and B is well approximated using a Poisson cumulative distribution function (CDF), taking the form

$$-\log\left(1 - PoissonCDF\left(X_{AB}; \lambda = \frac{N_A}{N} \times \frac{N_B}{N} \times N\right)\right),$$

where X_{AB} is the number of experiments that protein A and protein B co-purify in, N_A (resp. N_B) is the number of co-purifying pairs that protein A (resp. B) is observed in, and N is the total number of co-purifying pairs.

In the above form, λ factors as a rank-1 matrix, so that the method can be seen as modeling the co-occurrence matrix X_{AB} as the sum of a rank-1 "background" matrix (blurred by Poisson noise) and a sparse matrix indicating true interactions. Notably, X_{AB} ignores quantitative information, simply counting experiments in which proteins were co-purified. HGSCore [53] is an extension of the Hart et al. score that incorporates spectral count information through a transformation of the spectral counts (instead of directly using the co-occurrence matrix) and then analyzes the pseudo co-occurrence matrix in a similar manner. For the same reasons as above, we can view HGSCore as a rank-1 null model plus sparse true interactions, where the rank-1 component is estimated from a transformation of the spectral count data.

Similarly, SAINT [25] uses a probabilistic formulation to decompose a matrix of observed counts as a sum of: a rank-1 matrix, a sparse true interaction matrix, and generalized Poisson noise. Interestingly, SAINT decomposes the matrix of spectral counts—as opposed to co-occurrences—and has an entirely different justification for using a low rank model. Hart et al. and HGSCore assume that interaction partners are chosen at random in the null model, which gives rise to a low rank structure in the co-occurrence observations. Alternatively, SAINT assumes that contaminant proteins produce similar spectral counts across all bait experiments, which gives rise to a low rank structure in the spectral count observations. SAINT uses solely spoke evidence while Hart et al. and HGSCore use only co-occurrence evidence, suggesting that some combination of these approaches under a common framework may be an interesting direction for future investigation.

3.3.3 Moving toward complexes

As protein biology is ultimately driven by the interactions of protein complexes—not just pairwise protein interactions—recent work has begun inferring protein complexes directly from AP-MS data [112, 163, 113, 24, 50, 128]. Traditionally, methods have first inferred PPIs and then clustered proteins into complexes (e.g., Guruharsha et al. [53]); however, information may be lost in this two-step procedure that first post-processes the data into high-confidence pairwise interactions. As with matrix models, some recent methods that bypass this first step have considered only binary experimental data [163, 50], whereas others have successfully used spectral count information [112, 113, 128, 24]. A similar sampling approach could be used to extend methods that consider only binary data to leverage spectral counts.

3.4 Conclusions

As large-scale AP-MS experiments have become more common, an opportunity to leverage indirect co-occurrence information for PPI inference has arisen. Our sampling framework harnesses existing matrix methods for PPI inference that could previously only be applied to binary interaction data, achieving robust improvements across a range of data sets and enabling comparable or better performance versus current state-of-the-art methods. This framework extends the existing body of work on binary interaction analysis to apply to richer spectral count data now commonly available. Moreover, it is sufficiently general to have potential for future application in related protein interaction inference studies.

3.5 Methods

3.5.1 AP-MS data sets

The main data set we analyzed, DPiM, is a large-scale AP-MS study of the *Drosophila* proteome with 3485 experiments, which collectively pulled down 4927 distinct proteins ([53], Table S1). The DPiM data set is unique among publicly available AP-MS data sets because of its large size, which gives us confidence that the results we observed are not the result of random noise or overfitting. We also tested our approach on two medium-scale AP-MS data sets. One is another *Drosophila* study that focused on the MAPK pathway [130]; this data set contained 63 experiments, which collectively pulled down 1078 distinct proteins and included 9 control experiments. The other is a human data set referred to as TIP49 and originally published in Sardiu et al. [112]. We obtained the interaction data set, consisting of 35 experiments, which collectively pulled down 1207 distinct proteins and included 9 control experiments, from Choi et al. ([25], Table S1).

3.5.2 Validation data sets

To validate *Drosophila* PPI inferences, we used the data sets in the DroID database [90]. We excluded the Perrimon co-AP complex and DPiM co-AP complex data sets to avoid contaminating our test sets with training data, leaving 7 other PPI data sets that we used in the above validation procedure. The validation set contained 58,657 interactions supported by at least one source, 3,310 interactions supported by at least two sources, 289 interactions supported by at least three sources, and 67 interactions supported by at least four sources.

To validate human PPI inferences, we used BioGRID v3.1.79 [125], which contains 40,680 interactions supported by at least one source, 11,054 interactions supported by at least two sources, 4,879 interactions supported by at least three sources, and 2,271 interactions supported by at least four sources.

3.5.3 Implementation

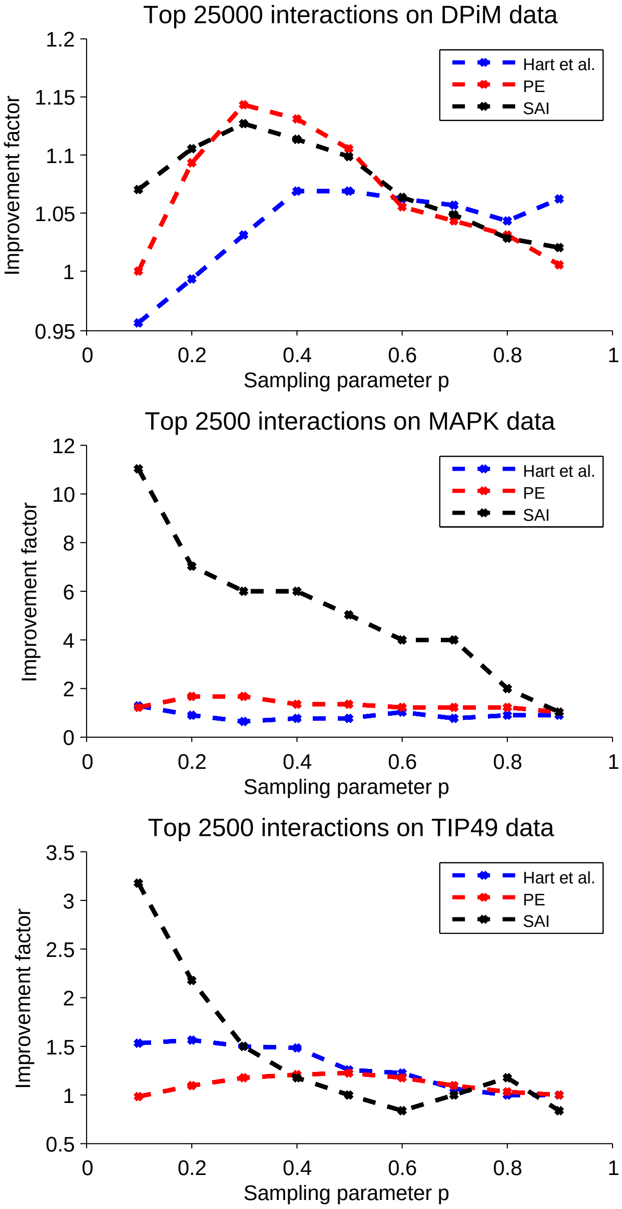
We re-implemented the SAI [48], PE [28], Hart et al. [54], and HGSCore [53] methods; each is described in its reference but code is not provided. The PE score uses two parameters, r, representing the probability of detecting a true association in a purification experiment, and n_{pseudo}, the number of pseudocounts added for each prey.

Since Collins et al. [28] estimates values of $r = 0.51, 0.62$, and 0.265 on three example data sets and suggests using $n_{\text{pseudo}} = 20, 10$, or 5, we set $r = 0.3$ and $n_{\text{pseudo}} = 10$. We downloaded and ran SAINT [25] with default parameters.

We also implemented the C2S score [149] but found its performance to be highly sensitive to the *tpr* (true positive rate) parameter; some values of *tpr*—including the default 0.6 in at least one of our tests—result in inferred values of the probabilistic parameters r_{bp} and r_{pp} that exceed 1, causing improper values in subsequent calculations (e.g., logarithms of negative numbers). We therefore excluded C2S from our analysis.

When we applied our sampling framework to data sets containing replicates, we treated columns corresponding to replicates independently. When we tested all of the methods on data sets containing controls, only SAINT, which explicitly models control data, used the controls.

Figure 3-6: Sensitivity of performance to sampling parameter p. We plot the improvement in performance, as a function of p, achieved by applying our sampling approach vs. applying methods to direct binarizations of spectral count data. Performance is measured using the same setup as in Figure 3-5. For figure readability, we show results for just the validation sets consisting of interactions supported by at least 3 pieces of evidence; similar results hold for the other validation sets.



Chapter 4

Inferring interactors from LUMIER using mixture models

Abstract

We describe a novel method for determining significant protein interactions from raw LUMIER data that corrects for spatial biases that occur in high-throughput LUMIER screens. We apply this method to a large LUMIER screen with 60 preys and 800 baits to characterize chaperone, co-chaperone, and client interactions. We show that our method is able to recover significantly more true interactions than previous methods. From this data, we assemble a comprehensive network of chaperone, co-chaperone, and client interactions that reveals new insights into co-chaperone specificity.

4.1 Introduction

[1]The crowded intracellular milieu poses major challenges to protein folding in vivo. Intrinsic and extrinsic stress can easily derail the finely tuned cellular protein homeostasis (proteostasis) network, leading to protein misfolding and aggregation. To cope, cells have evolved mechanisms to maintain proteostasis and to protect themselves from environmental insults. Indeed, a substantial fraction of the cellular proteome is

[1] This section is adapted from a manuscript to appear in *Cell* as "A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways" from Mikko Taipale, George Tucker, Jian Peng, Irina Krykbaeva, Zhen-Yuan Lin, Brett Larsen, Hyungwon Choi, Bonnie Berger, Anne-Claude Gingras, and Susan Lindquist. We contributed the computational processing of the interactions detected by LUMIER. In the subsequent subsections, we describe the application of our method to this data set.

dedicated to maintaining proteostasis [98].

The proteostasis network is intimately and very broadly linked to human disease. In common and rare diseases alike, it has emerged as a central modifier of disease progression and severity. Perturbation of the proteostasis network has been implicated in many if not most diseases ranging from neurodegeneration, to cancer, to Mendelian disorders, thereby contributing immensely to human disease burden [88, 146]. At the same time, preclinical models and more recent clinical results with drugs that target central modules of the network, such as proteasome or Hsp90 inhibitors, have shown that targeting the network has high therapeutic potential [138]. However, it is clear that we need a more detailed understanding of the proteostasis network to understand how exactly it is perturbed in disease and to develop more effective and specific therapeutics.

Chaperones are the most prominent class of proteins that shape the proteostasis network. They transiently bind thousands of substrate proteins (clients) in the cell and promote their folding, trafficking, and degradation [109]. The three major chaperone families chaperonins, Hsp70, and Hsp90 have distinct mechanisms of action and modes of client protein recognition. Chaperonins such as GroEL recognize and encapsulate proteins that are kinetically trapped in partially folded molten globule conformations, whereas Hsp70 binds short, hydrophobic peptide motifs that are often exposed during translation and in partially or fully unfolded proteins [55]. In contrast, most Hsp90 clients are almost completely folded but often require this chaperone for the final steps of folding, such as ligand or substrate binding [132].

As a result of these fundamental mechanistic differences in client protein recognition, chaperone families have distinct client preferences. Recent systematic proteomic approaches have started to uncover the in vivo client protein ensembles of each chaperone family [14, 69, 133, 150, 165]. However, previous studies have employed widely varying methods and model organisms, making it a challenge to quantitatively compare results and integrate them into a coherent model. Perhaps more importantly, however, chaperones do not function in isolation in vivo. Rather, they dynamically associate with a diverse set of cofactors. These factors, collectively referred to as co-chaperones, provide a host of auxiliary functions to chaperones, ranging from regulating the rate of client release to recruiting specific client proteins to the core chaperone [34, 65].

Detailed in vitro studies have revealed how co-chaperones interact with chaperones and regulate their function [95, 122, 123], but we have such information for only a few co-chaperones. Moreover, a growing body of evidence suggests co-chaperones play much more than a supportive role in protein homeostasis. For example, some co-chaperones possess intrinsic chaperone activity [40, 71, 166], whereas others independently regulate cellular processes that are distinct from those of canonical chaperones [35, 161]. Yet, both the client-protein specificity and possible chaperone-independent functions of most co-chaperones remain enigmatic.

Chaperone interactions are difficult to assay by standard methods because they interact so diversely, are highly abundant, interact with proteins that may be expressed at much lower abundance, and interact with a diverse set of co-chaperones to perform their function that are not present in binary assays such as yeast two-hybrid. Luminescence-based mammalian interactome mapping (LUMIER) [6] and its extension LUMIER with bait control (BACON) [133] are well suited to detecting these interactions. Here, we have taken a systematic and integrative approach, surveying the physical interaction landscape of all known Hsp90 co-chaperones and several known Hsp70 co-chaperones. We combine mass spectrometry and quantitative LUMIER assays to characterize the client protein specificity of co-chaperones and how they are integrated into the proteostasis network. Our analysis confirms the existence of two partially overlapping networks of chaperone/client interactions, centered on Hsp90 and Hsp70. It populates these networks with new members and dramatically increases their connectivity, while suggesting unique functions for most co-chaperones and identifying several new domain-specific co-chaperones.

In this chapter, we describe the method we developed to determine significant protein interactions from LUMIER data. The method corrects for spatial biases that occur in high-throughput LUMIER screens. We apply this method to a large LUMIER screen with 60 preys and 800 baits to characterize chaperone and co-chaperone

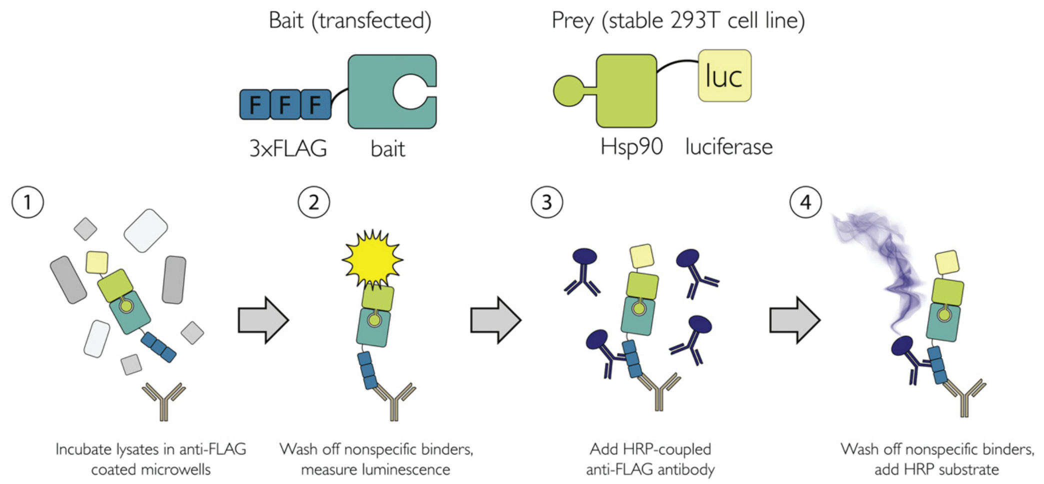

Bait (transfected)
Prey (stable 293T cell line)
F F F
3xFLAG
bait
luc
Hsp90
luciferase
1
2
3
4
Incubate lysates in anti-FLAG coated microwells
Wash off nonspecific binders, measure luminescence
Add HRP-coupled anti-FLAG antibody
Wash off nonspecific binders, add HRP substrate

Figure 4-1: Steps of the LUMIER with BACON assay[133]. 3 x FLAG tagged bait protein is coexpressed with renilla lucerifase tagged prey protein (HSP90 in this case). (1) Cells are lysed and the lysate is incubated in anti-FLAG coated microwells. (2) Nonspecific binders are washed away and interaction strength is measured by luminescence. (3 & 4) Bait abundance is measured by ELISA. Adapted from Figure 1 in [133].

interactions. We show that our method is able to recover significantly more true interactions than previous methods, and this allows us to construct a comprehensive network of chaperone and co-chaperone interactions. Finally, we discuss the insights gained from mapping chaperone, co-chaperone, and client protein interactions.

4.2 LUMIER

LUMIER is a co-affinity purification assay that uses luminescence to measure interaction strength between a pair of proteins called the bait and prey proteins. Renilla luciferase, an enzyme that emits light, is fused to a protein of interest, called a prey protein. In each interaction test, the prey protein is coexpressed with a FLAG-tagged protein, called a bait protein. The FLAG tag is a polypeptide tag added to a protein to enable efficient purification of the tagged protein. This allows us to co-purify the bait protein as well as any prey protein that interacts with the bait protein. Then, we can measure the luminescence to quantify the abundance of the bound prey pro-

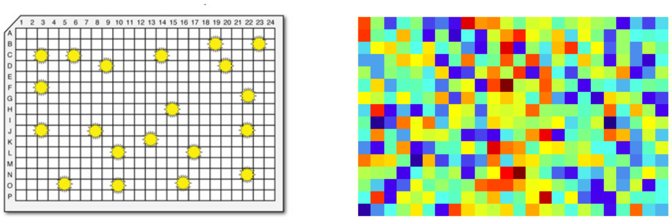


Figure 4-2: Visualization of luminescence readout from LUMIER. The LUMIER experiments were performed in parallel on 384-well plates (left panel). The luminescence can be visualized as a heatmap (right panel), where red indicates high luminescence and blue indicates low luminescence. In our experiments, all wells on the same plate contain the same prey while the bait is varied from well to well. Adapted from Figure 1 in [133].

tein and determine if an interaction occurred. LUMIER with bait control (BACON) [133] adds a step to quantify the abundance of the bait protein by enzyme-linked immunosorbent assay (ELISA) using a separate FLAG antibody. The bait abundance can be used to filter out bait proteins that failed to express and to normalize the LUMIER scores. Because interactions are probed *in-vivo*, we can detect interactions that involve additional protein partners and interactions that are contingent on post-translation modifications.

High-throughput LUMIER experiments are conducted in parallel on plates with hundreds of wells containing separate interaction test (Figure 4-2). In the experiments that we analyzed, every well on the same plate had the same prey protein while the bait proteins varied from well to well.

Previously, LUMIER has been used to map the transforming growth factor-β (TGBβ) pathway, to map HSP90 client interactions [133], to assay small-molecule binding to kinases [134], and to map other interactions.

4.3 Methods

As in all interaction assays, the interaction scores, in this case luminescence, have to be compared to a negative control to identify significant interactions. Even when no interaction occurs, the prey protein may bind at a low affinity to the FLAG-tag or the

FLAG antibody, which can cause spurious luminescence. We used statistical methods to model the background luminescence distribution and to identify significant interactions.

We use a two-step procedure to remove spatial bias and background batch effects before scoring LUMIER experiments. In many experiments, we observe that raw luminescence values exhibit spatial gradients across plates (Fig. 4-3). These spatial gradients may be caused by unavoidable temperature gradients that affect the kinetics of the reactions. Although the gradient changes smoothly, the differences between wells in a plate can be significant and may result in false positives. Unlike other assays with similar biases, such as microarrays, true interactions only increase the luminescence, hence simple averaging approaches to estimate the background would overestimate the bias. In the first step, we propose to explicitly model and then subtract out the smooth spatial bias. This approach also removes batch effects that influence a whole plate, allowing us to pool plates from the same prey.

All wells, even those without interactions, exhibit background levels of luminescence. For those plates without significant spatial bias, we empirically find that the background luminescence is well fit by a log-normal distribution, motivating modeling the luminescence values on a log scale. To determine significance above random background luminescence, we explicitly model the background log-luminescence with a normal distribution. In the following subsections, we provide detailed descriptions of the models and scoring procedure.

4.3.1 Spatial Bias Model

Specifically, we modeled the log-luminescence values with a Gaussian process mixture model to account for the spatial bias as well as true interactions. The observed luminescence is a combination of background luminescence modulated by a spatial bias and potentially luminescence from true interactions. We assume that the background log-luminescence is normally distributed, which is consistent with control experiments. We also noticed that the mean background log-luminescence depends on the identity of the prey protein, but is consistent on all plates of the same prey protein.

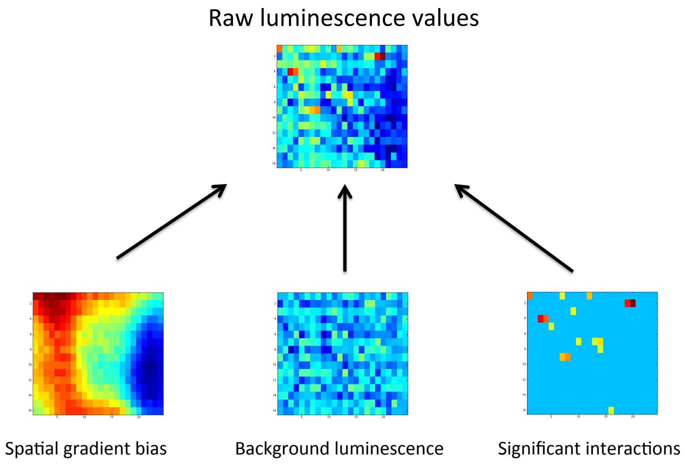


Figure 4-3: Mixture model for data normalization. The raw luminescence values (top panel) are composed of: a background log-luminescence drawn from a prey-dependent normal distribution (middle panel), a Gaussian variable which measures the spatial bias and correlation between neighboring wells on a plate (left panel), and an effect variable for each well which accounts for the actual interaction strength (right panel). The effects are visualized as heatmaps, where red indicates high luminescence and blue indicates low luminescence.

Similarly, the variance of the background distribution varies from prey to prey but is empirically similar across plates for the same prey. Based on these observations, we used a mixture model with three components: a background log-luminescence drawn from a prey-dependent Gaussian distribution, a Gaussian variable which measures the spatial bias, and an effect variable which accounts for the actual interaction strength. Our model flags outliers as potential interactions without biasing our estimate of the background.

Explicitly, our model is generative (i.e., it specifies the process generating the LUMIER data). For each prey g, we have a Gaussian distribution $N(m_g, \rho^2)$ from which the background log-luminescence μ_{gp} is drawn, which corresponds to the batch bias for plate p. m_g is the prey-dependent mean value; and ρ^2 is the background variance that controls the differences between plate-specific batch effects. In addition to this background noise model, we introduce a Gaussian process to account for the spatial bias on a plate. The spatial bias $b_{gp\cdot}$ for all wells on a plate p of prey g is drawn from a correlated multivariate normal distribution $N(0, K)$, where K specifies the smoothness or correlation among the spatial biases on neighboring wells. We used a squared exponential covariance matrix, estimated from the control plates.

For each plate p of prey g, we also have a mixture parameter π_{gp}, which is the proportion of true interactions on the plate. This parameter is drawn from a Beta distribution with prior parameter a and b. Then for each well w on the plate, we have a binary variable z_{gpw} drawn from a Bernoulli distribution with the mixture parameter π_{gp}. If the binary variable $z_{gpw} = 1$, we add a parameter modeling the interaction strength that represents the average binding strength between bait and prey; otherwise, the observed log-luminescence from the well is assumed to come solely from the background noise and the spatial bias.

Our model can be fully specified as (Figure 4-3)

$$x_{gpw} \sim (1 - z_{gpw})N(b_{gpw} + \mu_{gp}, \sigma_g^2) + z_{gpw}N(b_{gpw} + \mu_{gp} + \mu_e, \sigma_g^2 + \sigma_e^2),$$

where σ_g^2 is the variance of the background log-luminescence distribution for prey g, and μ_e and σ_e^2 are the parameters for the log-luminescence distribution for the significant or true biological interactions.

Inferring parameters in the model

Using the Expectation-Maximization algorithm (EM), we estimated parameters in the model by maximizing the log-likelihood of the observed data. The overall log-likelihood summed over each plate p, prey g and well w can be written as follows

$$\begin{aligned} L = &\sum_{gpw} z_{gpw} \left(\frac{-(x_{gpw} - \mu_{gp} - b_{gpw})^2}{2\sigma_g^2} - \frac{1}{2}\log(2\pi\sigma_g^2) \right) \\ &+ (1 - z_{gpw}) \left(\frac{-(x_{gpw} - \mu_{gp} - \mu_e - b_{gpw})^2}{2(\sigma_g^2 + \sigma_e^2)} - \frac{1}{2}\log(2\pi(\sigma_g^2 + \sigma_e^2)) \right) \\ &+ \sum_{gpw} z_{gpw} \log \pi_{gp} + (1 - z_{gpw}) \log(1 - \pi_{gp}) \\ &+ \sum_{gp} \left(\frac{-(\mu_{gp} - m_g)^2}{2\rho^2} - \frac{1}{2}\log(2\pi\rho^2) \right) - \sum_{gp} \frac{1}{2} b_{gp\cdot}^T K^{-1} b_{gp\cdot}, \end{aligned}$$

up to a constant that does not depend on the parameters. The first summation is the log-likelihood of the observed log-luminescence given our generative model, and the rest of the terms are the log-likelihood of the prior distribution of the parameters. In the model, z_{gpw} are unobserved variables and $\Theta = \{m_g, \mu_{gp}, b_{gpw}, \pi_{gp}, \sigma_g, \mu_e, \sigma_e\}$ are parameters that need to be estimated. To perform the EM algorithm, we iteratively replaced the z_{gpw} by their expectation given Θ and maximized the log-likelihood with respect to Θ.

Specifically, we optimized Θ by gradient descent. The gradient of the parameters can be computed efficiently in closed form

$$\begin{aligned}
\frac{\partial L}{\partial m_g} &= \sum_p \frac{\mu_{gp} - m_g}{\rho^2} \\
\frac{\partial L}{\partial \mu_e} &= \sum_{gpw} (1 - z_{gpw}) \frac{x_{gpw} - b_{gpw} - \mu_{gp} - \mu_e}{\sigma_g^2 + \sigma_e^2} \\
\frac{\partial L}{\partial \sigma_e} &= \sum_{gpw} (1 - z_{gpw}) \frac{\sigma_e}{\sigma_g^2 + \sigma_e^2} \left(\frac{(x_{gpw} - b_{gpw} - \mu_{gp} - \mu_e)^2}{\sigma_g^2 + \sigma_e^2} - 1 \right) \\
\frac{\partial L}{\partial \mu_{gp}} &= \sum_w z_{gpw} \frac{x_{gpw} - b_{gpw} - \mu_{gp}}{\sigma_g^2} + (1 - z_{gpw}) \frac{x_{gpw} - b_{gpw} - \mu_{gp} - \mu_e}{\sigma_g^2 + \sigma_e^2} - \frac{\mu_{gp} - m_g}{\rho^2} \\
\frac{\partial L}{\partial \sigma_g} &= \sum_{pw} z_{gpw} \frac{1}{\sigma_g} \left(\frac{(x_{gpw} - b_{gpw} - \mu_{gp})^2}{\sigma_g^2} - 1 \right) \\
&\quad + (1 - z_{gpw}) \frac{\sigma_g}{\sigma_g^2 + \sigma_e^2} \left(\frac{(x_{gpw} - b_{gpw} - \mu_{gp} - \mu_e)^2}{\sigma_g^2 + \sigma_e^2} - 1 \right).
\end{aligned}$$

Given these gradient calculations, we used quasi-Newton L-BFGS to optimize the parameters. Because π_{gp} is constrained to be a probability, we estimated it separately. Maximizing π_{gp} given the rest of Θ is straightforward

$$\pi_{gp} = \frac{\sum_w z_{gpw}}{W},$$

where W is the number of wells on a plate. We also estimated b_{gpw} separately because it has complex relations with the other parameters. Taking the gradient with respect to $b_{gp\cdot}$ gives

$$
\begin{aligned}
\nabla_{b_{gp\cdot}L} = diag(z_{gp\cdot}) \left(\frac{x_{gp\cdot} - \mu_{gp}}{\sigma_g^2} \right) + diag(1 - z_{gp\cdot}) \left(\frac{x_{gp\cdot} - \mu_{gp} - \mu_e}{\sigma_g^2 + \sigma_e^2} \right) \\
- \left[diag(z_{gp\cdot}/\sigma_g^2) + diag\left(\frac{1 - z_{gp\cdot}}{\sigma_g^2 + \sigma_e^2} \right) + K^{-1} \right] b_{gp\cdot}.
\end{aligned}
$$

This leads to a closed form update for b_{gpw}, which required solving a linear system of equations.

In summary, we repeatedly replaced the z_{gpw} by their expectation, maximized $\{m_g,\ \mu_{gp},\ \sigma_g, \mu_e, \sigma_e\}$, maximized π_{gp}, and maximized b_{gpw} until convergence.

4.3.2 Background Luminescence Model

After removing the spatial bias, we compared the interaction effect with the estimated background log-luminescence distribution. To do so, we used a previously described approach to compute Z-scores for each interaction [133]. Briefly, we estimated a mean and standard deviation parameter for each prey. We used the mode of the distribution of log-luminescence scores as the mean. Then, we estimated the standard deviation of the log-luminescence values as if all values above the mean were censored. When estimating the background luminescence distribution, considering only values below the mean is reasonable because true interactions only cause an increase in luminescence (as opposed to a decrease). Finally, to score interactions, we calculated the Z-score using the mean and standard deviation estimated above for each prey.

4.4 An application to mapping chaperone, co-chaperone, and client interactions

In the following subsections, we describe an application of this method to a large LUMIER screen with 60 preys and 800 baits to characterize chaperone, co-chaperone, and client interactions. We validated our method against BioGRID [19], a database of known protein interactions, and showed that it identifies more true interactions than previous methods.

4.4.1 Experiment setup

Our collaborators quantitatively assayed 800 known or putative clients for interaction with 60 chaperones, co-chaperones, and protein quality control factors. All interactions were assayed in duplicate. Lastly, all plates had at least one well with a 3xFLAG-tagged bait EGFP (which was not expected to interact with the prey) and at least three empty wells with no bait proteins to serve as negative controls.

4.4.2 Preprocessing

We found that a number of bait proteins failed to express. To be conservative in the reported interactions, we aggressively filtered interactions with low ELISA score. Specifically, we first quantile normalized ELISA scores across plates having the same bait protein configuration. Then we removed all interactions having ELISA score lower than the 95th (90th) quantile of the control ELISA scores for the chaperone, co-chaperone::client interactions (co-chaperone::beta-propeller interactions). We reduced the stringency for the co-chaperone::beta-propeller interaction experiments because we had fewer wells to estimate the control ELISA distribution from. We also flagged control wells that had abnormally high quantile normalized ELISA scores, manually checked, and removed mislabeled wells.

4.4.3 Validation

To validate our method, we compared the inferred interactions against a high-confidence gold standard interaction dataset. We compared methods at several false discovery rate (FDR) cutoffs for the number of predicted interactions that were validated in the gold standard. Comparing methods at fixed FDR cutoffs allowed methods to predict novel interactions without suffering a penalty, as long as the method controlled FDR. This is particularly important because we expected LUMIER to discover many novel interactions that were not in the gold standard. We estimated the FDR of each method using control wells on each plate.

Specifically, to calculate an empirical upper bound for FDR, we scored both negative control and experiment LUMIER wells. Each assayed plate had a 3xFLAG-tagged EGFP and at least three empty wells with no bait protein as negative controls. From the control well scores, we calculated the number of wells passing different score thresholds, which allowed us to estimate the number of false discoveries. Explicitly, we upper bounded the false discovery rate by

$$FDR(t) \leq \frac{C(t)/N_c}{E(t)/N_e}$$

where $FDR(t)$ is the false discovery rate at a threshold t, $C(t)$ is the number of negative control wells with score greater than t, N_c is the number of control wells, $E(t)$ is the number of experiment wells with score greater than t, and N_e is the total number of experiment wells.

In other words, this quantifies the expected number of experiment wells that pass the score threshold under the conservative assumption that all of the experiment wells were not true interactions, hence giving us an upper bound on the FDR. This bound is conservative because we expect that many of the experiment wells passing the score threshold will be true interactions. Because the number of control and experiment wells passing large thresholds decreases quickly, we smoothed our estimate of the FDR using a Generalized Pareto distribution to model the tail distribution of the control wells and of the experiment well scores separately. We then estimated the quantities $C(t)/N_c$ and $E(t)/N_e$ using the inferred Generalized Pareto distributions.

We used BioGRID [19] as our gold standard set of protein interactions. Our method resulted in substantially more overlap with the gold standard than previous approaches (Figure 4-5).

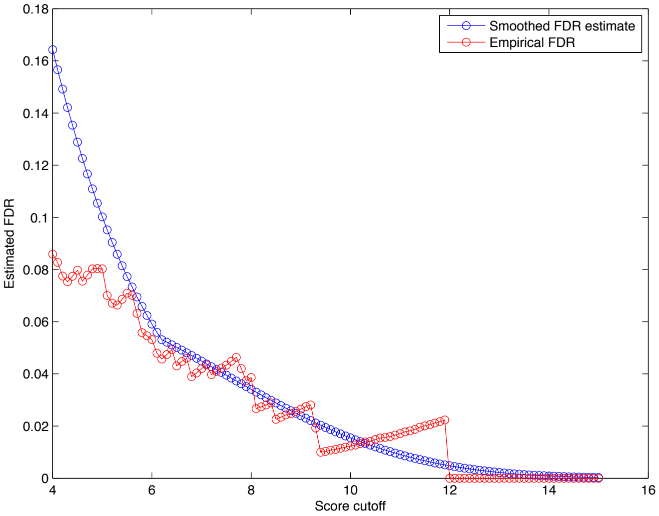


Figure 4-4: Estimation of the false discovery rate (FDR) for LUMIER. The red line indicates the empirical false discovery rate calculated based on control wells on each plate as described in the text. The blue line shows a smoothed fit. At LUMIER cutoff 7, the smoothed FDR is 0.044.

4.4.4 Results

We highlight some of the insights gained by mapping the interactions between chaperones, co-chaperones, and client proteins (Figure 4-6). Previous studies have primarily focused on the specificity of chaperone proteins, leaving co-chaperone specificity largely unexplored. Our systematic exploration of co-chaperone interactions identified highly specific connections between co-chaperones and particular biological processes: spindle assembly (BAG5 and MAD proteins), DNA replication (FKBP51 and the MCM complex), mRNA decapping (BAG4 and P bodies), retrograde signaling (NUDCD1 and COPI complex), and GPCR signaling (prefoldins and G protein γ subunits).

Co-chaperone interaction patterns revealed novel specificity for folding domains for a number of co-chaperones, in particular, for the poorly characterized NUDC family of co-chaperones. The evolutionarily related co-chaperones in the family recognize distinct but structurally homologous β-propeller domains.

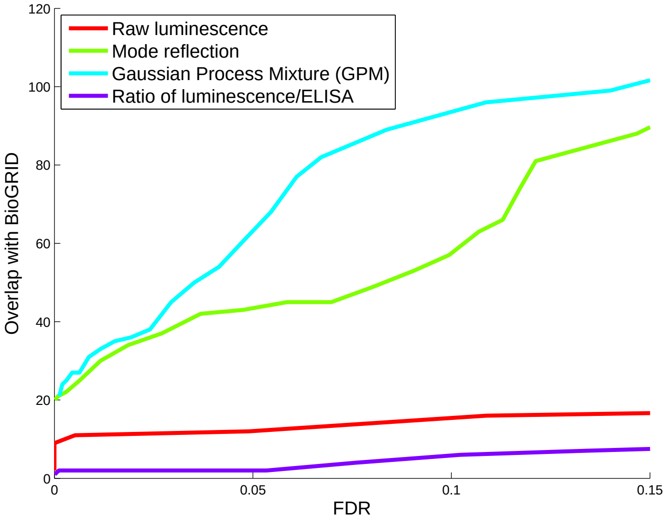


Figure 4-5: Comparison of methods for identifying significant interactions. Each method reports a list of significant interactions at various false discovery rate (FDR) cutoffs and we plot the number of significant interactions that overlap with BioGRID at these FDR cutoffs. Our method is plotted in cyan, green is a previous method [133], red is the raw readout, and purple is a simple manipulation of the raw readout.

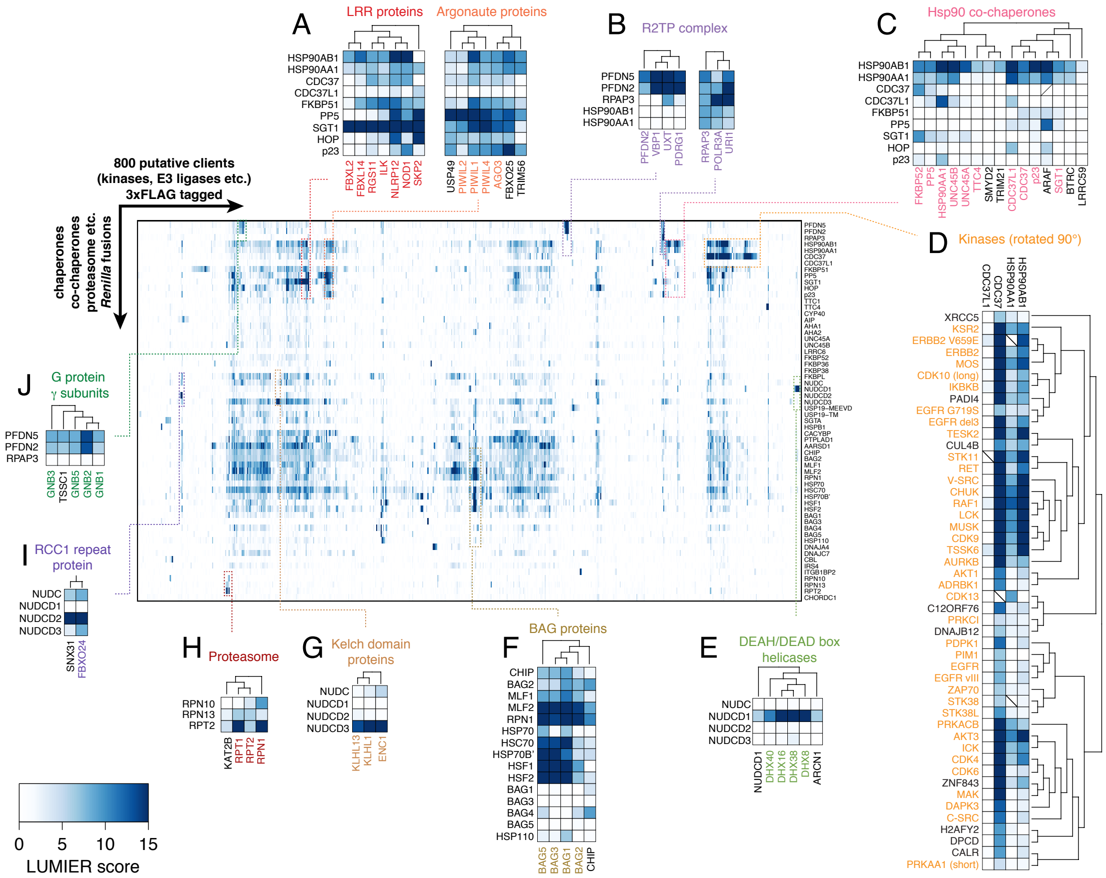


Figure 4-6: Quantitative view of the human protein folding landscape. 800 query proteins (arranged in columns) were assayed for interaction with 60 different chaperones, co-chaperones and quality control factors (rows) with a quantitative LUMIER assay. Query proteins were clustered based on their interaction profiles. Some of the biologically coherent clusters are highlighted in more detail. Proteins that share the same fold or are part of the same biological complex in each cluster are indicated in color. (A) LRR proteins (red) and Argonaute proteins (orange) form distinct clusters. LRR proteins interact strongly with SGT1, while Argonaute proteins associate with PP5. (B) The R2TP complex members (purple) forms two separate clusters. (C) Hsp90 co-chaperone cluster. (D) Kinases (orange) cluster together and interact specifically with CDC37 but not with CDC37L1. (E) NUDCD1 associates with DEAH/DEAD box helicases (green). (F) BAG proteins that cluster together interact strongly with Hsp70 proteins, Rpn1, Hsf1 and Hsf2. (G) Kelch domain protein cluster (brown) with NUDCD3. (H) Proteasome cluster. (I) RCC1 repeat protein FBXO24 (purple) interacts with NUDCD2. (J) G protein γ subunits (green) interact with prefoldins. From Taipale, Tucker, Peng, *et al.* "A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways" Cell, in press.

4.5 Conclusion

LUMIER presents an exciting opportunity to map interactions that may have been missed by standard assays. However, it suffers from systematic biases that can obfuscate true interactions. We introduced a statistical error model for LUMIER data that identifies many more true interactors than previous methods. We applied our method to a large LUMIER data set with 60 preys and 800 baits to construct an extensive map of chaperone, co-chaperone, and client protein interactions. This network revealed novel co-chaperone specificity and will provide an important resource for other researchers to understand the implications of disease perturbations on the protein homeostasis network.

Part II

Statistical genetics

Chapter 5

Mixed models with related individuals

Abstract

We calculate expected mixed model association statistics for samples with unrelated and related individuals. Surprisingly, our results suggest that standard mixed model association statistics may not be calibrated in samples with related individuals. Extensive simulations and analysis of data from the CARe cardiovascular consortium confirm that mixed models are inflated in a wide variety of relatedness regimes. We propose a two variance component mixed model that alleviates inflation and can improve power.

5.1 Introduction

Mixed models (MLM) are the state-of-the-art method for calculating association statistics in genome-wide association studies (GWAS). They are understood to correct for relatedness and population stratification at the same time as increasing power over linear regression [153]. In this chapter, we investigate the claim that mixed models correct for relatedness. Specifically, we expand upon [153] by providing an alternative derivation of the expected mixed model association statistic for unrelated individuals and then extend it to related individuals. MLM attempts to correct for relatedness by using an empirical estimate of kinship. Our theoretical results suggest that this may be insufficient. We investigate MLM statistics with related individuals through systematic simulations and find that MLM is inflated for a wide range of parameters. We propose a solution based on [160] and show that it substantially reduces the inflation in simulations and in tests with genotypes and phenotypes from the CARe consortium.

5.1.1 MLM statistics

First, we derive the MLM statistic for GWAS. We consider association testing with a quantitative trait, however future work might extend these results to the case-control setting via the liability threshold model [78]. For tractability, we consider the case where SNPs are unlinked. Additionally, for simplicity we assume there are no fixed effect covariates, however, all of these results naturally extend to the case with fixed effect covariates. Let N be the number of individuals in the study and M be the number of genotyped SNPs. The test SNP w and phenotype y are $N \times 1$ column vectors. The $N \times M$ genotype matrix W encodes the number of minor alleles (i.e., $0, 1, 2$) for each SNP and individual. For convenience, we'll assume that the genotype matrix has been normalized so that each SNP has mean 0 and variance 1.

The derivation of the MLM statistic starts by assuming a multivariate normal distribution for the phenotype, $y \sim N(w\beta, \Sigma)$ (for the moment we assume the covariance Σ is given, but we estimate it later). With a probabilistic model for y, we can calculate a Wald statistic to test the hypothesis that $\beta \neq 0$. When $\Sigma = I\sigma_e^2$, the standard linear regression Wald statistic hypothesis test is:

$$s_{linear} = \frac{\hat{\beta}^2}{\mathrm{var}(\hat{\beta})} = \frac{(w^T y)^2}{\sigma_e^2 w^T w},$$

which is distributed χ^2 with 1 degree of freedom (DOF) under the null hypothesis. When Σ is an arbitrary covariance matrix, we transform the phenotype to eliminate the correlation. Observe that if $\Sigma^{-1} = RR^T$, then $R^T y \sim N(R^T w\beta, I)$, so that the

MLM test statistic is:

$$s_{MLM} = \frac{(w^T R R^T y)^2}{w^T R R^T w} = \frac{(w^T \Sigma^{-1} y)^2}{w^T \Sigma^{-1} w}.$$

Now, we specify the covariance structure for y. We assume a linear model for the phenotype

$$y = w\beta + W\alpha + \epsilon,$$

where $\alpha \sim N(0, I\sigma_g^2/M)$ and $\epsilon \sim N(0, I\sigma_e^2)$. Marginalizing α gives

$$y \sim N(w\beta, \Sigma = WW^T\sigma_g^2/M + I\sigma_e^2).$$

σ_g^2 and σ_e^2 are typically estimated by restricted maximum likelihood (REML) [96]. REML adjusts for the loss in degrees of freedom due to fixed effect covariates and produces unbiased estimates of the variance parameters.

Standard linear algebra identities allow us to view the statistic from an alternative perspective. Using the Woodbury identity,

$$\Sigma^{-1} = \frac{1}{\sigma_e^2}\left(I - W\left(I\frac{\sigma_e^2}{\sigma_u^2} + W^T W\right)^{-1} W^T\right),$$

where $\sigma_u^2 = \sigma_g^2/M$. Define $R = W\left(I\frac{\sigma_e^2}{\sigma_u^2} + W^T W\right)^{-1} W^T$ as the ridge operator, in the sense that if $z \sim N(W\beta, \sigma_e^2)$ and $\beta \sim N(0, \sigma_u^2 I)$, $Rz = z_{ridge}$ is the MAP estimate or ridge regression estimate of z under this model. Thus, $\Sigma^{-1}z = \frac{1}{\sigma_e^2}(z - Rz) = \frac{1}{\sigma_e^2}(z - z_{ridge})$.

Using this notation, the MLM statistic can be rewritten as

$$s_{MLM} = \frac{1}{\sigma_e^2}\frac{(y^T(w - w_{ridge}))^2}{w^T(w - w_{ridge})} = \frac{1}{\sigma_e^2}\frac{(w^T(y - y_{ridge}))^2}{w^T(w - w_{ridge})},$$

which highlights the close connection between mixed models and ridge regression. In fact, the denominator is approximately constant in large data sets [131], so up to scaling s_{MLM} can be thought of as a linear regression Wald statistic on the ridge regression residual $(y - y_{ridge})$ instead of the phenotype (y).

In the derivations above, w was not included in Σ, so naively computing s_{MLM} would require computing and inverting a new Σ for each test SNP. A natural question arises: is it necessary to remove w from Σ? Empirically, it has been shown that including w in Σ reduces power [84]. If we include w in Σ, then we assume that $y \sim N(w\beta, ww^T\sigma_u^2 + \Sigma)$, so the MLM statistic including w (MLMi) is

$$\begin{aligned}
s_{MLMi} &= \frac{(w^T(ww^T\sigma_u^2 + \Sigma)^{-1}y)^2}{w^T(ww^T\sigma_u^2 + \Sigma)^{-1}w} \\
&= \frac{(w^T\Sigma^{-1}y)^2}{w^T\Sigma^{-1}w}\frac{1}{1 + w^T\Sigma^{-1}w\sigma_u^2} \\
&= s_{MLM}\frac{1}{1 + w^T\Sigma^{-1}w\sigma_u^2},
\end{aligned}$$

by applying the Woodbury identity. Σ is positive definite, so $\frac{1}{1+w^T\Sigma^{-1}w\sigma_u^2} < 1$, consistent with the empirical results that including w in Σ reduces power.

5.1.2 Expected statistics with unrelated individuals

Now, let us assume that the phenotypes y are actually generated from a mixed model, that is $y \sim N(w\beta, \Sigma)$. Plugging this value into the statistics gives the following after taking expectations with respect to the randomness from the normal distribution

$$s_{linear} = \frac{w^T\Sigma w}{N} + N\beta^2$$

$$s_{MLM} = 1 + w^T\Sigma^{-1}w\beta^2$$

$$s_{MLMi} = s_{MLM}\frac{1}{1 + w^T\Sigma^{-1}w\sigma_u^2},$$

using the fact that w was normalized to variance 1. The assumption that individuals are unrelated means that $\mathrm{Cov}(w)$ and $\mathrm{Cov}(W)/M$ are the identity. Assuming w and the columns of W are mutually independent (the unlinked SNPs assumption) and taking expectations, we see that

$$s_{linear} = 1 + N\beta^2$$

	Linear	MLM	MLMi
Exact	$\frac{w^T \Sigma w}{N} + N\beta^2$	$1 + w^T \Sigma^{-1} w \beta^2$	$\left(1 + w^T \Sigma^{-1} w \beta^2\right) \frac{1}{1 + w^T \Sigma^{-1} w \sigma_u^2}$
$\approx E[\cdot]$	$1 + N\beta^2$	$1 + F\beta^2$	$(1 + F\beta^2) \frac{1}{1 + F\sigma_u^2}$

Table 5.1: Wald statistics under data generated from MLM model, where F is given by Eq. 5.1.

and

$$s_{MLM} = 1 + E[trace(\Sigma^{-1})]\beta^2.$$

We can approximate the expectation of the trace using the Marchenko-Pastur distribution for the eigenvalues of a Wishart matrix [87]. First, we approximate

$$\frac{E[w^T \Sigma^{-1} w]}{N} = \frac{E[trace(\Sigma^{-1})]}{N} \approx \int \frac{\nu(x)}{x\sigma_g^2 + \sigma_e^2} dx$$

where ν is the eigenvalue density for a Wishart matrix (i.e. XX^T/M where X is a matrix whose entries are independent standard normals). ν is known in closed form when $N, M \to \infty$ at a finite ratio $N/M \in (0, 1]$, so we can evaluate the integral with this asymptotic density. This gives

$$E[trace(\Sigma^{-1})] \approx \frac{N}{2r\sigma_g^2} \left(-1 - \frac{\sigma_g^2}{\sigma_e^2}(1 - r) + \sqrt{\left(1 + \frac{\sigma_g^2}{\sigma_e^2} a\right)\left(1 + \frac{\sigma_g^2}{\sigma_e^2} b\right)} \right) \tag{5.1}$$

where $r = N/M, a = (1 + \sqrt{r})^2, b = (1 - \sqrt{r})^2$. If we replace the $w^T \Sigma^{-1} w$ terms with this expression, then we recover the results from [153]. Yang *et al.* show that this approximation is accurate and predicts that as N increases, the statistical power of mixed models can be much larger than linear regression [153].

5.1.3 Expected statistics with related individuals

The covariance matrix of the SNPs reflects the relatedness or pedigree structure of the individuals (e.g., siblings share half their of genomes on average, so the covariance between their normalized SNP vectors will be 0.5 on average). In particular, let $E[ww^T] = E[WW^T/M] = \theta$ denote the covariance structure. Now, the unlinked assumption means that the w and the columns of W are independent given θ. With related individuals, it will be important to model untyped SNPs as well as typed SNPs. So, we model the phenotype as

$$y = w\beta + W\alpha + U\gamma + \epsilon,$$

where $\gamma \sim N(0, I\sigma_h^2/M_h)$, M_h is the number of untyped or hidden causal SNPs, and $E[UU^T/M_h] = \theta$. In this case, after taking the expectation with respect to randomness in y, the MLM statistic is

$$\frac{w^T\Sigma^{-1}(WW^T\sigma_g^2/M + UU^T\sigma_h^2/M_h + I\sigma_e^2)\Sigma^{-1}w}{w^T\Sigma^{-1}w} + w^T\Sigma^{-1}w\beta^2.$$

Taking the expectation with respect to U, we get

$$\frac{w^T\Sigma^{-1}(WW^T\sigma_g^2/M + \theta\sigma_h^2 + I\sigma_e^2)\Sigma^{-1}w}{w^T\Sigma^{-1}w} + w^T\Sigma^{-1}w\beta^2.$$

Taking the expectation with respect to w gives

$$\begin{aligned} E&\left[\frac{w^T\Sigma^{-1}(WW^T\sigma_g^2/M + \theta\sigma_h^2 + I\sigma_e^2)\Sigma^{-1}w}{w^T\Sigma^{-1}w}\right] + E[w^T\Sigma^{-1}w]\beta^2 \\ &\approx \frac{E[w^T\Sigma^{-1}(WW^T\sigma_g^2/M + \theta\sigma_h^2 + I\sigma_e^2)\Sigma^{-1}w]}{E[w^T\Sigma^{-1}w]} + E[w^T\Sigma^{-1}w]\beta^2 \\ &= \frac{\mathrm{trace}(\Sigma^{-1}(WW^T\sigma_g^2/M + \theta\sigma_h^2 + I\sigma_e^2)\Sigma^{-1}\theta)}{\mathrm{trace}(\Sigma^{-1}\theta)} + \mathrm{trace}(\Sigma^{-1}\theta)\beta^2. \end{aligned}$$

When $\beta = 0$, the statistic should be χ^2 distributed with 1 degree of freedom, so the mean value should be exactly 1. This form of the Wald statistic suggests that inflation or deflation in the mean of the Wald statistic is due to mis-estimating the covariance of y. For example, suppose we are using the linear regression statistic where $\Sigma = I\hat{\sigma_e^2} \approx I(\sigma_g^2 + \sigma_h^2 + \sigma_e^2)$, a clear mis-estimate of the covariance of y. Then, the expected statistic is approximately

$$\frac{\mathrm{trace}(WW^T\sigma_g^2/M + \theta\sigma_h^2 + I\sigma_e^2)\theta)}{\hat{\sigma_e^2}\,\mathrm{trace}(\theta)} + \frac{\mathrm{trace}(\theta)}{\hat{\sigma_e^2}}\beta^2,$$

which after taking the expectation with respect to W and approximating the expectation of a ratio with the ratio of expectations

$$\begin{aligned}\frac{\mathrm{trace}(\theta^2)(\sigma_g^2+\sigma_h^2)+N\sigma_e^2}{N(\sigma_g^2+\sigma_h^2+\sigma_e^2)} + &\frac{N}{\sigma_g^2+\sigma_h^2+\sigma_e^2}\beta^2 \\ &= \sum_{i\neq j}\theta_{ij}^2\frac{\sigma_g^2+\sigma_h^2}{N(\sigma_g^2+\sigma_h^2+\sigma_e^2)} + 1 + \frac{N}{\sigma_g^2+\sigma_h^2+\sigma_e^2}\beta^2 \\ &= NS\frac{(\sigma_g^2+\sigma_h^2)}{(\sigma_g^2+\sigma_h^2+\sigma_e^2)} + 1 + \frac{N}{\sigma_g^2+\sigma_h^2+\sigma_e^2}\beta^2,\end{aligned}$$

where $S = \sum_{i\neq j}\theta_{ij}^2/N^2$ measures the relatedness in the dataset. As a result of the relatedness, the statistic is inflated by $NS\frac{(\sigma_g^2+\sigma_h^2)}{(\sigma_g^2+\sigma_h^2+\sigma_e^2)}$. This is consistent with a more general analysis of linear regression Wald statistics for linked markers[1].

Previous studies have shown that when σ_g^2 is estimated in samples with related individuals, $\sigma_g^2 \leq \hat{\sigma_g^2} \leq \sigma_g^2 + \sigma_h^2$. In this case, Σ may not match the covariance of y, potentially resulting in miscalibrated statistics. Inspired by [160], we propose introducing a second variance component in Σ. Instead of using $\Sigma = WW^T\sigma_g^2/M + I\sigma_e^2$, we estimate the covariance structure by $\Sigma = WW^T\sigma_g^2/M + \hat{\theta}\sigma_h^2 + I\sigma_e^2$ where θ is estimated by retaining the entries of WW^T/M above a threshold. Recovering θ can be viewed as a sparse covariance estimation problem, where thresholding approaches have been shown to produce good results [9]. We expect that the two variance component mixed model statistics will not be inflated in samples with related individuals because it more closely models the covariance structure of y. In the following sections, we describe the proposed two variance component MLM statistic in detail and describe the results of extensive simulations and applications to real genotypes and phenotypes from the CARe consortium.

[1] Hilary Finucane, personal communication.

5.2 Results

5.2.1 Simulated genotypes and phenotypes

We conducted extensive simulations with randomly-generated genotypes and phenotypes to understand inflation and power for mixed model association statistics with related individuals. We systematically varied the number of related individuals, the degree of relatedness, the number of markers in the genome, and the heritability of the trait. Specifically, we simulated 1000 individuals, where some pairs of individuals were related (50, 125, 250, and 500 pairs) and the rest of the individuals were unrelated (leaving 900, 750, 500, and 0 unrelated individuals, respectively). The pairs of individuals shared between 0 and 0.5 of their genomes in expectation. Additionally, we varied the number of markers (1,000, 5,000, 10,000, and 20,000 SNPs) and generated unlinked markers for simplicity. To simulate markers, we randomly generated minor allele frequencies uniformly in $[0.05, 0.5]$ and sampled genotypes from a binomial distribution. For pairs of related individuals and for each haplotype, with probability equal to the relatedness, the pair shared an allele drawn randomly, otherwise the alleles for the pair were drawn independently. We generated 100 candidate causal SNPs and 500 candidate null SNPs for testing in the same way. We used an infinitesimal model to generate the phenotype. In particular, we generated effect sizes for the observed SNPs from $N(0, h_g^2/M)$ where M is the number of SNPs in the simulation. We also generated effect sizes for the candidate causal test SNPs from $N(0, (h^2 - h_g^2)/100)$. Because the model does not include the candidate causal test SNPs, these SNPs effectively served as untyped causal loci. Finally, we formed the phenotype by multiplying the effects with the genotypes and adding independent noise distributed as $N(0, (1 - h^2)I)$.

From the definition of the Wald statistic for mixed models, under the null hypothesis (i.e., the probability model for y when $\beta = 0$), the statistic is distributed as a scaled χ^2 with 1 degree of freedom assuming that Σ is fixed. In theory, Σ depends on y and the scaling constant depends on the test SNP, however, for large sample sizes, Σ is well estimated and the constant does not depend strongly on the identity of the test SNP. This justifies measuring inflation by the mean Wald statistic on the null SNPs and measuring power by the mean Wald statistic on the causal SNPs divided by the mean Wald statistic on the null SNPs.

Contrary to the belief that mixed models correct for relatedness [153], we found that for many parameter settings, the mixed model statistic is significantly inflated (Figure 5-1) and in some cases, the inflation is quite substantial (e.g., mean Wald statistic 1.10 ± 0.01 with 500 sib-pairs, $N/M = 1$, $h^2 = 0.75, h_g^2 = 0.4$), whereas the two variance component mixed model alleviates the inflation (mean Wald statistic 1.01 ± 0.01 for the same simulation). Unsurprisingly, as we increased the relatedness (either by increasing the number of related individuals or the strength of relatedness), the inflation grew. As we increased the ratio individuals to markers, the inflation increased as well, suggesting that as sample sizes increase in real data sets, relatedness may pose a significant challenge. For phenotypes with moderate heritability ($h^2 = 0.5$), we found no substantial power difference between the one and two variance component mixed models. For phenotypes with larger heritability $h^2 = 0.75$, we found that the two variance component model increased power as relatedness increased (at most 3% and 6% improvements when $h_g^2 = 0.4$ and $h_g^2 = 0.25$, respectively). In all cases, the two variance component model alleviated inflation and maintained or increased power compared to the standard MLM (Figure 5-2).

To investigate the effects of using the thresholded estimator for θ, we performed simulations using a two variance component model that used the true pedigree matrix. We found that there were not substantial differences between using the thresholded estimator or the true pedigree matrix (Figure 5-2).

5.2.2 CARe genotypes

Next we explored simulations with real genotypes from the CARe cardiovascular consortium. The dataset includes samples from $8,367$ African-American individuals. After QC filtering (described in [79]), $770,390$ SNPs remained. Because the individuals were admixed, all subsequent analysis factored out the first 5 principal components to avoid confounding from population structure. To avoid problems due

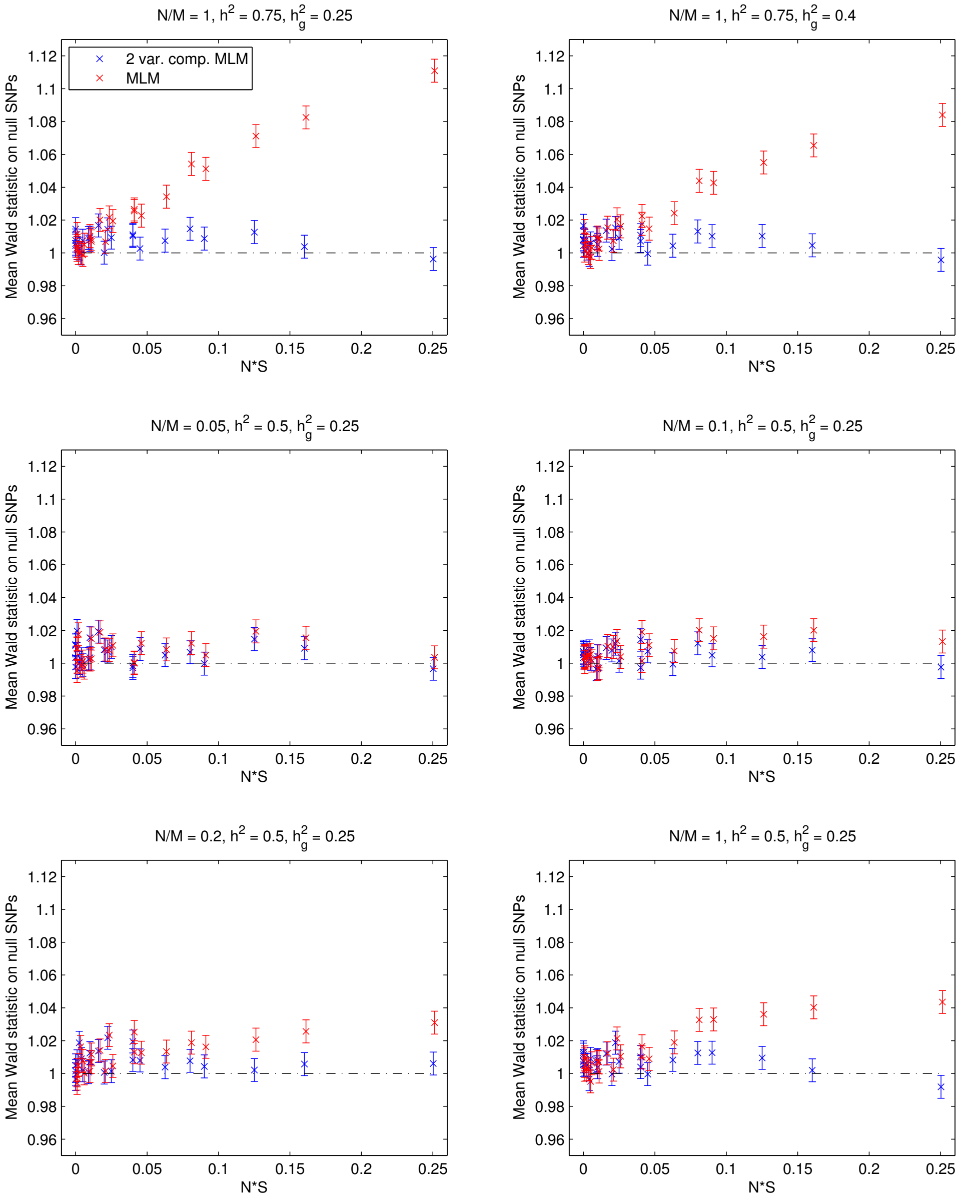


Figure 5-1: Calibration for mixed models. We plot the mean Wald statistic on null SNPs for the standard mixed model (MLM) and the two variance component model (2 var. comp. MLM) against $N * S$, where $S = \sum_{i \neq j} \theta_{ij}^2 / N^2$ measures the amount of relatedness in the data. Points are the mean over 100 simulations and standard errors were $\approx \pm 0.007$. We varied N/M, h^2, and h_g^2.

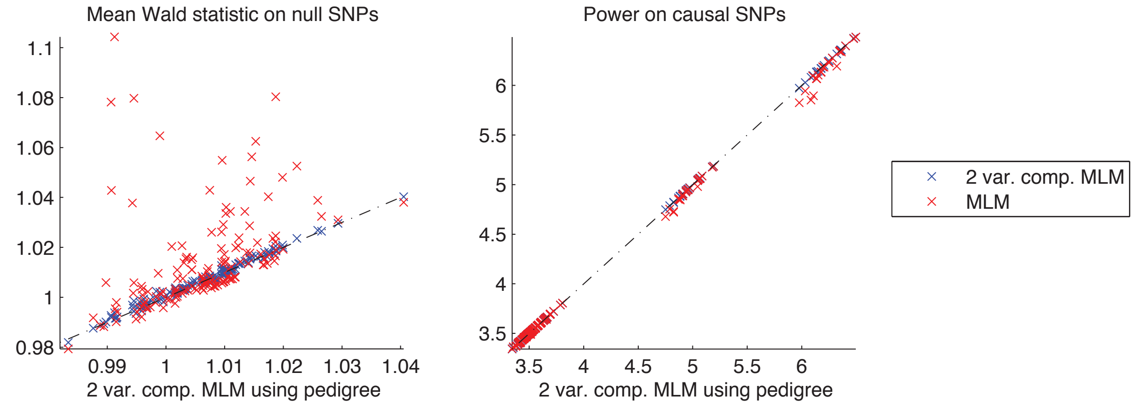


Figure 5-2: Inflation and power for mixed models. We plot inflation (left) and power (right) for the standard mixed model (MLM) and the two variance component model (2 var. comp. MLM) against a two variance component model that uses the true pedigree matrix (θ) instead of the thresholded estimator. Each point is the mean over 50 simulations. The thresholded estimator performs nearly identically to using the true relatedness matrix, whereas the standard MLM inflates statistics in many cases and can have decreased power.

to linkage between the candidate test SNPs and the GRM SNPs [84], we used SNPs from chromosome 1 as candidate causal SNPs and SNPs from chromosome 2 as candidate null SNPs. We explored varying ratios of N/M by reducing the number of observed SNPs. To generate phenotypes, we generated effect sizes for the observed SNPs from $N(0, h_g^2/M)$. Then, we randomly selected 250 candidate causal SNPs from chromosome 1 and generated effect sizes from $N(0, (h^2 - h_g^2)/250)$. Finally, we multiplied the effect sizes with the genotypes and added noise from $N(0, (1 - h^2)I)$.

Consistent with the previous simulation, the standard mixed model inflated statis-

Observed SNPs	# of SNPs	MLM	2 var. comp. MLM	threshold (t)
Chrom. 3 - 22	615,445	1.013 (0.002)	1.000 (0.002)	0.0239843
Chrom. 3 - 6	195,333	1.024 (0.002)	1.002 (0.002)	0.0505165
Chrom. 3 - 4	99,690	1.028 (0.002)	1.003 (0.002)	0.0807943
Chrom. 22	9,713	1.036 (0.002)	1.014 (0.002)	0.387328

Table 5.2: Mean Wald statistics on candidate null SNPs for simulations with CARe genotypes. Mean values over 100 simulations are reported with standard error in parenthesis. The two variance component model used the specified threshold to estimate the relatedness matrix.

tics and the two variance component model alleviated inflation (Table 5.2). Importantly, these results suggest that the levels of relatedness that are required for inflation are present in typical data sets. In the last simulation, where only markers from chromosome 22 are observed, the two variance component model appears to be inflated. Given the large threshold chosen (> 0.38), we hypothesize that the number of markers was too small to distinguish relatedness in the data from noise in the GRM, causing an incomplete correction. Finally, we found that the power of the two variance component model was similar and at least as great as the standard mixed model in all cases.

5.2.3 CARe phenotypes

Finally, we calculated mixed model statistics for the CARe phenotypes: body mass index (BMI), height, low density lipoprotein cholesterol (LDL), and high density lipoprotein cholesterol (HDL) (Table 5.3). We adjusted for age, sex, and the top 5 PCs. Because we do not know the causal and null SNPs, we calculated the average Wald statistic over all SNPs in a leave-one-chromosome-out fashion; noting that we expect the statistics to be larger than 1 due to polygenicity [153]. The average Wald statistics are higher for standard mixed models than the two variance component model, consistent with the previous simulations. This suggests that mixed model association statistics calculated on the CARe data using standard mixed models are slightly inflated.

Phenotype	N	MLM	2 var. comp. MLM	h^2_{pseudo}	h^2_g	h^2
BMI	8148	1.025	1.039	0.37	0.20	0.45
height	8148	1.050	1.060	0.40	0.29	0.43
LDL	5311	1.018	1.028	0.35	0.18	0.47
HDL	5031	1.034	1.051	0.46	0.23	0.62

Table 5.3: Mean Wald statistics over all SNPs for MLM and two variance component MLM. We list the number of individuals N with recorded phenotypes. We also list h^2_{pseudo}, the estimate of heritability using MLM, which is known to be inflated in samples with related individuals [160], and the estimates of heritability from the two variance component model.

5.3 Statistical Methods

In this section, we describe the MLM and two variance component MLM statistics.

5.3.1 MLM statistics

We mean centered the phenotype y, covariates X, and genotypes W. Additionally, we normalized each genotype by dividing by $\sqrt{2\hat{p}(1-\hat{p})}$ where $\hat{p}$ is the estimated minor allele frequency. Then the phenotype is modeled as

$$y = Xb + W\alpha + \epsilon,$$

where $\alpha \sim N(0, \sigma_g^2/M), \epsilon \sim N(0, \sigma_e^2 I)$, and b is a vector of weights for the covariates. This model naturally leads to an association statistic based on the Wald statistic.

To calculate the association statistic for SNP w, we added w as a fixed effect covariate to the previous model and tested whether its coefficient is significantly different than 0. Specifically, consider the model

$$y = w\beta + Xb + W\alpha + \epsilon,$$

where β is the coefficient for the test SNP. We estimated σ_g^2 and σ_e^2 by REML. The fixed effect coefficients (β, b) are estimated by maximum likelihood.

It is straightforward to construct the Wald statistic to test whether $\beta \neq 0$. Let $V = \hat{\sigma_g^2} WW^T/M + \hat{\sigma_e^2} I$ and $Q = [w; X]$. Then $\hat{\beta}$ is equal to the first entry of $(Q^T V^{-1} Q)^{-1} Q^T V^{-1} y$ and $\text{var}(\hat{\beta})$ is equal to the first entry of $(Q^T V^{-1} Q)^{-1}$. The test statistic is

$$\frac{\hat{\beta}^2}{\text{var}(\hat{\beta})},$$

which is χ^2 distributed with 1 degree of freedom under the null distribution.

To avoid proximal contamination [84], we used a leave-one-chromosome-out procedure [153]. For each test SNP w (which is not necessarily in W), we excluded the chromosome including that SNP from the genotypes used to calculate the GRM and calculated the Wald statistic for w with this GRM. We did this efficiently be precomputing and storing the GRM excluding each chromosome in turn.

5.3.2 Two variance component MLM statistics

In the two variance component model, we estimated θ by $(WW^T/M)_{>t}$ where $(\cdot)_{>t}$ denotes the matrix where entries less than or equal to t have been set to 0. Here and in other places where we formed a GRM, we projected out the first 5 principal components to avoid confounding from ancestry [100]. We optimized t to reproduce the covariance structure of the test SNPs. Specifically, let Z be the matrix of test SNPs (i.e., the SNPs on the chromosome we are testing and as result W is the matrix of SNPs on all other chromosomes). We set t to the minimizer of

$$||ZZ^T/M_z - (WW^T/M)_{>t}||_2^2$$

where M_z is the number of SNPs in Z and $||\cdot||_2$ is the Frobenius norm. Then under the phenotype model

$$y = N(w\beta + Xb, WW^T/M\sigma_g^2 + (WW^T/M)_{>t}\sigma_h^2 + I\sigma_e^2)$$

we estimated σ_g^2, σ_h^2, and σ_e^2 by REML. Then we proceeded as in the MLM with V now equal to

$$V = WW^T/M\hat{\sigma_g^2} + (WW^T/M)_{>t}\hat{\sigma_h^2} + I\hat{\sigma_e^2}.$$

5.4 Conclusion

Through extensive simulations and tests on CARe genotypes and phenotypes, we showed that standard mixed models can be miscalibrated in a wide range of relatedness settings. For current sample sizes and levels of relatedness, the inflation is small. However, our simulations suggest that as sample sizes increase, relatedness will play a larger role in inflating test statistics. The two variance component model effectively alleviates the inflation at no cost to power and in some cases can increase power over standard mixed models. Because of this, we expect the two variance component model to become increasingly relevant as sample sizes increase.

Chapter 6

Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select

Abstract

[1]Using a reduced subset of SNPs in a linear mixed model can improve power for genome-wide association studies, yet this can result in insufficient correction for population stratification. We propose a hybrid approach using principal components that does not inflate statistics in the presence of population stratification and improves power over standard linear mixed models.

6.1 Introduction

In recent years, there has been extensive research on linear mixed models (LMM) to calculate genome-wide association study (GWAS) association statistics [67, 66, 117, 168, 131, 153]. While linear mixed models implicitly assume that all SNPs have an effect on the phenotype (an infinitesimal genetic architecture), it is widely believed

[1] This chapter is adapted from "Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select" by George Tucker, Alkes L. Price, and Bonnie Berger appearing in Genetics [140].

that disease phenotypes do not follow an infinitesimal model and that modeling a genetic architecture where most SNPs have negligible effect and some have modest effect (a non-infinitesimal genetic architecture) would increase power. As a step in that direction, Listgarten et al. [84, 83] recently developed the state-of-the-art FaST-LMM Select method, which constructs a genetic relationship matrix (GRM) from a subset of top associated SNPs that are more likely to be causal. However, as a recent Perspective paper [153] shows, limiting the GRM to a subset of SNPs can result in insufficient correction for population stratification, leading to significantly inflated statistics and false positive associations (Tables 6.1, 6.2 and Appendix B.1).

As a solution to this problem, we propose PC-Select, a novel hybrid approach that includes the principal components (PCs) of the genotype matrix as fixed effects in FaST-LMM Select. PC-Select leverages the advantages of the FaST-LMM Select framework while correcting for population stratification. The two main steps of FaST-LMM Select are ranking SNPs by linear regression p-values to form the GRM with the top ranked SNPs and then calculating association statistics in a mixed model framework using this GRM. We used the top 5 PCs[2] as fixed effects in both of these steps (See **Methods**). As a result, PC-Select yields non-inflated test statistics in the presence of population stratification and maintains high power to detect causal SNPs.

6.2 Results

To examine inflation and power, we followed the simulation procedure in [153] and generated data sets each containing 10,000 SNPs for 1,000 individuals. To avoid a loss in power for LMM that can occur when candidate SNPs are included in the GRM[3] [84, 153], we separately simulated a set of candidate SNPs to compute test statistics. We sampled individuals from two populations with $F_{st} = 0.05$, ancestral

[2] We follow the recommendations in the literature [100] and use a fixed number of PCs. We have found that 5 PCs is generally sufficient to correct for stratification in simulated and real data sets. Alternatively, the number of PCs may be selected through cross validation or Tracy-Widom statistics [97].

[3] Both PC-Select and FaST-LMM Select avoid this by removing the candidate SNP and nearby SNPs from the GRM when computing the association statistic.

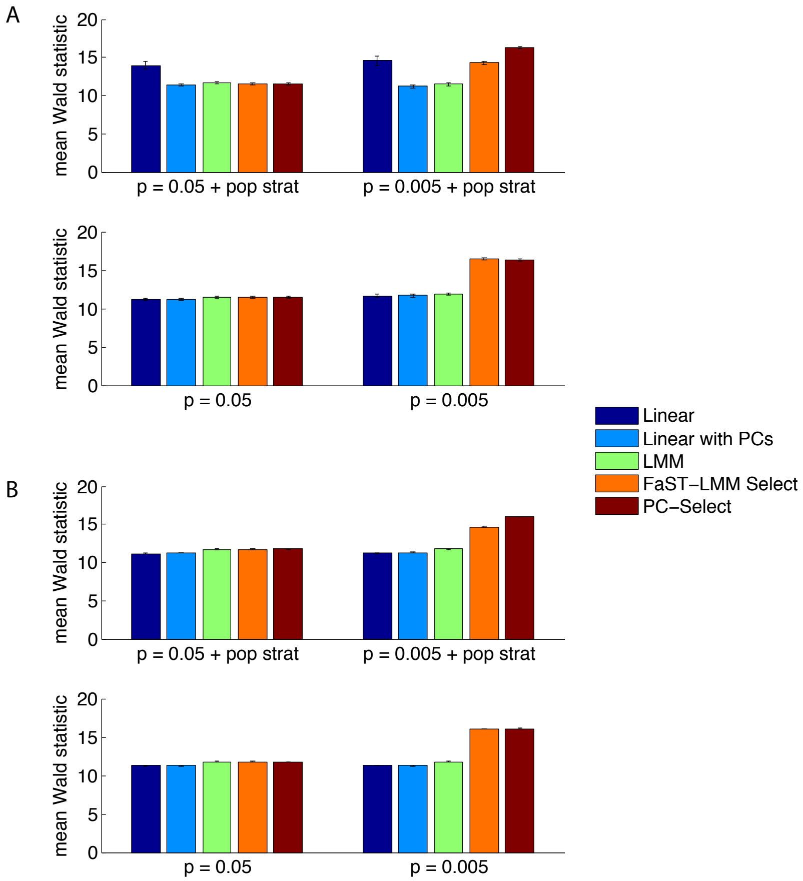


Figure 6-1: Comparison of power for linear regression, linear regression with PCs, standard LMM, FaST-LMM Select, and PC-Select on simulated genotypes and phenotypes (A) and real genotypes and simulated phenotypes (B) with and without population stratification as the fraction of casual SNPs ($p = 0.05, 0.005$) varies. To measure power, we plot the mean Wald statistic on test causal SNPs. In all cases, PC-Select has the highest power of the methods that do not inflate statistics.

minor allele frequencies uniform in $[0.1, 0.5]$, and mean phenotypic difference 0.25 standard deviations. To simulate causal SNPs in the GRM, we selected a fraction $p = 0.05$ or 0.005 of the SNPs at random and sampled Gaussian effect sizes (variance equal to 0.5 divided by the number of casual SNPs in the GRM) for these SNPs. We generated 500 candidate test null SNPs that were not causal, and to measure inflation we calculated λ_{GC}, the median Wald statistic on these SNPs divided by the theoretical median under the null distribution [31]. To investigate power, we generated 50 causal candidate SNPs with normally distributed effect sizes (variance equal to 0.5 divided by the number of causal candidate SNPs) and measured mean Wald statistic on these SNPs. We split the variability from causal SNPs evenly between the GRM and the causal candidate SNPs. We repeated all simulations 100 times and report the mean and standard error.

mean λ_{GC} (std. error)	pop. strat. $p = 0.05$	pop. strat. $p = 0.005$	$p = 0.05$	$p = 0.005$
Linear regression	3.8 (0.4)	4.5 (0.5)	1.01 (0.01)	1.01 (0.01)
Linear reg. with PCs	1.02 (0.01)	1.03 (0.01)	1.01 (0.01)	1.02 (0.01)
LMM	1.01 (0.01)	1.02 (0.01)	1.01 (0.01)	1.01 (0.01)
FaST-LMM Select	1.04 (0.01)	1.26 (0.03)	1.01 (0.01)	0.99 (0.01)
PC-Select	1.01 (0.01)	1.01 (0.01)	1.01 (0.01)	0.99 (0.01)

Table 6.1: Extent of null statistic inflation as measured by λ_{GC} (median Wald statistic on test null SNPs divided by the theoretical median under the null distribution [31]). We tabulate λ_{GC} for linear regression, linear regression with PCs, standard LMM, FaST-LMM Select, and PC-Select on simulated genotypes and phenotypes with and without population stratification as the fraction of casual SNPs ($p = 0.05, 0.005$) varies. Values shown are mean λ_{GC} over 100 simulations with standard error in parenthesis. FaST-LMM Select inflates statistics in the presence of population stratification when few SNPs are causal ($p = 0.005$), which may result in false positives.

We found that when few SNPs were causal ($p = 0.005$), FaST-LMM Select inflated null statistics in the presence of population stratification ($\lambda_{GC} = 1.26 \pm 0.03$), whereas PC-Select was properly calibrated ($\lambda_{GC} = 1.01 \pm 0.01$) (Table 6.1). Moreover, FaST-LMM Select lost power in the presence of population stratification (as measured by the mean Wald statistic on causal SNPs: 14.3 ± 0.2 with stratification versus 16.4 ± 0.1 without), whereas PC-Select's power in simulations with and without population stratification was not significantly different (16.3 ± 0.1 versus 16.3 ± 0.1) (Figure 6-1). Thus, even though PC-Select corrected for stratification, this advantage did not come at the expense of power. This gain is likely because the PCs reduce noise in selecting subsets of SNPs for the GRM in the presence of population stratification. In addition, PC-Select chose fewer SNPs than FaST-LMM Select to include in the GRM (over 100 simulations, mean SNPs chosen: $\sim$20 versus $\sim$240, Figure B-4), yielding potential computational savings. When many SNPs were causal ($p = 0.05$), both methods used nearly all SNPs in the GRM (over 100 simulations, mean SNPs chosen:

~9,400 and ~8,800 out of 10,000, respectively), achieving similar performance to standard LMM.

We also investigated a recent extension of FaST-LMM Select, the *genard* method [57] that fits a data-adaptive low-rank GRM; however, we found that it did not have increased power over LMM in our simulations (Figure B-5), which is consistent with previous simulations in a similar context [57].

Next, we evaluated inflation and power on real genotypes with simulated phenotypes in a similar manner. We analyzed 5,000 individuals randomly subsampled from a multiple sclerosis (MS) study genotyped on Illumina arrays [114] made available via Welcome Trust Case Control Consortium 2 (WTCCC2) (See **Methods**). As before, we separated GRM SNPs and candidate SNPs to avoid proximal contamination and provide a fair comparison of methods. We randomly sampled 50,000 SNPs for the GRM from chromosomes 3 to 22, 250 causal SNPs from chromosome 1, and 500 null SNPs from chromosome 2. To simulate environmental variance aligned with population structure, we added 0.25 times the first PC (after the PC had been normalized to variance 1) to each individual's phenotype. Otherwise, we generated phenotypes as before and report simulations over 200 randomly generated phenotypes.

mean λ_{GC} (std. error)	pop. strat. $p = 0.05$	pop. strat. $p = 0.005$	$p = 0.05$	$p = 0.005$
Linear regression	1.58 (0.02)	1.55 (0.02)	1.03 (0.01)	1.04 (0.01)
Linear reg. with PCs	1.01 (0.01)	1.00 (0.01)	1.01 (0.01)	1.02 (0.01)
LMM	1.02 (0.01)	1.01 (0.01)	1.00 (0.01)	1.02 (0.01)
FaST-LMM Select	1.02 (0.01)	1.06 (0.01)	1.00 (0.01)	1.02 (0.01)
PC-Select	1.01 (0.01)	1.01 (0.01)	1.00 (0.01)	1.01 (0.01)

Table 6.2: Extent of null statistic inflation as measured by λ_{GC}. We tabulate λ_{GC} for linear regression, linear regression with PCs, standard LMM, FaST-LMM Select, and PC-Select on real genotypes and simulated phenotypes with and without population stratification as the fraction of casual SNPs ($p = 0.05, 0.005$) varies. Values shown are mean λ_{GC} over 200 simulations with standard error in parenthesis. FaST-LMM Select inflates statistics in the presence of population stratification when few SNPs are causal ($p = 0.005$), which may result in false positives.

We again found that when few SNPs were causal ($p = 0.005$), FaST-LMM Select inflated null statistics in the presence of population stratification ($\lambda_{GC} = 1.06 \pm 0.01$), whereas PC-Select was properly calibrated ($\lambda_{GC} = 1.01 \pm 0.01$) (Table 6.2). Moreover, FaST-LMM Select lost power in the presence of population stratification (as measured by the mean Wald statistic on causal SNPs: 14.64 ± 0.05 with stratification versus 16.02 ± 0.05 without); in contrast, PC-Select's power in simulations with and without population stratification was not significantly different (16.02 ± 0.05 versus 16.08 ± 0.05) (Figure 6-1). In all of our simulations, PC-Select produced non-inflated statistics and high power.

Finally, we analyzed data from 10,204 MS cases and 5,429 controls genotyped on Illumina arrays [114] made available via WTCCC2 (See **Methods**). The cases and controls were not matched for ancestry and thus exhibited substantial population stratification. Evaluated over all SNPs, PC-Select had $\lambda_{GC} = 1.24$ and FaST-LMM Select had $\lambda_{GC} = 1.20$. Due to polygenicity, we expect λ_{GC} on all markers to be larger than 1. On the same data, [153] report $\lambda_{GC} = 1.23$ and 1.20 for linear regression with PCs and LMM, respectively, which they show is consistent with polygenicity. To evaluate power, we considered Wald statistics at 75 known associated SNPs (See **Methods**, Table B.1 for Wald statistics). PC-Select consistently gave larger Wald statistics than FaST-LMM Select (63 of 75 markers, $P = 2 \times 10^{-9}$; mean Wald statistic 12.07 versus 11.30). Based on cross-validation, both PC-Select and FaST-LMM Select chose to use all markers. This may indicate that the disease is not caused by a small number of loci with large effects or that our sample size is too small to capture this effect. Although, PC-Select and FaST-LMM Select chose to use all SNPs and thus neither method inflated statistics, we emphasize that without a priori knowledge about the genetic architecture, PC-Select automatically tunes the number of SNPs to include in the GRM to optimize power and simultaneously protects against population stratification at no cost to power.

6.3 Discussion

Janss et al. caution against using PCs as fixed effects in combination with a random effect derived from the GRM when estimating heritability [63]. This may result in an ill-posed model because the PCs enter both as fixed effects and implicitly through the random effect. We avoid this issue when estimating variance components by using the PCs as fixed effects in a restricted maximum likelihood (REML) approach, which projects the genotype matrix into a subspace orthogonal to the PCs, effectively removing them from the random effect. We also note that population structure and PCs have previously been used successfully as fixed effects (or separate random effects) in mixed model settings to address confounding from population structure and from unusually differentiated markers [158, 164, 101, 129, 102].

Using PCs in a linear model does not correct for family relatedness and cryptic relatedness [101]. As suggested by [153], due to the large length of segments shared identical-by-descent, using a subset of SNPs may correct for cryptic relatedness. [84] show that using a subset of SNPs in the GRM does not inflate statistics on the WTCCC data, where inflation is likely primarily due to cryptic relatedness. We expect that PC-Select will not be inflated by cryptic relatedness for the same reasons. In most human data sets with unrelated individuals, family relatedness is not an issue; however, for data sets with strong family relatedness, we suspect there may be cases where both PC-Select and FaST-LMM Select inflate statistics.

PC-Select has the same asymptotic run-time as FaST-LMM Select, quadratic in the number of individuals and linear in the number of markers. In practice, the run-time for the additional step of computing the PCs for the genotype matrix is minimal because both methods require several spectral decompositions of matrices of nearly the same size for the cross-validation step. It should be noted that while the asymptotic run-time of PC-Select and FaST-LMM Select is the same as previously published exact LMM methods [82, 168], the actual run-time of both methods is ostensibly longer by a factor of 10 due to the cross-validation step. The cross-validation step is parallelizable, so in practice this is not a significant limitation.

Including PCs as fixed effects allows PC-Select to infer ancestry from all SNPs simultaneously, while at the same time maintaining the benefits of using a statistically-chosen subset of the SNPs to estimate the GRM [84, 83]. As we have shown, using a combination of PCs and a subset of SNPs in the GRM gives the best of both worlds.

6.4 Methods

6.4.1 MS dataset

We analyzed data from 10,204 MS cases and 5,429 controls (from NBS and 1958BC) genotyped on Illumina arrays made available to researchers via WTCCC2 (`http://wtccc.org.uk/ccc2/`). We follow the quality control standards in [153]. Although [114] analyzed UK and non-UK samples separately followed by meta-analysis in most of their analyses, the data made available to researchers includes both UK and non-UK cases but only UK controls. We retained all samples in order to maximize sample size. We considered markers that were present in each of MS, NBS and 1958 BC datasets and removed markers with $> 0.5\%$ missing data, $P < 0.01$ for allele frequency difference between NBS and 1958BC, $P < 0.05$ for deviation from Hardy-Weinberg equilibrium, $P < 0.05$ for differential missingness between cases and controls, or $MAF < 0.1\%$ in any dataset, leaving 360,557 markers. The 75 known associated markers were defined by including, for each MS-associated marker listed in the NHGRI GWAS catalogue (`http://genome.gov/gwastudies/`), a single best tag at $r^2 > 0.4$ from the set of 360,557 markers if available.

6.4.2 Statistical methods

PC-Select follows a similar framework as FaST-LMM Select [82, 84, 83]. For completeness, we list the steps and equations we used.

First, we describe a method for computing association statistics, then in subsequent sections we describe the steps of PC-Select.

Association statistics:

The phenotype y, covariates X, and genotypes W are mean centered. Additionally, each genotype is divided by $\sqrt{2\hat{p}(1-\hat{p})}$ where $\hat{p}$ is the estimated minor allele frequency. Then the phenotype is modeled as

$$y = X\alpha + u + \epsilon,$$

where $u \sim N(0, \sigma_g^2 K), \epsilon \sim N(0, \sigma_e^2 I)$, α is a vector of weights for the covariates, and K is the GRM. This model naturally leads to an association statistic based on the Wald statistic.

To calculate the association statistic for SNP w, we add w as a fixed effect covariate to the previous model and test whether its coefficient is significantly different than 0. Specifically, consider the model

$$y = w\beta + X\alpha + u + \epsilon,$$

where β is the coefficient for the test SNP. We estimate σ_g^2 and σ_e^2 by REML. The fixed effect coefficients (β, α) are estimated by maximum likelihood.

It is straightforward to construct the Wald statistic to test whether $\beta \neq 0$. Let $V = \sigma_g^2 K + \sigma_e^2 I$ and $Q = [w; X]$. Then $\hat{\beta}$ is equal to the first entry of $(Q^T V^{-1} Q)^{-1} Q^T V^{-1} y$ and $\text{var}(\hat{\beta})$ is equal to the first entry of $(Q^T V^{-1} Q)^{-1}$. The test statistic is

$$\frac{\hat{\beta}^2}{\text{var}(\hat{\beta})},$$

which is asymptotically χ^2 distributed with 1 degree of freedom.

Now we describe the PC-Select method:

Step 1: Extracting PCs:

We extract the top 5 PCs from a GRM formed using all of the genotype data, WW^T, to use as fixed effect covariates. We use X to denote the matrix of user specified covariates and the top 5 PCs.

Step 2: Ranking SNPs by linear regression:

Second, we rank the SNPs by a linear regression test statistic. Linear regression test statistics are calculated by fixing σ_g^2 to 0 and using the procedure described above to calculate Wald statistics.

Step 3: Determining the GRM:

As in FaST-LMM Select, PC-Select uses a subset of the SNPs that are likely to be causal. In this step, we determine k, the number of top SNPs (as ranked in Step 2) to include in the GRM. We use 10-fold cross-validation on predictive log-likelihood to choose the number of top SNPs.

We choose k from a list of user defined possibilities (e.g., $k \in \{$100, 1000, 3000, 10,000, 30,000, $\ldots\}$). First, we randomly divide individuals into 10 equal groups or folds. For each fold i, we form a test set from the individuals in fold i and use the rest of the individuals as a training set. For each choice of k, we consider a subset of the genotype matrix consisting only of the top k SNPs (the ranking of the SNPs is recomputed per fold using the training data). For notational simplicity, we will also refer to the reduced genotype matrix by W, and it will be clear from context if this refers to the full genotype matrix or a subset. Let W_i denote the genotypes from fold i and W_{-i} represent the genotypes from the rest of the folds (similarly for y and X). We wish to evaluate the predictive log-likelihood of y_i given the training information (y_{-i}, X_{-i}, X_i) to assess the predictive power of using only the top k SNPs in the GRM.

Specifically, to evaluate the predictive log-likelihood, we start by forming a GRM from the training set $W_{-i}W_{-i}^T$. Then we estimate σ_g^2 and σ_e^2 from the training set by REML. We estimate α by ML with these variance parameters fixed. Then under the model

$$y = X\alpha + u + \epsilon$$

where $u \sim N(0, \sigma_g^2 WW^T)$ and $\epsilon \sim N(0, \sigma_e^2 I)$, the predictive distribution of the phenotypes given the training parameters, $y_i | y_{-i}, W, \alpha, \sigma_g^2, \sigma_e^2$, is normally distributed with mean

$$\sigma_g^2 W_i W_{-i}^T \left(W_{-i} W_{-i}^T \sigma_g^2 + \sigma_e^2 I\right)^{-1} (y_{-i} - X_{-i}\alpha) + X_i \alpha$$

and covariance

$$W_i W_i^T \sigma_g^2 + \sigma_e^2 I - \sigma_g^2 W_i W_{-i}^T \left(W_{-i} W_{-i}^T \sigma_g^2 + \sigma_e^2 I\right)^{-1} W_{-i} W_i^T \sigma_g^2.$$

This can be evaluated efficiently using the spectral decompositions computed in the REML step [82, 84]. We average the predictive log-likelihood over each of the 10 folds and choose the k that gives the highest average log-likelihood.

Step 4: Calculating association statistics:

Finally, with the number of top SNPs to use in the GRM fixed, we calculate association statistics for each SNP. Let W be the genotype matrix using the top k SNPs chosen in the previous step. To avoid proximal contamination [84], we use a leave-one-chromosome-out procedure [153]. For each test SNP w (which is not necessarily in W), we exclude the chromosome including that SNP from the GRM and calculate the Wald statistic for w with this GRM. We do this efficiently be precomputing and storing the GRM excluding each chromosome in turn.

Chapter 7

Phenotype prediction using regularized regression on genetic data in the DREAM5 Systems Genetics B Challenge

Abstract

[1] A major goal of large-scale genomics projects is to enable the use of data from high-throughput experimental methods to predict complex phenotypes such as disease susceptibility. The DREAM5 Systems Genetics B Challenge solicited algorithms to predict soybean plant resistance to the pathogen *Phytophthora sojae* from training sets including phenotype, genotype, and gene expression data. The challenge test set was divided into three subcategories, one requiring prediction based on only genotype data, another on only gene expression data, and the third on both genotype and gene expression data. Here we present our approach, primarily using regularized regression, which received the best-performer award for subchallenge B2 (gene expression only). We found that despite the availability of 941 genotype markers and 28,395 gene expression features, optimal models determined by cross-validation experiments typically used fewer than ten predictors, underscoring the importance of strong regularization in noisy datasets with far more features than samples. We also present substantial analysis of the training and test setup of the challenge, identifying high variance in performance on the gold standard test sets.

[1] This chapter previously appeared in *PLoS One* (2011) as "Phenotype prediction using regularized regression on genetic data in the DREAM5 Systems Genetics B Challenge" by Po-Ru Loh, George Tucker, Bonnie Berger [85].

7.1 Introduction

Predicting complex phenotypes from genotype or gene expression data is a key step toward personalized medicine: the use of genomic data to improve the health of individuals, for instance by predicting susceptibility to disease or response to treatment [148, 106, 147, 104]. A pivotal early success in this field was the discovery of gene expression profiles for the classification and prognosis of breast cancer [52, 1, 142]. Improved technology and declining costs have since enabled ever-larger genetic screens and gene expression studies, allowing researchers to apply the power of genetic analysis of genome-wide gene expression [13, 116]. The difficulty has thus shifted to the algorithmic side: untangling complex associations and identifying small numbers of influential predictors of phenotypic effects amid a sea of largely unrelated measurements [29, 115]. One avenue of recent research has been the integration of distinct types of genomic data to enhance inference, including both linkage studies combining knowledge from different organisms [21, 37] and integrative analysis of distinct data types for the same organism [76, 20].

It is difficult to objectively measure progress on algorithmic challenges without standard benchmarks; within this context, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) initiative [103] aims to provide a fair comparison of methods and a clear sense of the reliability of the models. The fifth annual DREAM challenge held in 2010 included a Systems Genetics component with the goal of predicting disease susceptibility from (1) only genotype data, (2) only gene expression data, and (3) genotype and gene expression data. Through the challenge, the organizers hoped to identify the best predictive modeling approaches and to evaluate the benefits of learning from combined genotype and gene expression data [20].

As a top performer on the second part of the challenge, we were invited to present our results at the DREAM5 conference and contribute to the DREAM5 collection in PLoS ONE; this paper describes our approach. We provide a comparison of several regularized regression models and find comparable performance of elastic net, lasso, and best subset selection. We also carefully analyze the level of noise in the data and consequent variability in performance and offer practical suggestions for similar data analysis and data pre-processing.

7.2 Materials and Methods

7.2.1 Dataset and challenge setup

The data for this challenge were collected from a systems genetics experiment conducted at the Virginia Bioinformatics Institute [167]. Two inbred lines of soybean plants that differed substantially in susceptibility to a pathogen, *Phytophthora sojae*, were crossed and their offspring were inbred for more than 12 generations to produce a population of recombinant inbred lines (RILs). Individuals within each RIL exhibited almost no genetic variation, whereas distinct RILs displayed much genetic variation owing to their differing mixtures of parental genes. Each RIL was screened for 941 genetic variants and gene-expression profiled for 28,395 genes; gene expression was measured in uninfected plants because the goal of the challenge was to predict disease susceptibility using only information gathered under normal (healthy) conditions.

After infection with *P. sojae*, the plants were assayed for two continuous phenotypes, each a measurement of the amount of pathogen RNA in the infected tissue sample. The first phenotype measured the fraction of pathogen probe sets that yielded a detectable hybridization signal as determined by the MAS5 presence/absence call in the Affymetrix software used to analyze the data. The second phenotype measured the ratio between the sum of all background-subtracted soybean probe intensities and the sum of all background-subtracted pathogen probe intensities. We abbreviate the two phenotypes as P1 and P2.

The training data, from 200 RILs, thus consisted of a 200×941 boolean matrix of genotype values (denoting presence or absence of genotype variants), a $200 \times 28,395$ real matrix of gene expression values, and a 200×2 real matrix of phenotype values. Three distinct test sets of 30 RILs each were used for evaluating submissions; 30×941 genotype and/or $30 \times 28,395$ gene expression matrices were provided accord-

ing to the respective subchallenge conditions, and predictions of the corresponding (withheld) 30 × 2 phenotype matrices were solicited. At the end of the submission period, predictions were scored according to their Spearman (rank) correlations to the withheld "gold standard" phenotype data. All training and test data are available at the DREAM5 challenge website
(`http://wiki.c2b2.columbia.edu/dream/index.php/D5c3`).

7.2.2 Preliminary ranking of predictors by correlation

We began our analysis for this challenge by computing correlation coefficients of the genotype and gene expression training features against the two phenotype variables. The magnitudes of these correlations guided our choice of modeling technique; we also later used correlation-sorted rank lists to limit the scope of computationally intense calculations to those features most likely to be relevant.

On first glance the highest correlations, above 0.3 for the expression data (Table 7.1), appear promising. The significance of these correlations needs to be considered with the numbers of features in mind, however: 941 genotype and 28,395 gene expression markers. As a rough sanity check, we generated random matrices with sizes equal to those of the training predictor matrices and computed the correlation coefficients of these random features with the training phenotype data. This experiment revealed that in fact the training features as a whole are only very weakly correlated with the phenotypes: almost all correlations from the real training data are within 0.03 of the highest random correlations, and only one real correlation is substantially larger (the 0.34 observed in expression vs. phenotype 2). From the point of view of Bonferroni-corrected p-values, this largest correlation is significant with p-value 0.017; all other p-values exceed 0.1 upon applying the Benjamini-Hochberg multiple hypothesis correction [8].

These observations suggest that most features have little or no predictive power, and hence proper regularization is crucial for modeling this dataset. Additionally, the small difference between training correlations and the random background distribution indicate that the prediction task at hand is difficult; the amount of signal in the data is likely quite small.

In light of the above considerations, we sought to keep our modeling simple and chose regularized regression as our general approach. Before fitting the data, however, we needed to ensure that the relation between predictor and response variables was as linear as possible, and so we considered data transformations and basis expansions.

7.2.3 Rank transformation to reduce phenotype outliers

Upon plotting the phenotype training data, we discovered that the variance in the distribution of phenotype 1 is dominated by outliers. Among the 200 measurements of phenotype 1, the largest outlier is 5.83 sample standard deviations from the mean. Moreover, the seven most deviant samples account for more than half of the total variance. For phenotype 2, the largest outlier is a substantial 3.77 standard deviations above the mean but overall the distribution does not have unusually long tails compared to a normal distribution. A plot of the fractions of variance explained by increasing subsets of largest outliers in phenotype 1, phenotype 2, and random data illustrates this behavior (Figure 7-1).

Motivated by the Spearman correlation-based scoring scheme used in this challenge, which judges predictions based on ordering rather than absolute accuracy, we applied a rank transformation to phenotype 1 to remove the impact of outliers on regression models. More precisely, we replaced the numerical values of phenotype 1 measurements with their ranks among the 200 sorted samples. Because the approaches we applied minimized squared error (along with regularization terms), asking our models to predict ranks rather than actual values removed the heavy weight that outlier values would otherwise have received. Absolute predictions could of course be recovered by interpolation if desired.

7.2.4 Basis expansion to boolean combinations of genotype variables

With only binary genotype data available for prediction in subchallenge B1, we hypothesized that the true phenotypic response for a genotyped sample would be far from linear. The simplest possible example of a nonlinear effect is interaction between genotype markers: for instance, if two genes act as substitutes for one another, their function is only suppressed if both are turned off. Similarly, if two genes are critical to different parts of a pathway, turning off either one would impair its function.

With these examples in mind, we considered applying logic regression [107] to expand the set of features available to our linear models to include boolean combinations of each pair $\{A, B\}$ of genotype features:

$$A \wedge B, A \wedge \neg B, \neg A \wedge B, \neg A \wedge \neg B.$$

Note that the complements of these relations are implicitly included by a linear model as well, so together they cover all nontrivial binary boolean relations.

To gauge the efficacy of these combined features, we compared the largest fractions of variance explained by single boolean combination features (using single-variable least-squares regression) to the best fits obtained by two-variable regression on pairs of the original genotype features. Looking at the 20 best-performing regressions from each group (Figure 7-2), we see that the top boolean combinations outperform the best two-variable regressor pairs, suggesting that basis expansion in this manner does indeed improve our ability to fit the data.

An important caveat to keep in mind when interpreting these measurements is that the number of feature combinations considered is very large (nearly 2 million), thus allowing random chance to inflate best performances as in the case of correlations examined above. Nonetheless, we expect that the relative trends are still informative.

Upon closer inspection of the best boolean combination markers, we discovered that some were near-trivial due to linkage disequilibrium (Figure 7-3): for instance, we observed cases of nearby markers A and B having identical values for 198 out of 200 samples, so that the boolean combination $A \wedge \neg B$ was nonzero for only two samples. Such combinations are very noisy (and likely uninformative) predictors; we therefore limited the boolean features under consideration to those containing at least 20 nonzeros.

7.2.5 Regularized regression modeling

Having taken steps to linearize the predictor-response relationship, we applied regularized regression to model the data. Classical linear regression on a predictor matrix $X \in \mathbb{R}^{N \times p}$ and response vector $y \in \mathbb{R}^N$ assumes a model $y = X\beta + w$ (where w represents noise) and finds the coefficient vector $\hat{\beta} \in \mathbb{R}^p$ minimizing the sum of squared residuals $||y - X\hat{\beta}||_2^2$. In the highly underconstrained case ($p \gg N$), however, additional constraints must be imposed for there to be any hope of approximating β; often one assumes that β is sparse, in which case ℓ_1-minimization techniques may be applied [137]. In the context of our experimental setup this assumption means that most genetic markers and expression values are unrelated to phenotype, which seems reasonable.

Our main approach of choice was elastic net regression [169], which imposes constraints on model complexity by adding the following penalization term to the squared residuals being minimized:

$$\lambda \left(\alpha ||\beta||_1 + (1 - \alpha) \frac{||\beta||_2^2}{2} \right),$$

where $0 \leq \alpha \leq 1$ determines the weighting of the two terms and $\lambda > 0$ is the strength of the regularization. Note that $\alpha = 0$ produces the ridge regression penalty while $\alpha = 1$ gives the lasso; thus, in some sense elastic nets interpolate between ℓ_2- and ℓ_1-regularization. Elastic net regression can be computed efficiently; we used the `glmnet` package available for Matlab [45].

For the purpose of comparison, we also tried fitting the data with a simple best subset selection approach, which seeks to minimize squared error using only a limited number of regressors. (In the language of our above discussion, this constraint can equivalently be viewed as imposing an ℓ_0 penalty $\lambda||\beta||_0$.) Because best subset selection is a nonconvex combinatorial problem with exponential complexity, however, finding best subsets exactly was computationally intractable [46]; instead, we performed simulated annealing on a subset of likely candidate features (chosen by correlation-ranking within our cross-validation loop) to obtain a reasonable approximation.

Implementation details are as follows. For elastic net regression, we ran `glmnet` with $\alpha = 0, 0.1, 0.2, \ldots, 1$ and uniformly log-spaced regularization path and default values of all other parameters. The best pair of (α, λ) for the elastic net was then selected to achieve optimal cross-validation performance. For the lasso, we ran `glmnet` with $\alpha = 1$ and default values of all other parameters. In this case, `glmnet` automatically calculated a regularization path and we selected the least complex model achieving within one standard deviation of the best cross-validation performance. We used this value of λ for our final regression fit.

For best subset selection, we first filtered to the top 30 features with strongest correlations to phenotype (recomputed for each cross-validation training set). We then used simulated annealing to compute subsets of size 1–20 features obtaining approximately optimal linear fits to each training fold. The annealing procedure consisted of 5 runs of initialization with a random feature subset of the required size followed by 5000 iterations of attempted swaps, using a linear cooling schedule. Explictly, the acceptance probability of a swap was

$$\exp(5 \cdot (\text{fractional improvement in fit})/(\text{fraction of iterations left})),$$

capped at 1.

7.3 Results

7.3.1 Modest performance of all regression techniques on training dataset

We evaluated our regression methods using 7-fold cross-validation on the 200-sample training set, measuring goodness of fit with Spearman correlation to match the DREAM evaluation criterion. We chose to use 7 folds so that our cross-validation test sets during development would have approximately the same size as the 30-sample gold standard validation set, allowing us to also estimate the performance variance to be expected on the validation set. We applied each regression technique—elastic net, lasso, and approximate best subset selection with simulated annealing—to fit phenotype 1 (rank-transformed) and phenotype 2 individually, using sets of regressors corresponding to the three subchallenges of DREAM5 Systems Genetics B: genotype only (B1), gene expression only (B2), and both genotype and expression (B3). Within subchallenge B1, we ran two sets of model fits, one using only raw genotype markers as regressors and the other using the boolean basis expansion described in Methods.

Because of the relatively small number of samples and large number of predictors, the random assignment of samples to cross-validation folds caused substantial fluctuation in performance, even when averaging across folds. We overcame this difficulty by running multiple cross-validation tests for each model fit using different fold assignments in each run (20 replicates for elastic net and lasso and 5 replicates for best subset selection), thus obtaining both mean performances and estimates of uncertainty in each mean. We chose regularization parameters for each method in each situation to optimize mean performance; Figure 7-4 shows the results using these parameters.

Overall, the three regularized regression techniques perform quite comparably. Note that elastic net regression necessarily always performs at least as well as lasso (because lasso corresponds to the elastic net with parameter choice $\alpha = 1$); however, the performance difference is very small in all cases. Best subset selection appears to perform slightly better than the others in predicting phenotype 1 and somewhat worse in predicting phenotype 2.

Comparing the different regressor sets, subchallenge B1 with genotype data only is clearly the most difficult. The availability of gene expression data in subchallenges B2 and B3 dramatically boosts average Spearman correlations to the 0.25-0.3 range for phenotype 1 (though performance for phenotype 2 is largely unchanged in the 0.15-0.2 range typical for all other cases). Unfortunately, our regression models did not attain a performance increase from B2 to B3 with the inclusion of genotype data along with expression data, nor did boolean basis expansion appear to help with performance on B1.

7.3.2 Effectiveness of rank transformation on phenotype 1

Surprisingly, the rank transformation we applied to phenotype 1 turned out to have the greatest impact of the pre-regression data transformations we attempted. For the purpose of comparison, we performed the same model-fitting as above using raw (untransformed) values of phenotype 1. In all cases the rank transformation increases average Spearman correlations considerably (Table 7.2). For subchallenges B2 and B3, rank-transforming phenotype 1 more than doubles the correlation that would otherwise be achieved, though a look at scatter plots of predicted versus actual values (Figure 7-5) shows that our predictive power is still marginal: predictions are compressed toward the mean, as tends to occur when trying to apply regression to data that is difficult to model. The effectiveness of the rank transformation was unique to phenotype 1; in contrast, rank-transforming phenotype 2 had no significant effect.

7.3.3 Strong regularization in best-fit models

Taking a closer look at the optimal regularization parameters for elastic net, lasso, and approximate best subset selection, we discovered strikingly low model complexity prescribed by cross-validation in each case. As an example, the blue curves of Figure 7-6 plot average performance of lasso and best subset selection on subchallenge

B2 as a function of increasing model complexity. (Note that unlike typical cross-validation curves with error to be minimized on the vertical axis, our performance metric is Spearman correlation so we seek maxima.) The regularization parameter is particularly transparent for best subset selection (shown in the bottom two plots): in this case, regularization is explicitly manifested as the number of features to be used in the subset chosen for regression.

With lasso, we likewise see that performance drops off quickly as model complexity increases; here, the complexity parameter λ is less directly interpretable, but since the ℓ_1-minimization approach of lasso also results in sparse models, the result in this case as well is that lasso also recommends using only a handful of features. Even with elastic net regression, which tends to fit denser models due to the presence of an ℓ_2 "ridge" penalty, we find that optimal regularization parameter choices de-emphasize the ridge term, creating lasso-like model fits with α (the "lasso proportion") typically in the range 0.8 to 1.

To better understand the strong regularization, we provide heat maps displaying the feature weight distributions chosen by the elastic net to predict phenotype 1 (rank-transformed) and phenotype 2 for a set of cross-validation runs on subchallenge B2 (Figure 7-7). As expected, the few features chosen from the 28,395 available are typically among those predicted to be most informative according to correlation with phenotype (Table 7.1). The features assigned greatest weight are quite stable from fold to fold, while the choice of lower-weight features is noisier.

7.3.4 High variance in performance on individual cross-validation folds and test set

As mentioned earlier, our cross-validation analysis also allows us to estimate the accuracy to which algorithm performance can be measured using a 30-sample test set. Unfortunately, we find that this test size is insufficient for accurate evaluation: whereas the greatest-weight features selected by our models are relatively stable from fold to fold (Figure 7-7), the Spearman correlations obtained on the held-out test folds vary markedly. The blue error bars in Figure 7-6 display one standard deviation in the Spearman correlation between predicted and actual phenotype values from fold to fold; with 7-fold cross validation, each fold contains about 29 samples. These standard deviations mostly fall in the 0.15-0.2 range, in some cases exceeding the mean performance of even the best parameter choice.

The red curves of Figure 7-6 illustrate the variance in performance when models fit on the training data were applied to the actual 30-sample gold standard test set (released after the end of the DREAM5 challenge). As expected, test set performance strays substantially from the mean.

7.3.5 Official DREAM5 challenge results

Notwithstanding the caveat just discussed regarding uncertainty in results on a small test size, we include the final results from the DREAM5 Systems Genetics B challenge for completeness (Figure 7-8). Our team, identified by "orangeballs" and Team 754 in the published results, achieved the best performance on the subchallenge B2 test set. The overall distribution of Spearman correlations achieved by the various teams is in line with what we would expect given our analysis of our training results, with subchallenges B2 and B3 being more tractable than B1.

7.4 Discussion

While the performance achieved by our methods—indeed, by every team's methods—is modest, our work does highlight a few important lessons in statistical learning and in the setup of algorithmic benchmarking challenges such as DREAM. Regarding the first, our analysis did not lead us to a radically new and complex model for the genotype-phenotype relationship in *P. sojae*; on the contrary, we found that given the limitations of small sample size and noise in the training data, the best models we discovered were among the simplest we tried. Regularized least squares regression with careful cross-validation and linearization (using the rank transform we applied to phenotype 1) proved to be as effective an approach as any other we are aware of, and the noise-to-signal in the data was such that the best linear fits needed only a few well-chosen regressors.

One might hope that the transparency of such simple models can shed light on the underlying biological mechanism at work; while this may be possible, we also should caution against trying to glean more from the models than the data allow. Simplicity may be due to the involvement of only relatively few genes or just to the fact that heavy regularization makes models less prone to overfitting. In light of the noisiness of the dataset, we suspect the latter may be true. As a case in point, while we were disappointed that modeling pairwise interactions through boolean basis expansion did not improve fitting using the genotype data, we still find it quite plausible that such effects are at work and may aid modeling in situations when more data is available. With this dataset, our techniques were likely unable to discern these effects because the limited data size could not support the increased complexity that modeling interactions would entail.

Overall, while this contest was perhaps too ambitious for the data available, we feel it succeeded in stimulating research and discussion in the field. The original motivation of developing methodology for combining genotype and gene expression data to improve phenotype prediction remains a worthy goal and interesting open question.

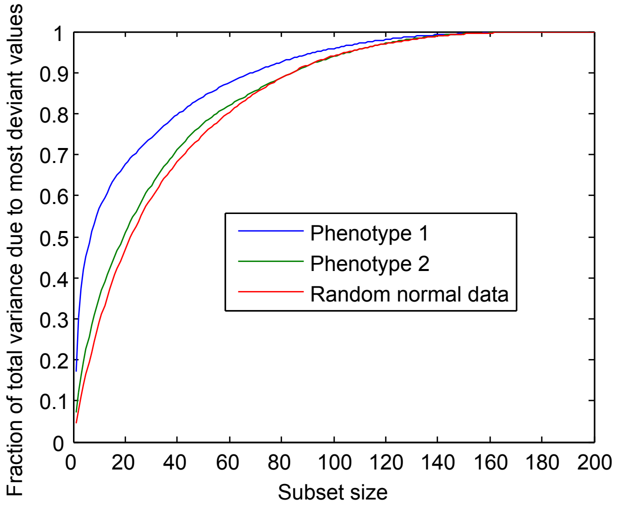


Figure 7-1: Large contribution of outliers to variance in phenotype 1. The largest seven outliers in phenotype 1 account for the bulk of the variance in the data; in contrast, the outlier distribution for phenotype 2 is similar to that of a random normal variable.

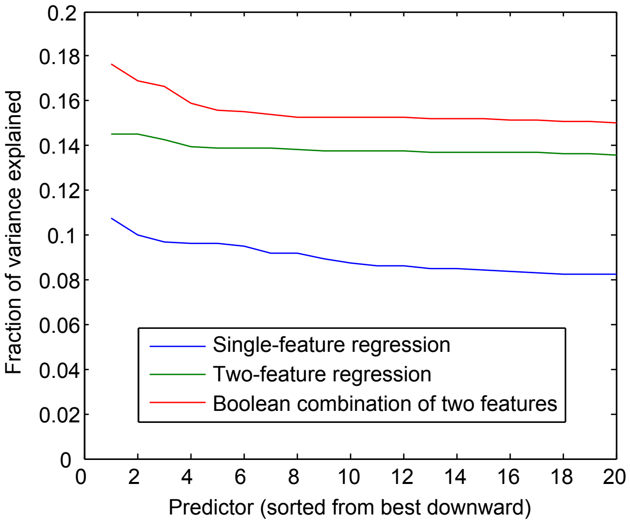


Figure 7-2: Single-variable, two-variable, and pairwise logic regression for phenotype 2. The plot compares the best least squares fits attainable under three model types: single-variable regression using each genotype feature independently (blue), two-variable regression using pairs of features at once (green), and single-variable regression using pairs of features combined through a binary boolean relation (red). The best single-variable fits using boolean combination features outperform the best two-variable regressions.

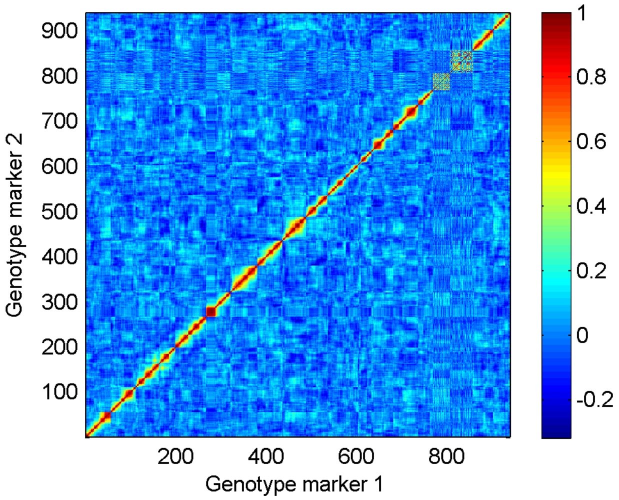


Figure 7-3: Correlation coefficients between genotype markers, displaying linkage disequilibrium. The heat map shows Pearson correlations between pairs of genotype markers; most pairs have only slightly positive or negative correlations attributable to chance, but groups of nearby markers exhibit distinctly positive correlations.

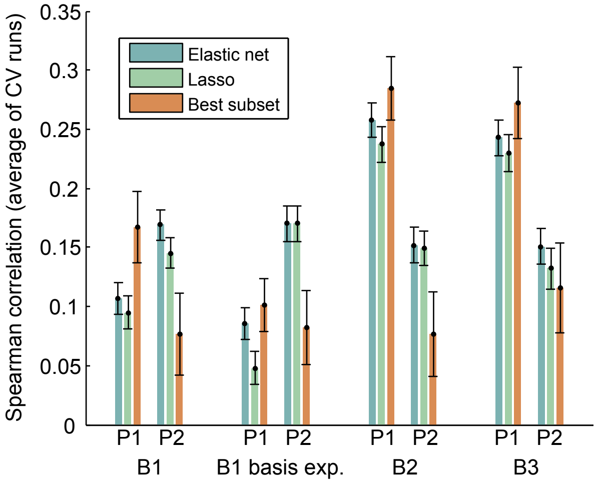


Figure 7-4: Goodness of fit of regularized regression models on training data using various regressor sets. We tested elastic net, lasso, and approximate best subset selection on phenotypes 1 and 2 using regressor sets derived from the DREAM5 subchallenges B1, B2, and B3. In each case the regularization parameter(s) were chosen to optimize average Spearman correlation. We ran multiple cross-validation tests with different random fold splits to reduce uncertainty in mean performance and enable comparison between methods; error bars show one standard deviation of confidence.

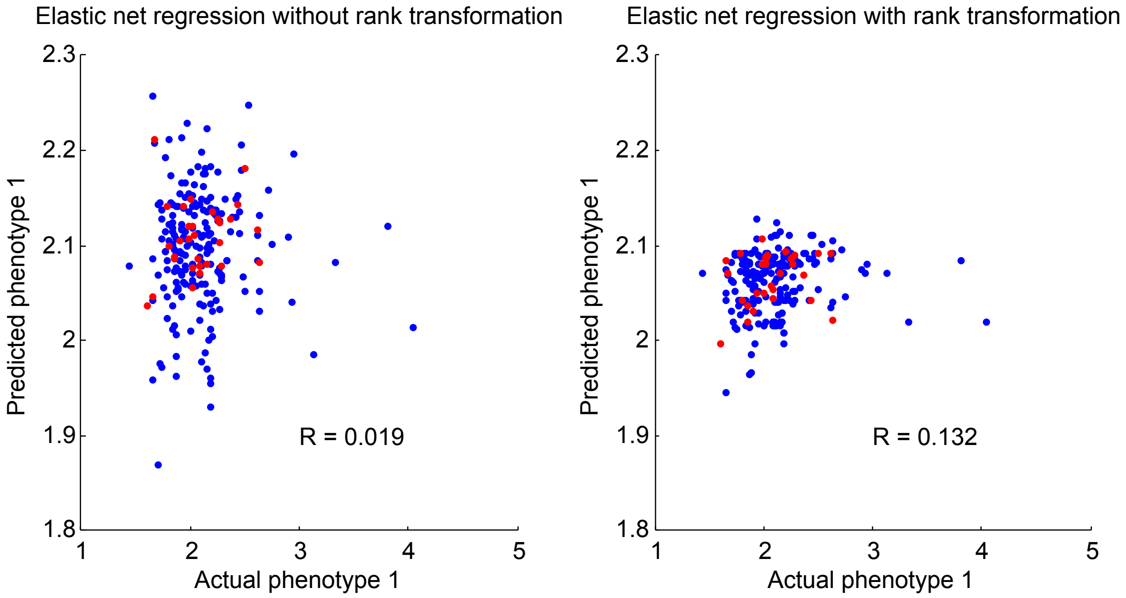


Figure 7-5: Example elastic net predictions versus actual values with and without rank transformation for subchallenge B2P1. Each scatter plot shows predictions from one cross-validation run on the training data (blue points) as well as predictions of the fitted model for the gold standard test set (red points). For the elastic net modeling on rank-transformed data (right plot), predictions of phenotype 1 values on an absolute scale were obtained by interpolation. The reported values of R are the Pearson correlation coefficients.

Top correlations	Genotype		Expression	
(absolute values)	Training	Random	Training	Random
Phenotype 1	0.2155	0.2404	0.3034	0.2835
	0.2122	0.2116	0.2976	0.2781
	0.2061	0.1862	0.2975	0.2749
	0.2054	0.1857	0.2963	0.2689
	0.2041	0.1851	0.2909	0.2611
Phenotype 2	0.2433	0.2127	0.3441	0.2777
	0.2261	0.2104	0.3084	0.2684
	0.2198	0.2053	0.2990	0.2679
	0.2181	0.1928	0.2824	0.2642
	0.2180	0.1926	0.2754	0.2619

Table 7.1: Highest absolute correlations of genotype and gene expression data to phenotype, versus random background. The top five correlations found in the training data are shown, as are the top five correlations against a random 0-1 matrix with the same dimensions as the genotype data and a random normal matrix replacing the gene expression data.

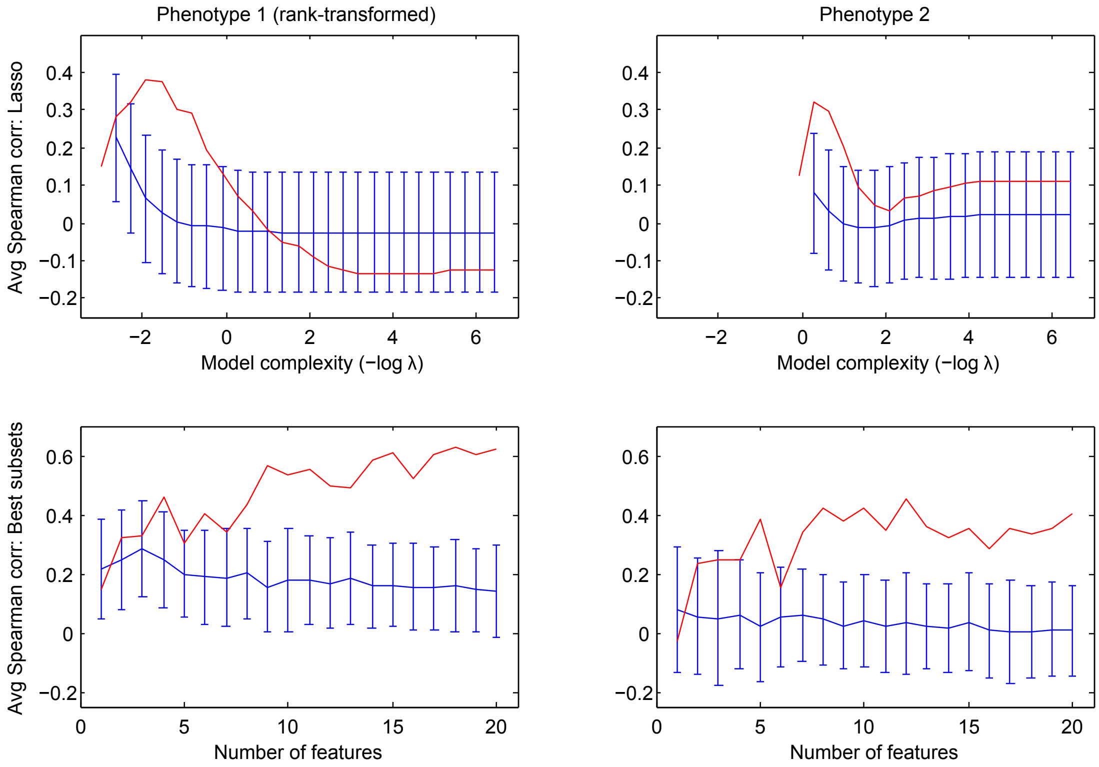


Figure 7-6: Variation in cross-validation and test set performance with model complexity for subchallenge B2. Each plot follows the performance of a regression model as complexity increases. For lasso (top plots), model complexity is determined by a regularization parameter λ; for best subset selection (bottom plots), complexity is defined as the number of features used. The blue curves show Spearman correlations averaged over cross-validation folds, each fold having approximately the same size as the gold standard test set. Performance varies dramatically from fold to fold; error bars show one standard deviation of the Spearman correlations achieved for different folds. The red curves follow performance of the models on the actual gold standard.

	Spearman corr. before and after transformation					
Subchallenge (regressors)	Elastic net		Lasso		Best subset	
B1 (genotype)	0.058	0.107	0.054	0.095	0.092	0.167
B1 (genotype with basis expansion)	0.042	0.085	0.011	0.048	0.025	0.102
B2 (expression)	0.099	0.257	0.094	0.237	0.111	0.285
B3 (genotype and expression)	0.090	0.243	0.077	0.230	0.092	0.272

Table 7.2: Improvement in goodness of fit with rank transformation on phenotype 1. Applying the rank transform to phenotype 1 increases average cross-validated Spearman correlations for all regression approaches and regressor sets we tested. The performance improvement is especially large for subchallenges B2 and B3, where gene expression data is available.

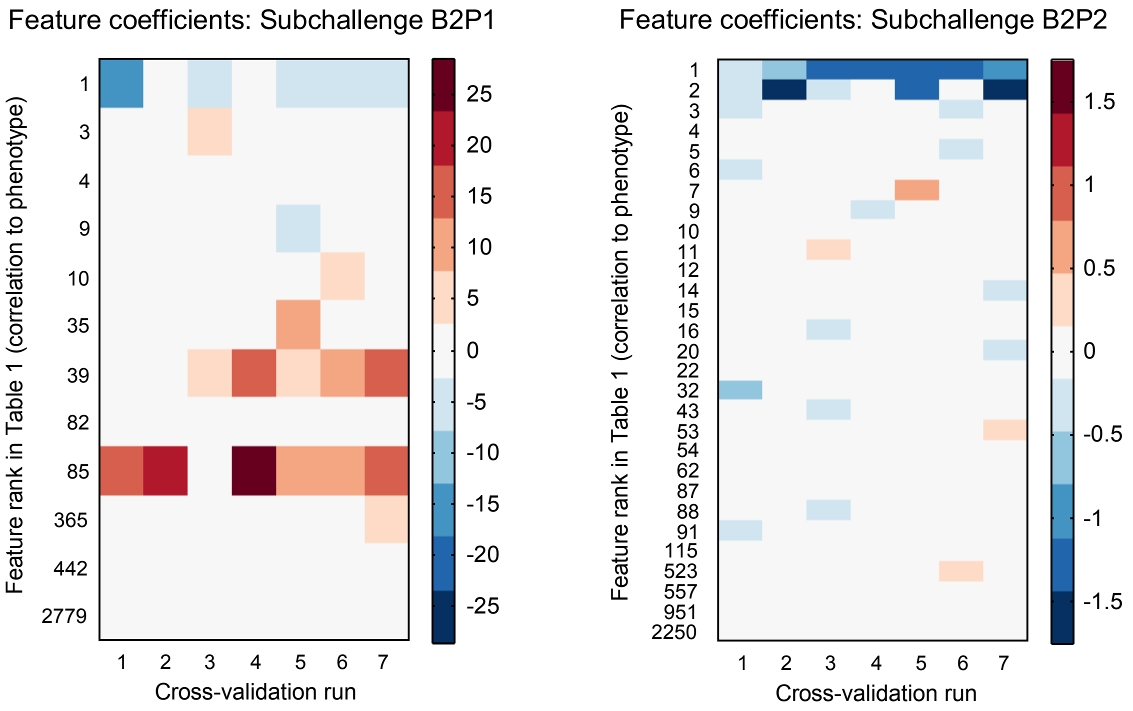


Figure 7-7: Stability of features and coefficients selected by elastic net regression for subchallenge B2. The heat maps show regression coefficients chosen by the best-fit elastic net models as each cross-validation fold is in turn held out of the training set. The features shown on the vertical axis are those having a nonzero coefficient in at least one of the seven runs; they are indexed by their rank in Table 1, correlation to the phenotype being predicted.

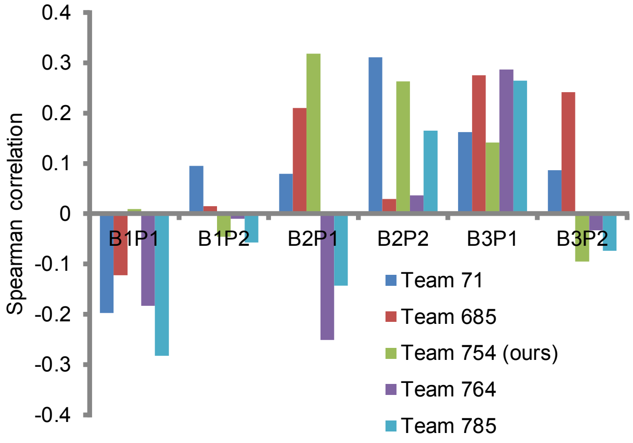


Figure 7-8: Final results of the five teams participating in all three DREAM5 Systems Genetics B subchallenges. All teams had difficulty even achieving consistently positive correlations; we suspect the main obstacles were the large amount of noise in the data and the small 30-sample gold standard evaluation sets. We achieved the best performance on the test set used for subchallenge B2 (prediction using gene expression data only).

Appendix A

Supporting Information for Incorporating quantitative mass spectrometry data in protein interaction analysis

Figure A-1: As the validation databases for protein interactions are not complete, we do not have true negative protein interactions, so we cannot form ROC curves. We, as is typical (see [77, 25]), plot percent of predicted interactions present in the respective validation set for a varying number of predicted interactions, which conveys conceptually similar information to a ROC curve. For this figure, all interactions supported by at least one external source are included in the validation set. Otherwise, the setup is the same as Figure 3-5.

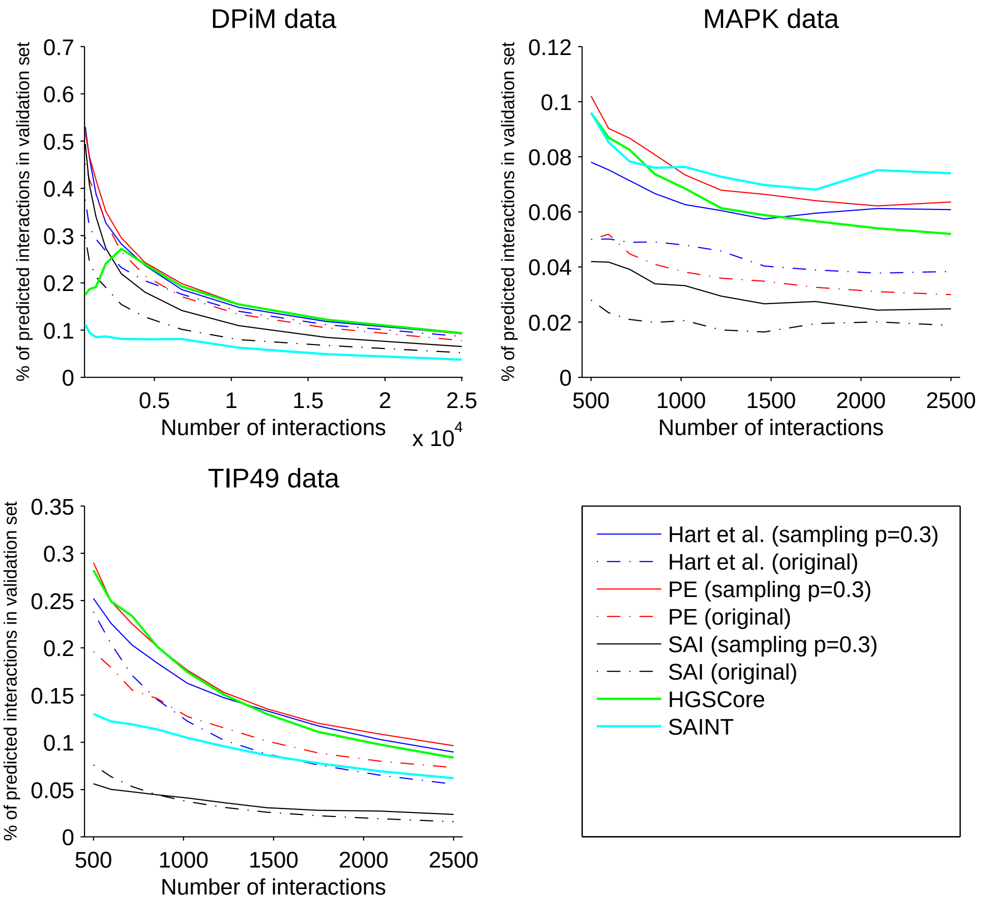


Figure A-2: Percent of predicted interactions present in the respective validation set for a varying number of predicted interactions. For this figure, all interactions supported by at least two external sources are included in the validation set. Otherwise, the setup is the same as Figure 3-5.

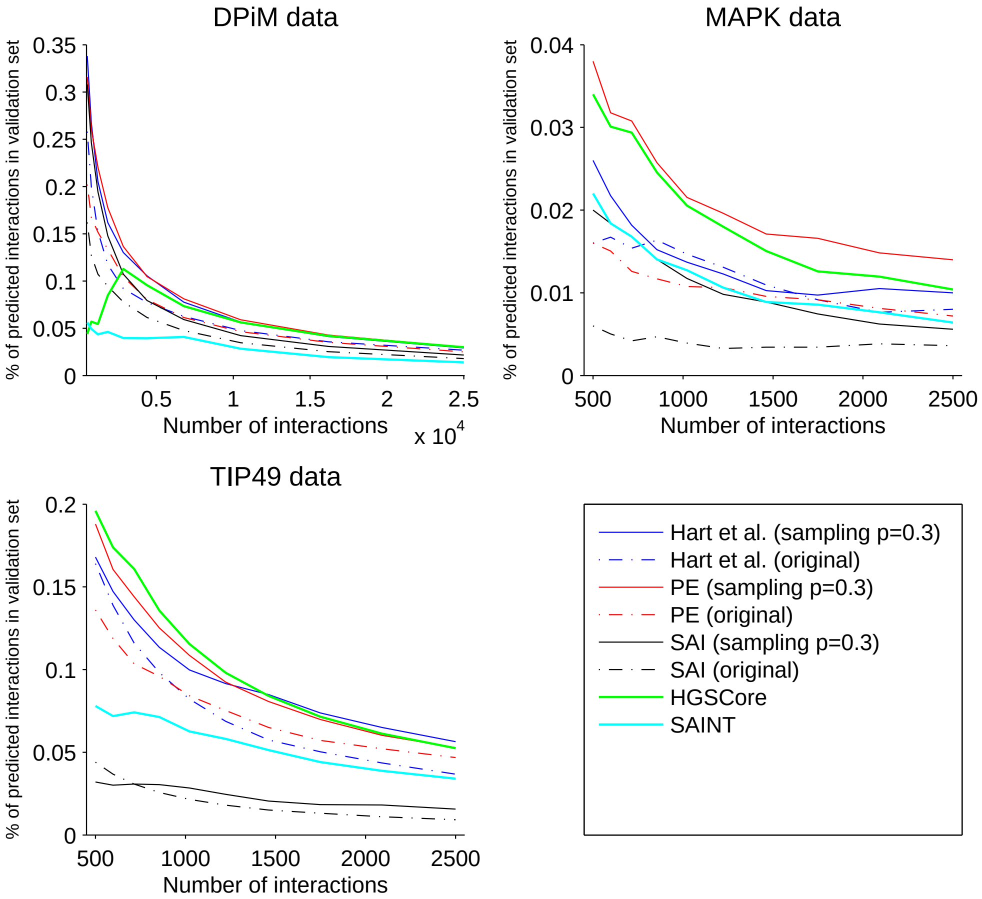


Figure A-3: Percent of predicted interactions present in the respective validation set for a varying number of predicted interactions. For this figure, all interactions supported by at least three external sources are included in the validation set. Otherwise, the setup is the same as Figure 3-5.

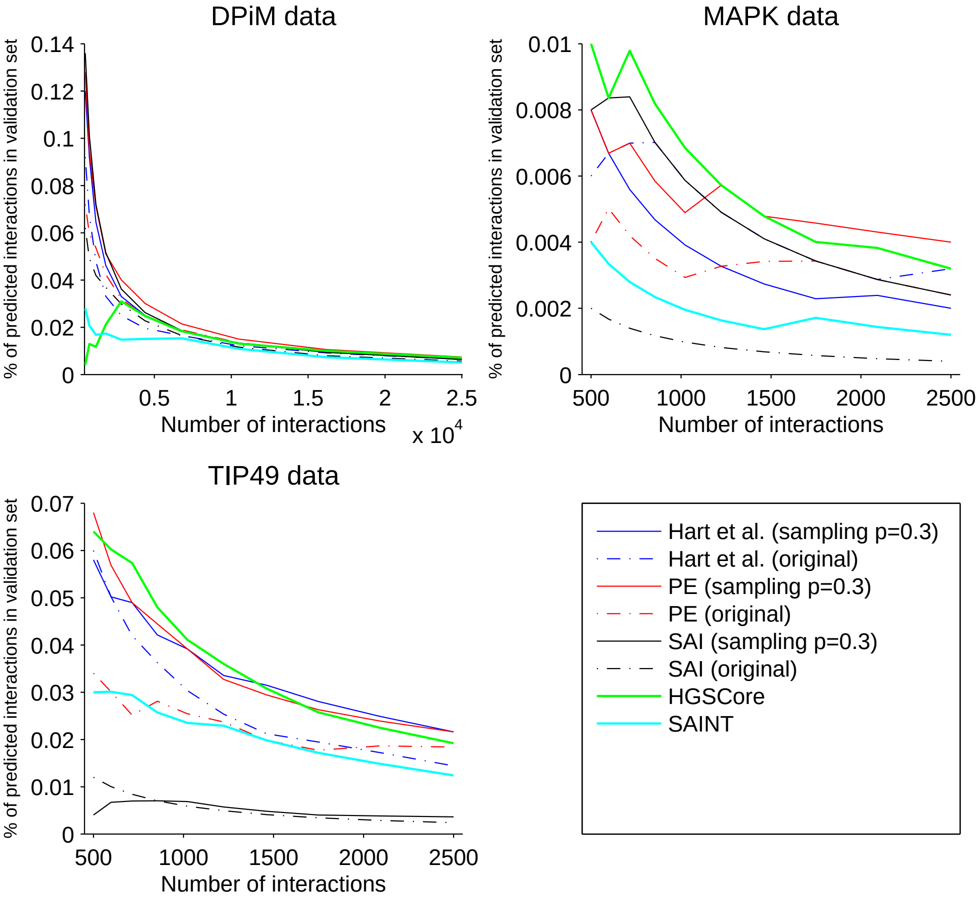


Figure A-4: Performance comparison of methods using $p = 0.2$ as the sampling parameter. The setup is otherwise the same as in Figure 3-5.

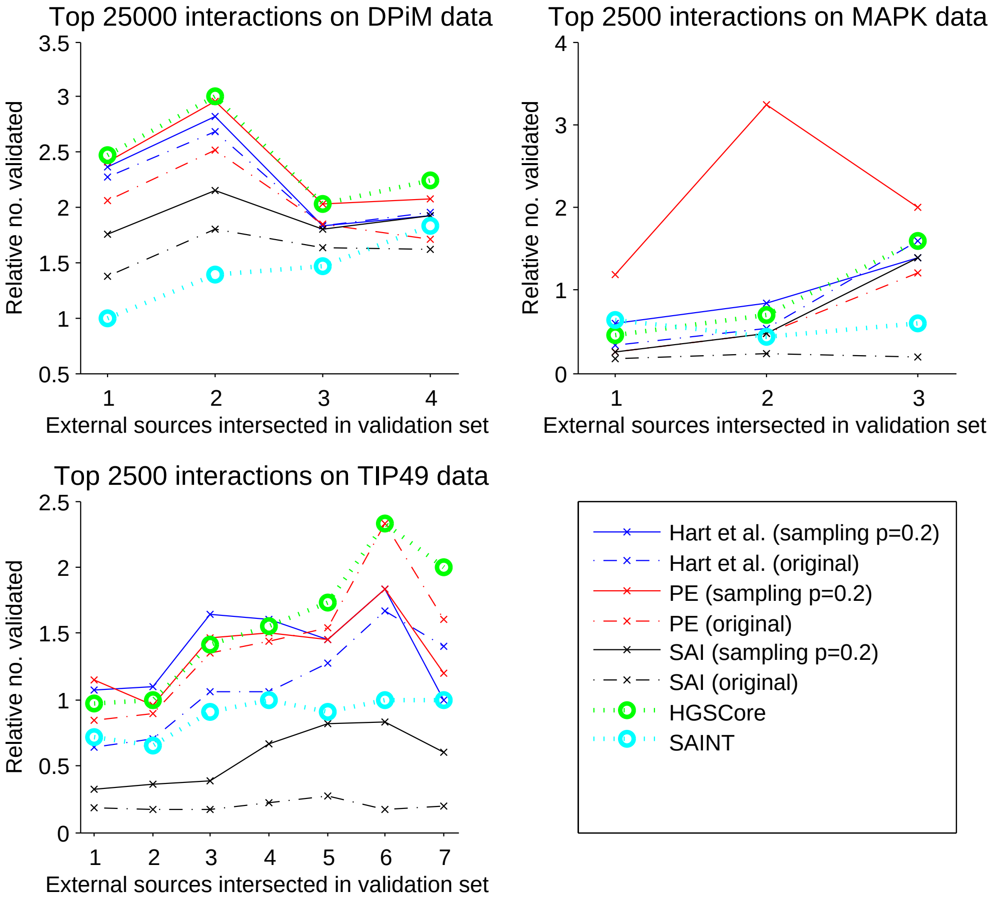


Figure A-5: Performance comparison of methods using $p = 0.5$ as the sampling parameter. The setup is otherwise the same as in Figure 3-5.

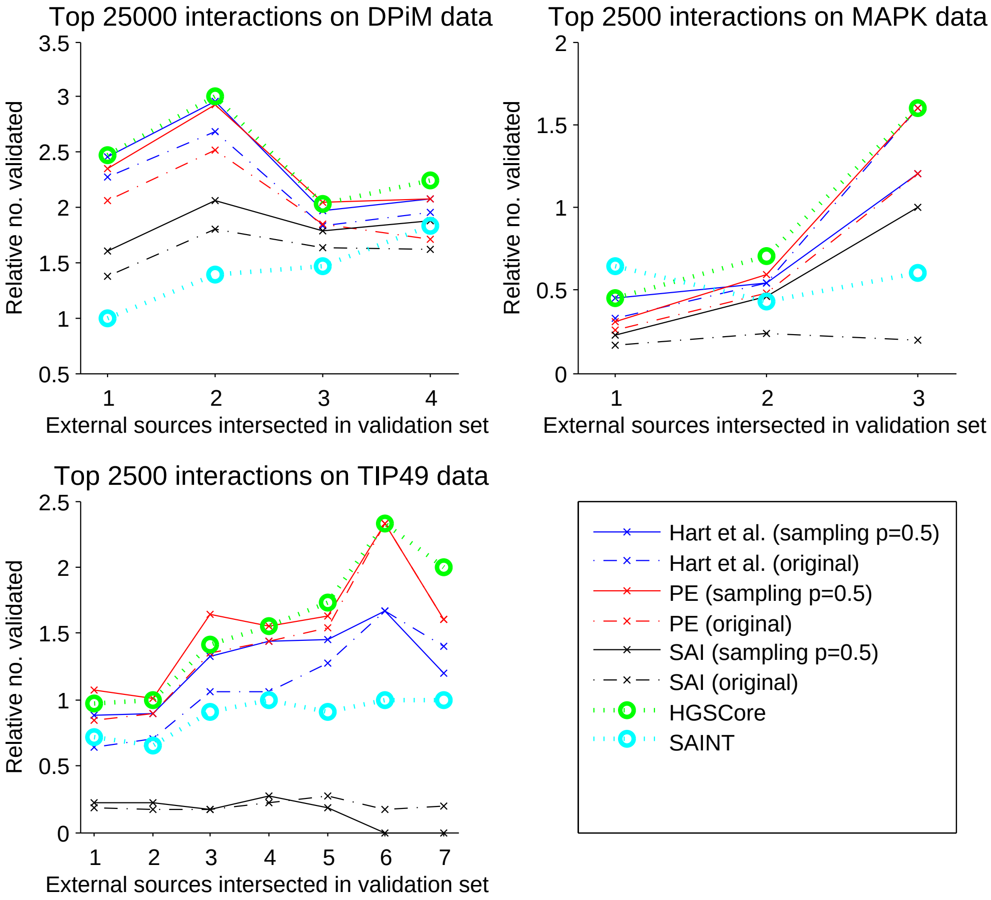


Figure A-6: Sensitivity of performance to sampling parameter p for higher-confidence predictions. Only the top 40% of predictions considered in Figure 3-6 are evaluated here. The setup is otherwise the same.

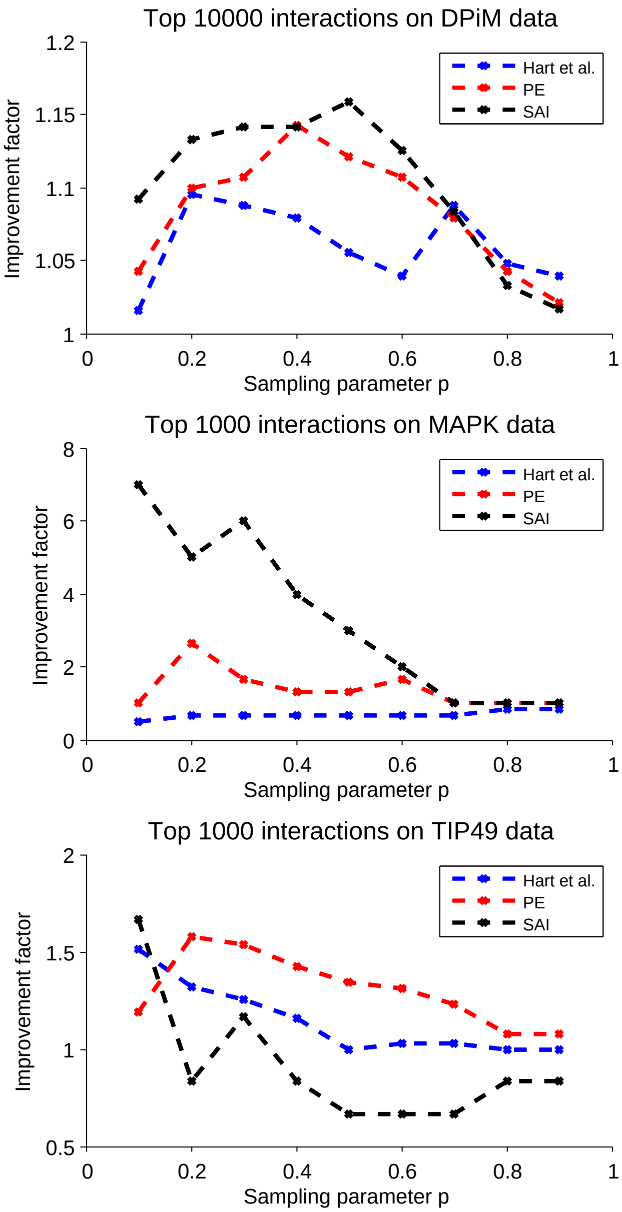
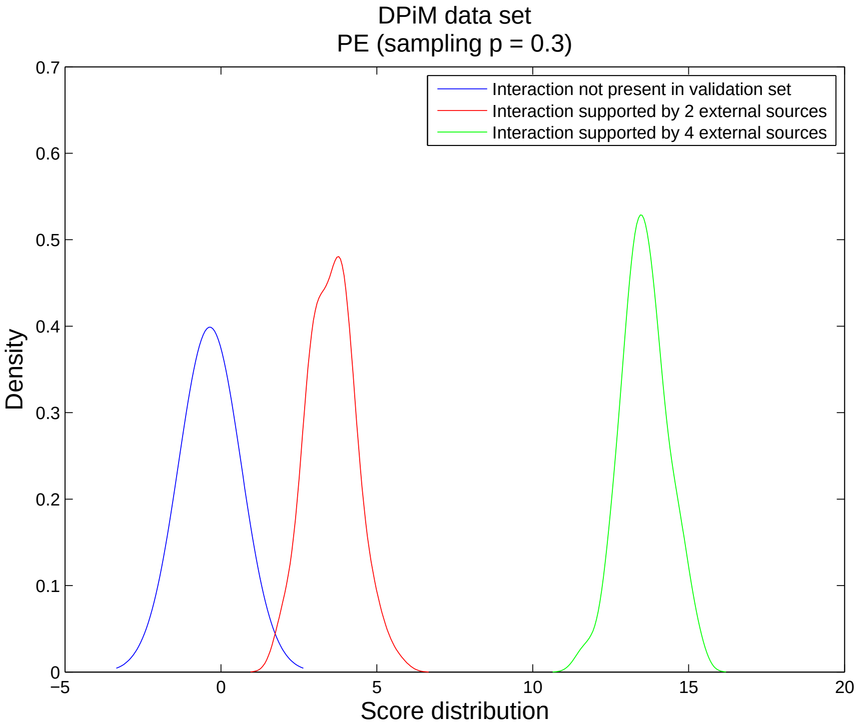


Figure A-7: Distribution of scores over multiple binary realizations representative interactions. The score distributions for low, medium and high confidence interactions overlap slightly indicating the importance of averaging over multiple samples.



Appendix B

Supporting Information for PC-Select

B.1 Model performance as the number of top SNPs to include in the GRM is varied.

We investigated model performance as the number of top SNPs, k, to include in the GRM is varied. In the following simulations, we compared using the top k SNPs in the GRM to a model using PCs with the top k SNPs. The following analysis explores the intermediate choice that the FaST-LMM Select and PC-Select methods have to make. Both methods use cross-validation predictive log-likelihood to choose k.

In the presence of population stratification and without causal SNPs, we found that no choice of top k SNPs is sufficient to correct for population stratification, except when all SNPs are used in the GRM (Figure B-1). This illustrates the tension between using a subset of SNPs in the GRM to increase power and the need to use all SNPs to correct for population stratification. On the other hand, when using PCs, statistics were not inflated for any choices of k.

In the absence of population stratification, including PCs does not compromise power. The power when using PCs with the top k SNPs is not significantly different than when using the top k SNPs (Figure B-2).

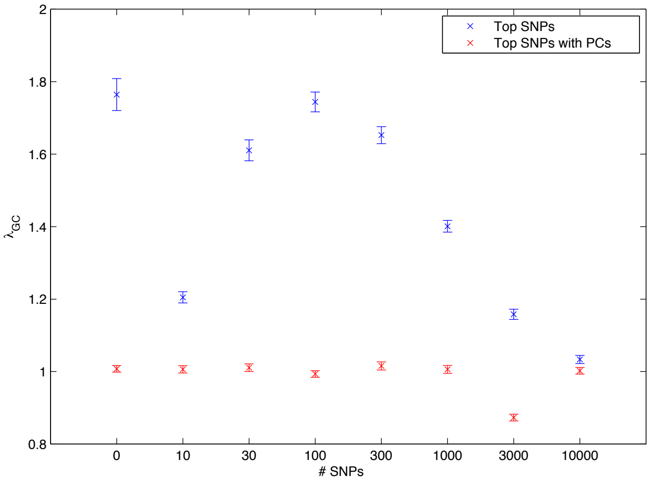


Figure B-1: Comparison of inflation when using the top k SNPs in the GRM and when using PCs with the top k SNPs in the GRM. Two populations are simulated with $F_{st} = 0.05$ and no SNPs are causal. Without PCs, the only choice of k that is not significantly inflated is using all SNPs. With PCs, no choice of k is inflated.

In the presence of population stratification and casual SNPs, we find that when few SNPs are causal ($p = 0.005$), using a subset of SNPs increases power over standard LMM as previously reported [83]. However, in this regime, using the top k SNPs inflates null statistics (Figure B-3). With PCs, there were choices of k that improved power over standard LMM, while at the same time avoiding inflating null statistics.

B.2 Implementation

We suggest implementing PC-Select by extracting PCs from the genotype data using EIGENSOFT [100] and then running FaST-LMM Select [82, 84, 83] with REML using the PCs as fixed effects.

For large datasets, we found that FaST-LMM Select exhausted our 170-GB memory limit, so we provide a memory efficient MATLAB implementation of the cross-validation step to select k. Then using GCTA [152], the SNPs can be sorted by linear regression p-value, a truncated GRM using the top k SNPs can be formed, and association statistics can be computed using GCTA mlma-loco with a GRM consisting only of the top k SNPs. In all steps, the PCs are included as fixed effects as well as

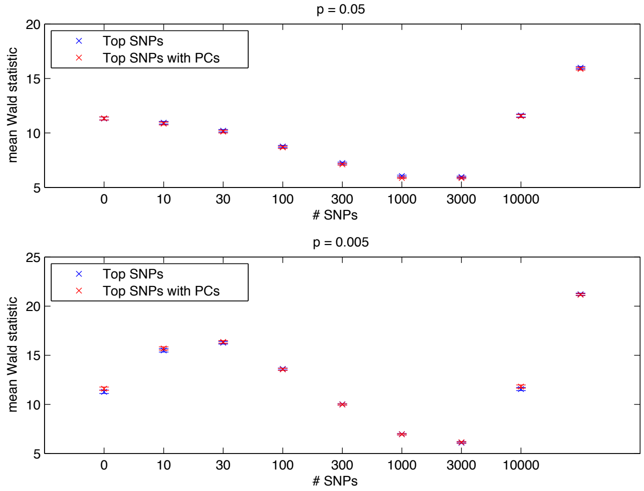


Figure B-2: Comparison of power when using the top k SNPs in the GRM and when using PCs with the top k SNPs in the GRM. A fraction $p = 0.05, 0.005$ of the SNPs were randomly chosen as causal and population stratification was not present. The last unlabeled points result from using only truly causal SNPs to construct the GRM. It represents the highest achievable score. In all cases, the power is not significantly different between the two methods.

any additional covariates.

EIGENSOFT is available at: `http://www.hsph.harvard.edu/alkes-price/software/`

FaST-LMM Select is available at: `http://research.microsoft.com/en-us/um/redmond/projects/mscompbio/fastlmm/`

MATLAB data simulators, analysis pipeline, and cross-validation implementation are available at: `http://groups.csail.mit.edu/cb/pc-select/`

GCTA is available at: `http://www.complextraitgenomics.com/software/gcta/download.html`

NHGRI SNP	Tag SNP	Chr.	Position	PC-Select	FaST-LMM Select
rs4648356	rs4648356	1	2699024	57.5983	54.1676
rs233100	rs233100	1	85544597	8.75607	8.02551
rs6604026	rs6604026	1	93076191	14.6014	13.1095
rs11581062	rs11581062	1	101180107	24.5684	24.2292
rs2300747	rs1335532	1	116902480	28.2122	25.6348
rs3761959	rs3761959	1	155935902	24.4998	21.3694
rs1323292	rs1323292	1	190807644	7.9819	7.87495
rs12466022	rs12466022	2	43212565	3.21049	2.91722
rs7595037	rs7595037	2	68500599	6.24591	6.28292
rs17174870	rs17174870	2	112381672	20.6672	19.6121
rs10201872	rs10201872	2	230814968	8.58927	7.21495
rs9821630	rs9821630	3	16945942	4.62683	4.48627
rs11129295	rs11129295	3	27763784	13.9527	11.6844
rs669607	rs669607	3	28046448	12.5163	10.2774
rs771767	rs771767	3	103231328	6.03805	5.39359
rs2293370	rs2293370	3	120702624	29.6184	27.8487
rs4285028	rs4285028	3	123143354	2.48924	1.93022
rs4308217	rs4308217	3	123275877	1.59951	1.47762
rs9282641	rs9282641	3	123279458	15.2174	14.883
rs908821	rs908821	3	142023408	1.65432	1.67819
rs1841770	rs1841770	3	149239376	0.148637	0.415421
rs2243123	rs2243123	3	161192345	5.57165	5.2983
rs10936599	rs10936599	3	170974795	20.7171	20.6344
rs228614	rs228614	4	103797685	1.69184	1.48077
rs12644284	rs12644284	4	154373450	0.621968	0.58592
rs7672826	rs7672826	4	182636689	0.195435	0.136328
rs6897932	rs6897932	5	35910332	26.7281	26.3093
rs756699	rs756699	5	133474474	4.17009	3.96739

NHGRI SNP	Tag SNP	Chr.	Position	PC-Select	FaST-LMM Select
rs1062158	rs1062158	5	141503184	10.9026	10.5473
rs2546890	rs2546890	5	158692478	15.9447	16.2163
rs4075958	rs4075958	5	176717118	10.6018	9.93475
rs11755724	rs11755724	6	7063989	4.02806	4.53184
rs11962089	rs11962089	6	105718913	0.43312	0.530643
rs802734	rs802734	6	128320491	7.09974	5.66315
rs9321490	rs9321490	6	135536568	9.43833	8.90168
rs11154801	rs11154801	6	135781048	27.7979	25.7072
rs17066096	rs17066096	6	137494601	14.9497	13.2316
rs13192841	rs13192841	6	138008907	4.6728	4.39535
rs1738074	rs1738074	6	159385965	26.3828	24.4227
rs6952809	rs6952809	7	2415019	11.6368	10.6263
rs758944	rs758944	7	75791233	4.57183	4.76084
rs354033	rs354033	7	148920397	7.50107	7.06653
rs1520333	rs1520333	8	79563593	7.31574	6.09822
rs2019960	rs2019960	8	129261453	0.60485	0.688187
rs2150702	rs2150702	9	5883861	2.07041	1.67457
rs1755289	rs1755289	9	17928351	1.85224	2.01786
rs290986	rs290986	9	92603357	15.5941	14.1678
rs3780792	rs3780792	9	135825164	0.2023	0.190039
rs3118470	rs3118470	10	6141719	25.7664	25.0488
rs1250550	rs1250550	10	80730323	13.7682	14.0924
rs7923837	rs7923837	10	94471897	8.35128	7.04458
rs650258	rs650258	11	60588858	17.1356	16.4211
rs4409785	rs4409785	11	94951070	10.2767	9.10435
rs630923	rs630923	11	118259563	8.59249	8.76583
rs1458175	rs1458175	12	40252128	0.261234	0.235203
rs703842	rs703842	12	56449006	27.2546	25.9998

NHGRI SNP	Tag SNP	Chr.	Position	PC-Select	FaST-LMM Select
rs9523762	rs9523762	13	92129887	0.258751	0.179013
rs4902647	rs4902647	14	68323944	6.32905	5.2318
rs2300603	rs2300603	14	75075310	16.8255	16.7838
rs2744148	rs2744148	16	1013553	9.70226	8.50068
rs7200786	rs7200786	16	11085302	40.8143	38.0727
rs386965	rs386965	16	78210042	26.3899	24.2915
rs13333054	rs13333054	16	84568534	12.8357	12.5916
rs4792814	rs4792814	17	40758788	4.16238	3.62007
rs180515	rs180515	17	55379057	16.859	16.1936
rs12456021	rs12456021	18	54364370	5.15011	4.5952
rs7238078	rs7238078	18	54535172	5.53865	4.79417
rs1077667	rs1077667	19	6619972	32.0466	29.9332
rs874628	rs874628	19	18165700	23.1209	22.2579
rs7255066	rs7255066	19	49837943	6.19959	5.82356
rs307896	rs307896	19	52353333	11.9913	11.3425
rs281380	rs281380	19	53906282	16.3993	14.4309
rs397020	rs397020	20	1153886	5.67704	5.94638
rs2283792	rs2283792	22	20461125	7.36146	6.70206
rs140522	rs140522	22	49318132	10.4006	9.15617

Table B.1: Wald statistics for 75 published associated markers in the MS data set.

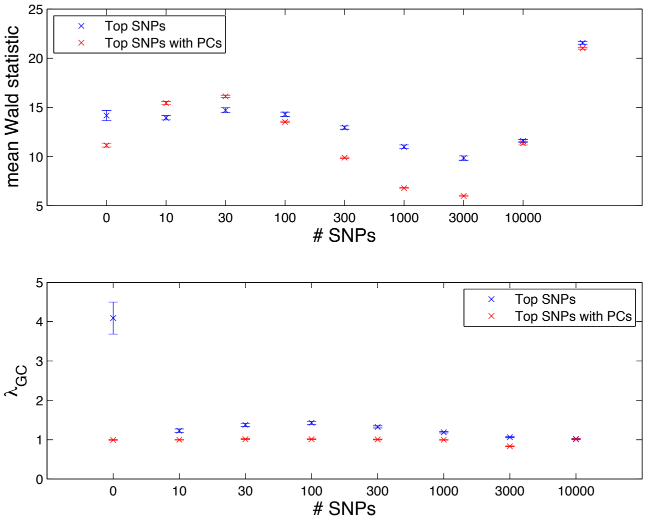


Figure B-3: Comparison of power and λ_{GC} when using the top k SNPs in the GRM and when using PCs with the top k SNPs in the GRM. Two populations were simulated with $F_{st} = 0.05$ and a randomly chosen fraction $p = 0.005$ of SNPs were chosen as causal. The top subplot measures power by mean Wald statistic on test causal SNPs and the bottom subplot measures inflation by λ_{GC} on an independent set of null test SNPs. Whenever using the top k SNPs without PCs has higher power than using PCs, it also exhibits significant inflation of λ_{GC}.

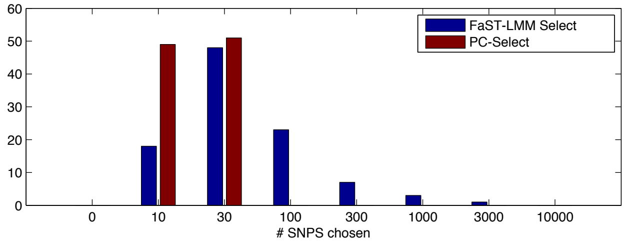


Figure B-4: Comparison of number of SNPs chosen by Fast-LMM Select and PC-Select. The histogram shows the choices made by each method over 100 simulations with population stratification and $p = 0.005$. On average PC-Select chooses fewer SNPs to include in the GRM.

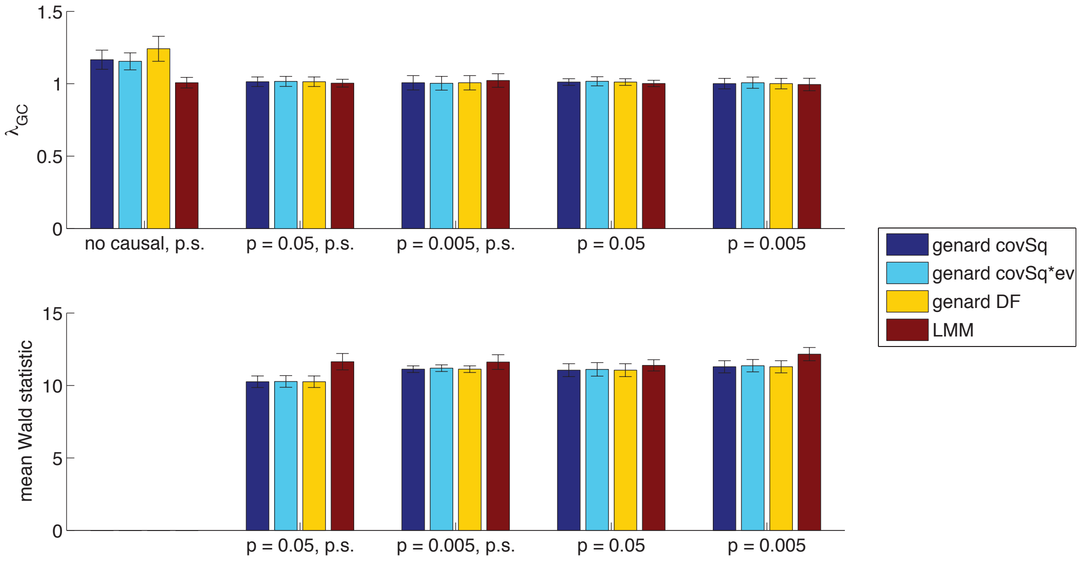


Figure B-5: Comparison of λ_{GC} and power for the genard method [57] and standard LMM on simulations with and without population stratification (abbreviated p.s.) as the fraction of casual SNPs (no causal, $p = 0.05, 0.005$) varies. As recommended by the author of the genard method, model complexity is selected by BIC and PCs are ordered by squared correlation to the phenotype (covSq), squared correlation to the phenotype multiplied by the eigenvalue (covSq*ev), and effective degrees of freedom (DF). In these simulations, genard does not provide a benefit over standard LMM.

Bibliography

[1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, February 2000.

[2] Bruno Aranda, P Achuthan, Yasmin Alam-Faruque, I Armean, Alan Bridge, C Derow, Marc Feuermann, AT Ghanbarian, Samuel Kerrien, Jyoti Khadake, et al. The intact molecular interaction database in 2010. *Nucleic acids research*, 38(suppl 1):D525–D531, 2010.

[3] H. Asha, I. Nagy, G. Kovacs, D. Stetson, I. Ando, and C. R. Dearolf. Analysis of ras-induced overproliferation in drosophila hemocytes. *Genetics*, 163(1):203–15, 2003.

[4] Dariel Ashton-Beaucage, Christian M Udell, Hugo Lavoie, Caroline Baril, Martin Lefrançois, Pierre Chagnon, Patrick Gendron, Olivier Caron-Lizotte, Éric Bonneil, Pierre Thibault, et al. The exon junction complex controls the splicing of *MAPK* and other long intron-containing transcripts in *Drosophila*. *Cell*, 143(2):251–262, 2010.

[5] S. Bandyopadhyay, C. Y. Chiang, J. Srivastava, M. Gersten, S. White, R. Bell, C. Kurschner, C. H. Martin, M. Smoot, S. Sahasrabudhe, D. L. Barber, S. K. Chanda, and T. Ideker. A human map kinase interactome. *Nature Methods*, 7(10):801–5.

[6] Miriam Barrios-Rodiles, Kevin R Brown, Barish Ozdamar, Rohit Bose, Zhong Liu, Robert S Donovan, Fukiko Shinjo, Yongmei Liu, Joanna Dembowy, Ian W Taylor, et al. High-throughput mapping of a dynamic signaling network in mammalian cells. *Science*, 307(5715):1621–1625, 2005.

[7] C. Behrends, M. E. Sowa, S. P. Gygi, and J. W. Harper. Network organization of the human autophagy system. *Nature*, 466(7302):68–76.

[8] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[9] Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[10] D. A. Blanchard, S. Mouhamad, M. T. Auffredou, A. Pesty, J. Bertoglio, G. Leca, and A. Vazquez. Cdk2 associates with map kinase in vivo and its nuclear translocation is dependent on map kinase activation in il-2-dependent kit 225 t lymphocytes. *Oncogene*, 19(36):4184–9, 2000.

[11] T. Bouwmeester, A. Bauch, H. Ruffner, P. O. Angrand, G. Bergamini, K. Croughton, C. Cruciat, D. Eberhard, J. Gagneur, S. Ghidelli, C. Hopf, B. Huhse, R. Mangano, A. M. Michon, M. Schirle, J. Schlegl, M. Schwab, M. A. Stein, A. Bauer, G. Casari, G. Drewes, A. C. Gavin, D. B. Jackson, G. Joberty, G. Neubauer, J. Rick, B. Kuster, and G. Superti-Furga. A physical and functional map of the human tnf-alpha/nf-kappa b signal transduction pathway. *Nature Cell Biology*, 6(2):97–105, 2004.

[12] Ashton Breitkreutz, Hyungwon Choi, Jeffrey R Sharom, Lorrie Boucher, Victor Neduva, Brett Larsen, Zhen-Yuan Lin, Bobby-Joe Breitkreutz, Chris Stark, Guomin Liu, et al. A global protein kinase and phosphatase interaction network in yeast. *Science*, 328(5981):1043–1046, 2010.

[13] Rachel B. Brem, Gal Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, 2002.

[14] Giulia Calloni, Taotao Chen, Sonya M Schermann, Hung-chun Chang, Pierre Genevaux, Federico Agostini, Gian Gaetano Tartaglia, Manajit Hayer-Hartl, and F Ulrich Hartl. Dnak functions as a central hub in the *E. coli* chaperone network. *Cell reports*, 1(3):251–264, 2012.

[15] Benjamin J Cargile, Jonathan L Bundy, and James L Stephenson. Potential for false positive identifications from large databases through tandem mass spectrometry. *Journal of Proteome Research*, 3(5):1082–1085, 2004.

[16] Carolina Cela and Marta Llimargas. Egfr is essential for maintaining epithelial integrity during tracheal remodelling in drosophila. *Development*, 133(16):3115–3125, 2006.

[17] S. E. Celniker, L. A. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. MacAlpine, G. Micklem, F. Piano, M. Snyder, L. Stein, K. P. White, and R. H. Waterston. Unlocking the secrets of the genome. *Nature*, 459(7249):927–30, 2009.

[18] Arnaud Ceol, Andrew Chatr Aryamontri, Luana Licata, Daniele Peluso, Leonardo Briganti, Livia Perfetto, Luisa Castagnoli, and Gianni Cesareni. Mint, the molecular interaction database: 2009 update. *Nucleic acids research*, 38(suppl 1):D532–D539, 2010.

[19] Andrew Chatr-aryamontri, Bobby-Joe Breitkreutz, Sven Heinicke, Lorrie Boucher, Andrew Winter, Chris Stark, Julie Nixon, Lindsay Ramage, Nadine Kolas, Lara ODonnell, et al. The biogrid interaction database: 2013 update. *Nucleic acids research*, 41(D1):D816–D823, 2013.

[20] Bo-Juen Chen, Helen C Causton, Denesy Mancenido, Noel L Goddard, Ethan O Perlstein, and Dana Pe'er. Harnessing gene expression to identify the genetic basis of drug resistance. *Molecular Systems Biology*, 5:310, January 2009.

[21] Yanqing Chen, Jun Zhu, Pek Y. Lum, Xia Yang, Shirly Pinto, Douglas J. MacNeil, Chunsheng Zhang, John Lamb, Stephen Edwards, Solveig K. Sieberts, Amy Leonardson, Lawrence W. Castellini, Susanna Wang, Marie-France Champy, Bin Zhang, Valur Emilsson, Sudheer Doss, Anatole Ghazalpour, Steve Horvath, Thomas A. Drake, Aldons J. Lusis, and Eric E. Schadt. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435, March 2008.

[22] S. Chien, L. T. Reiter, E. Bier, and M. Gribskov. Homophila: human disease gene cognates in drosophila. *Nucleic acids research*, 30(1):149–51, 2002.

[23] Hyungwon Choi, Timo Glatter, Mathias Gstaiger, and Alexey I Nesvizhskii. Saint-ms1: Protein–protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *Journal of Proteome Research*, 11(4):2619–2624, 2012.

[24] Hyungwon Choi, Sinae Kim, Anne-Claude Gingras, and Alexey I Nesvizhskii. Analysis of protein complexes through model-based biclustering of label-free quantitative ap-ms data. *Molecular Systems Biology*, 6(1), 2010.

[25] Hyungwon Choi, Brett Larsen, Zhen-Yuan Lin, Ashton Breitkreutz, Dattatreya Mellacheruvu, Damian Fermin, Zhaohui S Qin, Mike Tyers, Anne-Claude Gingras, and Alexey I Nesvizhskii. Saint: probabilistic scoring of affinity purification-mass spectrometry data. *Nature Methods*, 8(1):70–73, 2010.

[26] H. R. Christofk, M. G. Vander Heiden, M. H. Harris, A. Ramanathan, R. E. Gerszten, R. Wei, M. D. Fleming, S. L. Schreiber, and L. C. Cantley. The m2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature*, 452(7184):230–3, 2008.

[27] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, et al. Integration of biological networks and gene expression data using cytoscape. *Nature Protocols*, 2(10):2366–2382, 2007.

[28] Sean R Collins, Patrick Kemmeren, Xue-Chu Zhao, Jack F Greenblatt, Forrest Spencer, Frank CP Holstege, Jonathan S Weissman, and Nevan J Krogan. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Molecular & Cellular Proteomics*, 6(3):439–450, 2007.

[29] Alberto de la Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

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

[31] B Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[32] Georg Dietzl, Doris Chen, Frank Schnorrer, Kuan-Chung Su, Yulia Barinova, Michaela Fellner, Beate Gasser, Kaolin Kinsey, Silvia Oppel, Susanne Scheiblauer, et al. A genome-wide transgenic rnai library for conditional gene inactivation in drosophila. *Nature*, 448(7150):151–156, 2007.

[33] Christophe J Echeverri, Philip A Beachy, Buzz Baum, Michael Boutros, Frank Buchholz, Sumit K Chanda, Julian Downward, Jan Ellenberg, Andrew G Fraser, Nir Hacohen, et al. Minimizing the risk of reporting false positives in large-scale rnai screens. *Nature methods*, 3(10):777–779, 2006.

[34] Frank J Echtenkamp and Brian C Freeman. Expanding the cellular molecular chaperone network through the ubiquitous cochaperones. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1823(3):668–673, 2012.

[35] Frank J Echtenkamp, Elena Zelin, Ellinor Oxelmark, Joyce I Woo, Brenda J Andrews, Michael Garabedian, and Brian C Freeman. Global functional map of the p23 molecular chaperone reveals an extensive cellular network. *Molecular Cell*, 43(2):229–241, 2011.

[36] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–8, 1998.

[37] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S. Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G. Bragi Walters, Steinunn Gunnarsdottir, Magali Mouy, Valgerdur Steinthorsdottir, Gudrun H. Eiriksdottir, Gyda Bjornsdottir, Inga Reynisdottir, Daniel Gudbjartsson, Anna Helgadottir, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Unnur Styrkarsdottir, Solveig Gretarsdottir, Kristinn P. Magnusson, Hreinn Stefansson, Ragnheidur Fossdal, Kristleifur Kristjansson, Hjortur G. Gislason, Tryggvi Stefansson, Bjorn G. Leifsson, Unnur Thorsteinsdottir, John R. Lamb, Jeffrey R. Gulcher,

Marc L. Reitman, Augustine Kong, Eric E. Schadt, and Kari Stefansson. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428, March 2008.

[38] B. Estrada, S. E. Choe, S. S. Gisselbrecht, S. Michaud, L. Raj, B. W. Busser, M. S. Halfon, G. M. Church, and A. M. Michelson. An integrated strategy for analyzing the unique developmental programs of different myoblast subtypes. *PLoS Genetics*, 2(2):e16, 2006.

[39] Marie Evangelista, Tze Yang Lim, James Lee, Leon Parker, Amir Ashique, Andrew S Peterson, Weilan Ye, David P Davis, and Frederic J de Sauvage. Kinome sirna screen identifies regulators of ciliogenesis and hedgehog signal transduction. *Science Signaling*, 1(39):ra7, 2008.

[40] Brian C Freeman, David O Toft, and Richard I Morimoto. Molecular chaperone machines: chaperone activities of the cyclophilin cyp-40 and the steroid aporeceptor-associated protein p23. *Science*, 274(5293):1718–1720, 1996.

[41] A. Friedman and N. Perrimon. A functional rnai screen for regulators of receptor tyrosine kinase and erk signalling. *Nature*, 444(7116):230–4, 2006.

[42] A. Friedman and N. Perrimon. High-throughput approaches to dissecting mapk signaling pathways. *Methods*, 40(3):262–71, 2006.

[43] A. Friedman and N. Perrimon. Genetic screening for signal transduction in the era of network biology. *Cell*, 128(2):225–31, 2007.

[44] Adam A Friedman, George Tucker, Rohit Singh, Dong Yan, Arunachalam Vinayagam, Yanhui Hu, Richard Binari, Pengyu Hong, Xiaoyun Sun, Maura Porto, et al. Proteomic and functional genomic landscape of receptor tyrosine kinase and ras to extracellular signal-regulated kinase signaling. *Science Signaling*, 4(196):rs10, 2011.

[45] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[46] George M. Furnival and Robert W. Jr. Wilson. Regression by leaps and bounds. *Technometrics*, 16(4):499–511, 1974.

[47] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–6, 2006.

[48] Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars Juhl Jensen, Sonja Bastuck, Birgit Dümpelfeld, Angela Edelmann, Marie-Anne Heurtier, Verena Hoffman, Christian Hoefert, Karin Klein, Manuela Hudak, Anne-Marie Michon, Malgorzata Schelder, Markus Schirle, Marita Remor, Tatjana Rudi, Sean Hooper, Andreas Bauer, Tewis Bouwmeester, Georg Casari, Gerard Drewes, Gitte Neubauer, Jens M. Rick, Bernhard Kuster, Peer Bork, Robert B. Russell, and Giulio Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

[49] Anne-Claude Gavin, Markus Bösche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jörg Schultz, Jens M. Rick, Anne-Marie Michon, Cristina-Maria Cruciat, Marita Remor, Christian Höfert, Malgorzata Schelder, Miro Brajenovic, Heinz Ruffner, Alejandro Merino, Karin Klein, Manuela Hudak, David Dickson, Tatjana Rudi, Volker Gnau, Angela Bauch, Sonja Bastuck, Bettina Huhse, Christina Leutwein, Marie-Anne Heurtier, Richard R. Copley, Angela Edelmann, Erich Querfurth, Vladimir Rybin, Gerard Drewes, Manfred Raida, Tewis Bouwmeester, Peer Bork, Bertrand Seraphin, Bernhard Kuster, Gitte Neubauer, and Giulio Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[50] Guy Geva and Roded Sharan. Identification of protein complexes from co-immunoprecipitation data. *Bioinformatics*, 27(1):111–117, 2011.

[51] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–1736, 2003.

[52] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[53] K. G. Guruharsha, Jean-François Rual, Bo Zhai, Julian Mintseris, Pujita Vaidya, Namita Vaidya, Chapman Beekman, Christina Wong, Odise Cenaj, Emily McKillip, Saumini Shah, Mark Stapleton, Charles Yu, Bayan Parsa, Xiao Chen, Bhaveen Kapadia, K. VijayRaghavan, and Spyros Artavanis-Tsakonas. A protein complex network of *Drosophila melanogaster*. *Cell*, 147(3):690–703, 2011.

[54] G Traver Hart, Insuk Lee, and Edward M Marcotte. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, 8(1):236, 2007.

[55] F Ulrich Hartl, Andreas Bracher, and Manajit Hayer-Hartl. Molecular chaperones in protein folding and proteostasis. *Nature*, 475(7356):324–332, 2011.

[56] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary D. Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier, Lingyun Yang, Cheryl Wolting, Ian Donaldson, Sren Schandorff, Juanita Shewnarane, Mai Vo, Joanne Taggart, Marilyn Goudreault, Brenda Muskat, Cris Alfarano, Danielle Dewar, Zhen Lin, Katerina Michalickova, Andrew R. Willems, Holly Sassi, Peter A. Nielsen, Karina J. Rasmussen, Jens R. Andersen, Lene E. Johansen, Lykke H. Hansen, Hans Jespersen, Alexandre Podtelejnikov, Eva Nielsen, Janne Crawford, Vibeke Poulsen, Birgitte D. Sørensen, Jesper Matthiesen, Ronald C. Hendrickson, Frank Gleeson, Tony Pawson, Michael F. Moran, Daniel Durocher, Matthias Mann, Christopher W. V. Hogue, Daniel Figeys, and Mike Tyers. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):180–183, 2002.

[57] Gabriel E Hoffman. Correcting for population structure and kinship using the linear mixed model: Theory and extensions. *PloS ONE*, 8(10):e75707, 2013.

[58] Yanhui Hu, Ian Flockhart, Arunachalam Vinayagam, Clemens Bergwitz, Bonnie Berger, Norbert Perrimon, and Stephanie E Mohr. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*, 12(1):357, 2011.

[59] Shao-shan Carol Huang and Ernest Fraenkel. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Science Signaling*, 2(81):ra40, 2009.

[60] R. A. Hunt, W. Edris, P. K. Chanda, B. Nieuwenhuijsen, and K. H. Young. Snapin interacts with the n-terminus of regulator of g protein signaling 7. *Biochemical and biophysical research communications*, 303(2):594–9, 2003.

[61] Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

[62] Stefanie Jäger, Peter Cimermancic, Natali Gulbahce, Jeffrey R. Johnson, Kathryn E. McGovern, Starlynn C. Clarke, Michael Shales, Gaelle Mercenne, Lars Pache, Kathy Li, Hilda Hernandez, Gwendolyn M. Jang, Shoshannah L. Roth, Eyal Akiva, John Marlett, Melanie Stephens, Iván D'Orso, Jason Fernandes, Marie Fahey, Cathal Mahon, Anthony J. O'Donoghue, Aleksandar Todorovic, John H. Morris, David A. Maltby, Tom Alber, Gerard Cagney, Frederic D.

Bushman, John A. Young, Sumit K. Chanda, Wesley I. Sundquist, Tanja Kortemme, Ryan D. Hernandez, Charles S. Craik, Alma Burlingame, Andrej Sali, Alan D. Frankel, and Nevan J. Krogan. Global landscape of hiv-human protein complexes. *Nature*, 481(7381):365–370, 2011.

[63] Luc Janss, Gustavo de los Campos, Nuala Sheehan, and Daniel Sorensen. Inferences from genomic models in stratified populations. *Genetics*, 192(2):693–704, 2012.

[64] Katherine C Jordan, Steven D Hatfield, Michael Tworoger, Ellen J Ward, Karin A Fischer, Stuart Bowers, and Hannele Ruohola-Baker. Genome wide analysis of transcript levels after perturbation of the egfr pathway in the drosophila ovary. *Developmental dynamics*, 232(3):709–724, 2005.

[65] Harm H Kampinga and Elizabeth A Craig. The hsp70 chaperone machinery: J proteins as drivers of functional specificity. *Nature Reviews Molecular Cell Biology*, 11(8):579–592, 2010.

[66] Hyun Min Kang, Jae Hoon Sul, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 2010.

[67] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[68] C. S. Karapetis, S. Khambata-Ford, D. J. Jonker, C. J. O'Callaghan, D. Tu, N. C. Tebbutt, R. J. Simes, H. Chalchal, J. D. Shapiro, S. Robitaille, T. J. Price, L. Shepherd, H. J. Au, C. Langer, M. J. Moore, and J. R. Zalcberg. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine*, 359(17):1757–65, 2008.

[69] Michael J Kerner, Dean J Naylor, Yasushi Ishihama, Tobias Maier, Hung-Chun Chang, Anna P Stines, Costa Georgopoulos, Dmitrij Frishman, Manajit Hayer-Hartl, Matthias Mann, et al. Proteome-wide analysis of chaperonin-dependent protein folding in *Escherichia coli*. *Cell*, 122(2):209–220, 2005.

[70] Samuel Kerrien, Sandra Orchard, Luisa Montecchi-Palazzi, Bruno Aranda, Antony F Quinn, Nisha Vinod, Gary D Bader, Ioannis Xenarios, Jérôme Wojcik, David Sherman, et al. Broadening the horizon–level 2.5 of the hupo-psi format for molecular interactions. *BMC Biology*, 5(1):44, 2007.

[71] Yoko Kimura, Suzanne L Rutherford, Yoshihiko Miyata, Ichiro Yahara, Brian C Freeman, Lin Yue, Richard I Morimoto, and Susan Lindquist. Cdc37 is a molecular chaperone with specific functions in signal transduction. *Genes & Development*, 11(14):1775–1785, 1997.

[72] L. Kockel, K. S. Kerr, M. Melnick, K. Bruckner, M. Hebrok, and N. Perrimon. Dynamic switch of negative feedback regulation in drosophila akt-tor signaling. *PLoS Genetics*, 6(6):e1000990.

[73] W. Kolch and A. Pitt. Functional proteomics to dissect tyrosine kinase signalling pathways in cancer. *Nature Reviews Cancer*, 10(9):618–29.

[74] I. Kratchmarova, B. Blagoev, M. Haack-Sorensen, M. Kassem, and M. Mann. Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. *Science*, 308(5727):1472–7, 2005.

[75] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–43, 2006.

[76] Zoltan Kutalik, Jacques S. Beckmann, and Sven Bergmann. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nature Biotechnology*, 26(5):531–539, May 2008.

[77] Mathieu Lavallée-Adam, Philippe Cloutier, Benoit Coulombe, and Mathieu Blanchette. Modeling contaminants in ap-ms/ms experiments. *Journal of Proteome Research*, 10(2):886–895, 2010.

[78] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.

[79] Guillaume Lettre, Cameron D Palmer, Taylor Young, Kenechi G Ejebe, Hooman Allayee, Emelia J Benjamin, Franklyn Bennett, Donald W Bowden, Aravinda Chakravarti, Al Dreisbach, et al. Genome-wide association study of coronary heart disease and its risk factors in 8,090 african americans: the nhlbi care project. *PLoS Genetics*, 7(2):e1001300, 2011.

[80] Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, 2009.

[81] M. Lin, D. R. Sutherland, W. Horsfall, N. Totty, E. Yeo, R. Nayar, X. F. Wu, and A. C. Schuh. Cell surface antigen cd109 is a novel member of the alpha(2) macroglobulin/c3, c4, c5 family of thioester-containing proteins. *Blood*, 99(5):1683–91, 2002.

[82] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

[83] Christoph Lippert, Gerald Quon, Eun Yong Kang, Carl M Kadie, Jennifer Listgarten, and David Heckerman. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific Reports*, 3, 2013.

[84] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. Improved linear mixed models for genome-wide association studies. *Nature Methods*, 9(6):525–526, 2012.

[85] Po-Ru Loh, George Tucker, and Bonnie Berger. Phenotype prediction using regularized regression on genetic data in the dream5 systems genetics b challenge. *PloS ONE*, 6(12):e29095, 2011.

[86] M. B. Major, B. S. Roberts, J. D. Berndt, S. Marine, J. Anastas, N. Chung, M. Ferrer, X. Yi, C. L. Stoick-Cooper, P. D. von Haller, L. Kategaya, A. Chien, S. Angers, M. MacCoss, M. A. Cleary, W. T. Arthur, and R. T. Moon. New regulators of wnt/beta-catenin signaling revealed by integrative molecular screening. *Science Signaling*, 1(45):ra12, 2008.

[87] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[88] Marc L Mendillo, Sandro Santagata, Martina Koeva, George W Bell, Rong Hu, Rulla M Tamimi, Ernest Fraenkel, Tan A Ince, Luke Whitesell, and Susan Lindquist. Hsf1 drives a transcriptional program distinct from heat shock to support highly malignant human cancers. *Cell*, 150(3):549–562, 2012.

[89] Martin Lee Miller, Lars Juhl Jensen, Francesca Diella, Claus Jorgensen, Michele Tinti, Lei Li, Marilyn Hsiung, Sirlester A Parker, Jennifer Bordeaux, Thomas Sicheritz-Ponten, et al. Linear motif atlas for phosphorylation-dependent signaling. *Science Signaling*, 1(35):ra2, 2008.

[90] Thilakam Murali, Svetlana Pacifico, Jingkai Yu, Stephen Guest, George G Roberts, and Russell L Finley. Droid 2011: a comprehensive, integrated resource for protein, transcription factor, rna and gene interactions for drosophila. *Nucleic acids research*, 39(suppl 1):D736–D743, 2011.

[91] Alexey I Nesvizhskii. Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. *Proteomics*, 12(10):1639–1655, 2012.

[92] J. Q. Ni, M. Markstein, R. Binari, B. Pfeiffer, L. P. Liu, C. Villalta, M. Booker, L. Perkins, and N. Perrimon. Vector and parameters for targeted transgenic rna interference in drosophila melanogaster. *Nature Methods*, 5(1):49–51, 2008.

[93] Stéphane Ory, Ming Zhou, Thomas P Conrads, Timothy D Veenstra, and Deborah K Morrison. Protein phosphatase 2a positively regulates ras signaling by dephosphorylating ksr1 and raf-1 on critical 14-3-3 binding sites. *Current Biology*, 13(16):1356–1364, 2003.

[94] Svetlana Pacifico, Guozhen Liu, Stephen Guest, Jodi R Parrish, Farshad Fotouhi, and Russell L Finley. A database and tool, im browser, for exploring and integrating emerging gene and protein interaction data for drosophila. *BMC bioinformatics*, 7(1):195, 2006.

[95] Barry Panaretou, Giuliano Siligardi, Philippe Meyer, Alison Maloney, Janis K Sullivan, Shradha Singh, Stefan H Millson, Paul A Clarke, Soren Naaby-Hansen, Rob Stein, et al. Activation of the atpase activity of hsp90 by the stress-regulated cochaperone aha1. *Molecular Cell*, 10(6):1307–1318, 2002.

[96] H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[97] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.

[98] Evan T Powers and William E Balch. Diversity in the origins of proteostasis networksa driver for protein function in evolution. *Nature Reviews Molecular Cell Biology*, 14(4):237–248, 2013.

[99] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772, 2009.

[100] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

[101] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[102] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. Response to sul and eskin. *Nature Reviews Genetics*, 14(4):300–300, 2013.

[103] Robert J Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K Sorger, Leonidas G Alexopoulos, Xiaowei Xue, Neil D Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE*, 5(2):e9202, January 2010.

[104] Matthew V. Rockman. Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature*, 456(7223):738–744, 2008.

[105] Assen Roguev, Dale Talbot, Gian Luca Negri, Michael Shales, Gerard Cagney, Sourav Bandyopadhyay, Barbara Panning, and Nevan J Krogan. Quantitative genetic-interaction mapping in mammalian cells. *Nature methods*, 10(5):432–437, 2013.

[106] A.D. Roses. Pharmacogenetics and drug development: the path to safer and more effective drugs. *Nature Reviews Genetics*, 5(9):645–56, 2004.

[107] Ingo Ruczinski, Charles Kooperberg, and Michael LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, September 2003.

[108] AI Saeed, Vasily Sharov, Joe White, Jerry Li, Wei Liang, Nirmal Bhagabati, J Braisted, M Klapa, T Currier, M Thiagarajan, et al. Tm4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2):374, 2003.

[109] Helen Saibil. Chaperone machines for protein folding, unfolding and disaggregation. *Nature Reviews Molecular Cell Biology*, 14(10):630–642, 2013.

[110] Alok J Saldanha. Java treeviewextensible visualization of microarray data. *Bioinformatics*, 20(17):3246–3248, 2004.

[111] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(suppl 1):D449–D451, 2004.

[112] Mihaela E Sardiu, Yong Cai, Jingji Jin, Selene K Swanson, Ronald C Conaway, Joan W Conaway, Laurence Florens, and Michael P Washburn. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proceedings of the National Academy of Sciences*, 105(5):1454–1459, 2008.

[113] Mihaela E Sardiu, Laurence Florens, and Michael P Washburn. Evaluation of clustering algorithms for protein complex and protein interaction network assembly. *Journal of Proteome Research*, 8(6):2944–2952, 2009.

[114] S Sawcer, G Hellenthal, M Pirinen, CC Spencer, NA Patsopoulos, L Moutsianas, A Dilthey, Z Su, C Freeman, SE Hunt, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–219, 2011.

[115] Eric E. Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj Guhathakurta, Solveig K. Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, Pek Y. Lum, Amy Leonardson, Rolf Thieringer, Joseph M. Metzger, Liming Yang, John Castle, Haoyuan Zhu, Shera F. Kash, Thomas A. Drake, Alan Sachs, and Aldons J. Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717, July 2005.

[116] Eric E. Schadt, Stephanie A. Monks, Thomas A. Drake, Aldons J. Lusis, Nam Che, Veronica Colinayo, Thomas G. Ruff, Stephen B. Milligan, John R. Lamb, Guy Cavet, Peter S. Linsley, Mao Mao, Roland B. Stoughton, and Stephen H. Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, March 2003.

[117] Vincent Segura, Bjarni J Vilhjálmsson, Alexander Platt, Arthur Korte, Ümit Seren, Quan Long, and Magnus Nordborg. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, 44(7):825–830, 2012.

[118] Young Sik Lee and Richard W Carthew. Making a better rnai vector for *Drosophila*: use of intron spacers. *Methods*, 30(4):322–329, 2003.

[119] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

[120] Mikiko C. Siomi and Haruhiko Siomi. Identification of components of rnai pathways using the tandem affinity purification method. In GordonG. Carmichael, editor, *RNA Silencing*, volume 309 of *Methods in Molecular Biology*, pages 1–9. Humana Press, 2005.

[121] T. Sjoblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, and V. E. Velculescu. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797):268–74, 2006.

[122] Holger Sondermann, Clemens Scheufler, Christine Schneider, Jörg Höhfeld, F-Ulrich Hartl, and Ismail Moarefi. Structure of a bag/hsc70 complex: convergent functional evolution of hsp70 nucleotide exchange factors. *Science*, 291(5508):1553–1557, 2001.

[123] Daniel R Southworth and David A Agard. Client-loading conformation of the hsp90 molecular chaperone revealed in the cryo-em structure of the human hsp90: Hop complex. *Molecular Cell*, 42(6):771–781, 2011.

[124] Mathew E Sowa, Eric J Bennett, Steven P Gygi, and J Wade Harper. Defining the human deubiquitinating enzyme interaction landscape. *Cell*, 138(2):389–403, 2009.

[125] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[126] B. Stefansson and D. L. Brautigan. Protein phosphatase 6 subunit with conserved sit4-associated protein domain targets iκbϵ. *J Biol Chem*, 281(32):22624–34, 2006.

[127] J. Stelling, U. Sauer, Z. Szallasi, 3rd Doyle, F. J., and J. Doyle. Robustness of cellular functions. *Cell*, 118(6):675–85, 2004.

[128] Alexey Stukalov, Giulio Superti-Furga, and Jacques Colinge. Deconvolution of targeted protein–protein interaction maps. *Journal of Proteome Research*, 11(8):4102–4109, 2012.

[129] Jae Hoon Sul and Eleazar Eskin. Mixed models can correct for population structure for genomic regions under selection. *Nature Reviews Genetics*, 14(4):300–300, 2013.

[130] Xiaoyun Sun, Pengyu Hong, M. Kulkarni, Young Kwon, and N. Perrimon. An advanced method for identifying protein-protein interaction by analyzing tap/ms data. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1–6, 2012.

[131] Gulnara R Svishcheva, Tatiana I Axenovich, Nadezhda M Belonogova, Cornelia M van Duijn, and Yurii S Aulchenko. Rapid variance components-based method for whole-genome association analysis. *Nature Genetics*, 2012.

[132] Mikko Taipale, Daniel F Jarosz, and Susan Lindquist. Hsp90 at the hub of protein homeostasis: emerging mechanistic insights. *Nature Reviews Molecular cell biology*, 11(7):515–528, 2010.

[133] Mikko Taipale, Irina Krykbaeva, Martina Koeva, Can Kayatekin, Kenneth D Westover, Georgios I Karras, and Susan Lindquist. Quantitative analysis of hsp90-client interactions reveals principles of substrate recognition. *Cell*, 150(5):987–1001, 2012.

[134] Mikko Taipale, Irina Krykbaeva, Luke Whitesell, Sandro Santagata, Jianming Zhang, Qingsong Liu, Nathanael S Gray, and Susan Lindquist. Chaperones as thermodynamic sensors of drug-target interactions reveal kinase inhibitor specificities in living cells. *Nature biotechnology*, 31(7):630–637, 2013.

[135] Chris Soon Heng Tan, Bernd Bodenmiller, Adrian Pasculescu, Marko Jovanovic, Michael O Hengartner, Claus Jorgensen, Gary D Bader, Ruedi Aebersold, Tony Pawson, and Rune Linding. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Science Signaling*, 2(81):ra39, 2009.

[136] Kristen Tenney, Mark Gerber, Anne Ilvarsonn, Jessica Schneider, Maria Gause, Dale Dorsett, Joel C Eissenberg, and Ali Shilatifard. Drosophila rtf1 functions in histone methylation, gene expression, and notch signaling. *Proceedings of the National Academy of Sciences*, 103(32):11970–11974, 2006.

[137] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[138] Jane Trepel, Mehdi Mollapour, Giuseppe Giaccone, and Len Neckers. Targeting the dynamic hsp90 complex in cancer. *Nature Reviews Cancer*, 10(8):537–549, 2010.

[139] George Tucker, Po-Ru Loh, and Bonnie Berger. A sampling framework for incorporating quantitative mass spectrometry data in protein interaction analysis. *BMC bioinformatics*, 14(1):299, 2013.

[140] George Tucker, Alkes L. Price, and Bonnie A. Berger. Improving the power of gwas and avoiding confounding from population stratification with pc-select. *Genetics*, 2014.

[141] Peter Uetz, Loic Giot, Gerard Cagney, Traci A. Mansfield, Richard S. Judson, James R. Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, Alia Qureshi-Emili, Ying Li, Brian Godwin, Diana Conover, Theodore Kalbfleisch, Govindan Vijayadamodar, Meijia Yang, Mark Johnston, Stanley Fields, and Jonathan M. Rothberg. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.

[142] Laura J. van 't Veer, Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, René Bernards, and Stephen H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, January 2002.

[143] A. Veraksa, A. Bauer, and S. Artavanis-Tsakonas. Analyzing protein complexes in drosophila with tandem affinity purification-mass spectrometry. *Dev Dyn*, 232(3):827–34, 2005.

[144] Alex von Kriegsheim, Daniela Baiocchi, Marc Birtwistle, David Sumpton, Willy Bienvenut, Nicholas Morrice, Kayo Yamada, Angus Lamond, Gabriella Kalna, Richard Orton, et al. Cell fate decisions are specified by the dynamic erk interactome. *Nature cell biology*, 11(12):1458–1464, 2009.

[145] F. Wan, D. E. Anderson, R. A. Barnitz, A. Snow, N. Bidere, L. Zheng, V. Hegde, L. T. Lam, L. M. Staudt, D. Levens, W. A. Deutsch, and M. J. Lenardo. Ribosomal protein s3: a kh domain subunit in nf-kappab complexes that mediates selective gene regulation. *Cell*, 131(5):927–39, 2007.

[146] Xiaodong Wang, John Venable, Paul LaPointe, Darren M Hutt, Atanas V Koulov, Judith Coppinger, Cemal Gurkan, Wendy Kellner, Jeanne Matteson, Helen Plutner, et al. Hsp90 cochaperone aha1 downregulation rescues misfolding of cftr in cystic fibrosis. *Cell*, 127(4):803–815, 2006.

[147] M. West, G. S. Ginsburg, A. T. Huang, and J. R. Nevins. Embracing the complexity of genomic data for personalized medicine. *Genome Research*, 16(5):559–566, May 2006.

[148] Andrea D. Weston and Leroy Hood. Systems biology, proteomics, and the future of health care: Toward predictive, preventative, and personalized medicine. *Journal of Proteome Research*, 3(2):179–196, 2004.

[149] Zhipeng Xie, Chee Keong Kwoh, Xiao-Li Li, and Min Wu. Construction of co-complex score matrix for protein complex prediction from ap-ms data. *Bioinformatics*, 27(13):i159–i166, 2011.

[150] Alice Y Yam, Yu Xia, Hen-Tzu Jill Lin, Alma Burlingame, Mark Gerstein, and Judith Frydman. Defining the tric/cct interactome links chaperonin function to stabilization of newly made proteins with complex topologies. *Nature structural & molecular biology*, 15(12):1255–1262, 2008.

[151] Takuya Yamamoto, Miki Ebisuya, Fumito Ashida, Kazuo Okamoto, Shin Yonehara, and Eisuke Nishida. Continuous erk activation downregulates antiproliferative genes throughout g1 phase to allow cell-cycle progression. *Current Biology*, 16(12):1171–1182, 2006.

[152] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[153] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–106, 2014.

[154] Esti Yeger-Lotem, Laura Riva, Linhui Julie Su, Aaron D Gitler, Anil G Cashikar, Oliver D King, Pavan K Auluck, Melissa L Geddie, Julie S Valastyan, David R Karger, et al. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature Genetics*, 41(3):316–323, 2009.

[155] Sailu Yellaboina, Asba Tasneem, Dmitri V Zaykin, Balaji Raghavachari, and Raja Jothi. Domine: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic acids research*, 39(suppl 1):D730–D735, 2011.

[156] A. S. Yoo, C. Bais, and I. Greenwald. Crosstalk between the egfr and lin-12/notch pathways in c. elegans vulval development. *Science*, 303(5658):663–6, 2004.

[157] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam,

N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–10, 2008.

[158] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, 2006.

[159] Xueping Yu, Joseph Ivanic, Anders Wallqvist, and Jaques Reifman. A novel scoring approach for protein co-purification data reveals high interaction specificity. *PLoS Computational Biology*, 5(9):e1000515, 2009.

[160] Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L Price. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genetics*, 9(5):e1003520, 2013.

[161] Elena Zelin, Yang Zhang, Oyetunji A Toogun, Sheng Zhong, and Brian C Freeman. The p23 molecular chaperone and gcn5 acetylase jointly modulate protein-dna dynamics and open chromatin status. *Molecular Cell*, 48(3):459–470, 2012.

[162] Qi Zeng, Jing-Ming Dong, Ke Guo, Jie Li, Hui-Xian Tan, Vicki Koh, Catherine J Pallen, Edward Manser, and Wanjin Hong. Prl-3 and prl-1 promote cell migration, invasion, and metastasis. *Cancer research*, 63(11):2716–2722, 2003.

[163] Bing Zhang, Byung-Hoon Park, Tatiana Karpinets, and Nagiza F Samatova. From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics*, 24(7):979–986, 2008.

[164] Keyan Zhao, María José Aranzana, Sung Kim, Clare Lister, Chikako Shindo, Chunlao Tang, Christopher Toomajian, Honggang Zheng, Caroline Dean, Paul Marjoram, et al. An arabidopsis example of association mapping in structured samples. *PLoS Genetics*, 3(1):e4, 2007.

[165] Rongmin Zhao, Mike Davey, Ya-Chieh Hsu, Pia Kaplanek, Amy Tong, Ainslie B Parsons, Nevan Krogan, Gerard Cagney, Duy Mai, Jack Greenblatt, et al. Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the hsp90 chaperone. *Cell*, 120(5):715–727, 2005.

[166] Meiying Zheng, Tomasz Cierpicki, Alexander J Burdette, Darkhan Utepbergenov, Paweł Ł Janczyk, Urszula Derewenda, P Todd Stukenberg, Kim A Caldwell, and Zygmunt S Derewenda. Structural features and chaperone activity of the nudc protein family. *Journal of Molecular Biology*, 409(5):722–741, 2011.

[167] Lecong Zhou, Santiago Mideros, Lei Bao, Regina Hanlon, Felipe Arredondo, Sucheta Tripathy, Konstantinos Krampis, Adam Jerauld, Clive Evans, Steven St Martin, MA Saghai Maroof, Ina Hoeschele, Anne Dorrance, and Brett Tyler. Infection and genotype remodel the entire soybean transcriptome. *BMC Genomics*, 10(1):49, 2009.

[168] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, 2012.

[169] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67(2):301–320, 2005.