Computational methods to explore chemical and genetic interaction networks for novel human therapies

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Raamesh Deshpande

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Chad L. Myers, Advisor

November, 2013

Acknowledgements

I thank my advisor Chad Myers for being a great teacher, mentor, and supervisor to me and providing an atmosphere conducive for conducting cutting-edge research. He has been very supportive and given me excellent advice throughout my Ph.D. I admire his patience, his communicate skills, his smile, and his ability to understand complicated concepts very quickly. The quality I like most in him and the one that makes him a great advisor and teacher is that he never refused a question. These qualities made working with him a great pleasure.

Working with my labmates, Ben, Elizabeth, Yungil, Roman, Jeremy, Rob, Justin, Scott, Carles, Seung Ho, Josh, has been very productive and at the same time a lot of fun. We often discussed a host of scientific topics from a big picture - “How to save the world?” - level questions to minutiae of statistics. These discussions were incredibly useful in brainstorming interesting research ideas as well as learning new concepts and getting doubts cleared. Also, my labmates gave useful comments on my practice talks which were helpful in improving my talks for conferences and other venues. I also thank my colleagues from Rui Kuang’s lab: Zhang Wei, Taehyun Hwang, Ze Tian and others; and Vipin Kumar’s lab: Gowtham Atluri, Gaurav Pandey, Wen Wang, Gang Fang, Sanjoy Dey and others. These groups work in related but slightly different fields of research and interacting with them was useful to refine my research and brought new set of ideas to me. It has been a great pleasure and privilege to work with all of you.

I thank my collaborators who were crucial for almost all of my projects. I have worked with Jeff Piotrowski, Sheena Li, Kerry Andrusiak, and Charles Boone from University of Toronto for the chemical genomics project; with Shikha Sharma and Wei-Shou Hu for the active subnetworks discovery project; Michael Asiedu, Mitchell Klebig and Dennis Wigle for the cancer synthetic lethality project; Anastasia Baryshnikova and Michael Costanzo for the genetic interactions project. Working with these collaborators who are experts in their respective fields has broadened my perspective and has been immensely rewarding.

I have been fortunate to have made a lot of friends during my stay at UMN and they have made my life memorable and interesting, and kept me sailing through my Ph.D. We partied, gossipped, visited places, shared food, had a lot of fun, and also had a lot of intellectual discussions. I thank GVM, Rasik, Ramya, Anu, Ankit, Avanish, Navneet, Rakesh, Srini, Taehyun, Wei, Preethi, Anagha, Mikhil, Pratap, Anup, Vinit, Sayan, Nandita, Varoon and Aruna. I also thank my friends from IIT and NCBS: Riju, Piyush, Pragati, Niraj, Abhishek and many others who kept in touch with me despite differences in space and time zone.

I also thank the computer science office and the operator staff who took care of administrative and computer details respectively. I especially thank Georganne Tolaas for helping me through the administrative details of Ph.D. process.

I give special thanks to my parents, Kalindi Deshpande and Shripad Deshpande, and my sister, Juhi Deshpande, for being supportive from the Ph.D. selection process to the completion of my Ph.D. I also thank my relatives Satish Patil and Vimala Sortur who encouraged me to pursue my Ph.D. dream, my grandmother, Akka, who kept me connected to the rest of my big family, and also other relatives who always wished big things for me.

Dedication

This dissertation is dedicated to my parents, Kalindi Deshpande and Shripad Deshpande, and my sister, Juhi Deshpande.

Abstract

Model organisms are often used as a test-bed for the development of new genomic technologies and computational approaches. For example, the yeast *Saccharomyces cerevisiae* was the first eukaryote to have its entire genome sequenced, paving the way for the sequencing of the human genome. Beyond genome sequencing, yeast and other model organisms have been extensively used for reverse genetics technology development. Reverse genetics is a general approach for studying biology where the genome is perturbed in precise ways (e.g. targeted gene deletion), to gain functional information about the perturbed genes from the resulting phenotypic changes. With developments of new genomic technologies, reverse genetics at a genome-wide scale has become a reality. This dissertation focuses on developing several computational methods for scaling up the reverse genetics experiments in model organisms as well as for exploring the generated genomics data with the ultimate goal of understanding and translating these data for use in applications for human therapeutics.

Towards this ultimate goal, along with my collaborators, we have screened comprehensive chemical genomics interactions data experiments for a large collection of natural products in model organism *S. cerevisiae*. Recent studies have proven the power of yeast chemical genomics approaches for mechanism of action (MOA) analysis. In particular, for compounds with a specific MOA, the chemical-genetic interaction profile, which is determined by testing yeast non-essential gene deletion mutants for compound hypersensitivity, should mimic the target gene's genetic interaction profile, as determined through global synthetic lethal genetic interaction screening. One limitation of the approach for screening compounds is that natural products are of limited availability and thus there simply may not be enough compound for full-genome chemical genetic profiling. To address this limitation, we developed a method, COMPRESS-GI (COMpress Profiles Related to Epistasis by Selecting Informative Genes), to identify a small subset of the deletion collection that remains highly informative for MOA analysis. We identified a diagnostic strain set comprising ~5% of the non-essential deletion mutant collection. We validated this approach, showing that our diagnostic set performs comparably to complete chemical-genetic profiles. We also demonstrate that our method provides substantial improvement over baseline strategies based on selection of either random genes or hubs. Our approach can also be generalized beyond the chemical genomics context. For example, the approach can be used for optimizing genetic interaction screens in new organisms as well as genetic interaction screens across multiple conditions.

One challenge we encountered while developing COMPRESS-GI algorithm was that Pearson correlation, a popular similarity measure in the genetic interaction community, performed close to background performance for thresholded genetic interaction data and for smaller profiles. The stability of the similarity measure especially on smaller profiles is extremely critical for COMPRESS-GI algorithm. We realized that no systematic study has been conducted to evaluate the strengths and weaknesses of different similarity measures for genetic interaction analysis. So, I conducted a comparison of different profile similarity measures for their abilities to discover functionally similar genes from the genetic interaction networks in various contexts such

as noise conditions, batch effects, thresholded genetic interactions and smaller profiles. I discovered that dot product, one of the simplest profile similarity measures, outperforms several sophisticated similarity measures for many of these contexts. Further, dot product satisfied the stability criterion essential for the COMPRESS-GI algorithm so I used dot product in the COMPRESS-GI algorithm.

Based on the COMPRESS-GI strategy, a large scale chemical genomics screening of more than 10,000 natural compounds was conducted in *S. cerevisiae*. In addition to being a stepping stone for chemical genomics experiments in human, the major aim of this experiment is to generate a knowledge base of drug-targets that could be translated to human. One way to translate the findings is to focus on drug-targets where the target gene is conserved in human. The conserved genes comprise over 50% of the yeast genome which means that majority of the yeast predictions can be used for discovery of therapeutically relevant compounds in human. While precise targets are discovered in few cases, for a large majority of cases, the predictions are at a much higher process level predictions. We need methods that can translate such process level predictions to human in an unbiased and comprehensive fashion. Biological networks such as protein interaction networks and functional linkage networks contain the information about gene relations which could be used for such unbiased translations. To deal with this problem as well as a more general problem of cross-species comparison of genomic data using networks, we developed neXus (Network - cross(X)-species - Search), a method for network-based exploration of genomic data across species. neXus is a first network based exploration method developed for the cross-species setting as well as it is a first method that can scale up to size of functional linkage networks which can contain up to several millions edges. Though developed for cross-species setting, we demonstrate the single species version of neXus performs better than other available single-species methods in terms of discovery of more number of subnetworks in the data relative to when the input is randomized.

One of the direct applications of discovery of targets for the drugs is that it will pave way to treat cancer with known susceptible targets. The mutations in cancer cells though provide growth advantage to cancer cells also create several susceptibilities which could be targeted to selectively kill the cancer cells and leaving the surrounding normal cell relatively unharmed. Synthetic lethality, which describes lethality of a double gene knock-out strain when either of the respective single gene knock-out strains are viable, provides a mean to discover targets specific to a cancer tumor. Targeting the synthetic lethal interactors of genes mutated in cancer will cause lethality to only cancer cells; however, only few synthetic lethal interactions are known in human. On the other hand, synthetic lethal interactions have been screened on a very large scale in model organisms which could be translated to human for discovery of cancer targets. We have tested the cross-species approach to predict cancer relevant synthetic lethal interactions in human at a proof of principle scale and discovered 5 out of 21 are conserved in human. Among them, we discovered a strong interaction between *SMARCB1*, a gene frequently mutated in rhabdoid sarcomas and *PSMA4*. We demonstrate that knock-down of *PSMA4* is lethal in tumors with endogenous *SMARCB1* mutation while a tumor with functional *SMARCB1* is relatively unaffected.

My contributions impact several different aspects of personalized medicine strategy including drug discovery and cancer-target discovery. Drug discovery has become increasingly expensive in the recent years and developing newer ways for drug discovery has become essential. We aim to break the trend of increasing costs by developing a novel drug-discovery approach inspired by system biology, first, in model organisms, and then later in human cell lines. I worked on functional genomic optimization strategies to increase the throughput of the drug discovery approach in model organisms. One the other hand I have also worked on cancer target discovery where the knowledge of drug-targets can be directly applied. I, along with my collaborators, have demonstrated the approach of using synthetic lethality interactions from model organisms to discover cancer targets in human is feasible. I foresee that in future personalized medicine strategy for cancer treatment is fully developed where given a cancer tumor, the cancer genome is sequenced, mutations identified and used to predict tumor-specific cancer targets using which appropriate drugs can be used that inhibit the discovered cancer targets.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

We are in an exciting era of computational biology where proper understanding of the generated data has become the bottleneck rather than the data generation. Several types of new genomics data are being generated, each of which uniquely contributes to the understanding of biological systems. For example, quantitative expression measures capture genes that are actively transcribed in an organism for a particular condition, protein interaction networks have helped understand how proteins bind to form larger structures to perform their function, while genetic interaction networks have revealed how genes are dependent on each other for different phenotypes. Understanding these datasets is certainly a challenge, but in several cases, the rate of data generation has exceeded the growth of computers, and this has created scalability related issues. A historically successfully example of keeping up with the data generation is BLAST[1], which evolved from a global alignment algorithm to a local alignment algorithm that can scale up with the data generated and at the same time provide reliable results. Many other fields of computational biology now face a similar challenge and a better understanding of the operations used in our analyses and development of scalable methods than enable their application on the fast-growing collections of data has become a necessity.

In addition to sheer volume of data, computational and systems biology face the daunting challenge of integrating diverse types of genomic-scale data to generate a holistic understanding of the biological system. Data integration in computational biology can be broadly classified into two types: integrating different datasets to generate new perspectives, and comparison of analogous datasets across different contexts. An example of integrating different datasets is integration of expression data for a certain condition with protein interaction network to discover functional modules active in that context. In contrast, a parallel data integration scenario is comparing analogous genomic datasets across different contexts such as different cancer cell lines or different species to understand what is common and unique about the datasets and contexts. For example, contrasting the expression profiled in stem cells and differentiated cells will help us identify what is unique in each of the two cell types. Another example of analogous data integration is comparing biological networks across species, which will help in discovering conserved functional modules across species.

I have addressed several of these problems in this dissertation; however before delving into these specific challenges, I will first introduce some of the genomics datasets and resources which are used in the subsequent chapters. After introduction of the resources, I will describe the organization of the dissertation.

1.1 Relevant genomics resources

1.1.1 Genetic interaction data

. Reverse genetics refers to experiments that involve making a specific perturbation to a genetic element in an organism and associating the changes in the outcome in the organism with the perturbed genetic element. Genes are the most natural genetic element for such reverse genetic experiments. So in many model organisms, collections of strains that lack specific genes have been created [2-4] and studied with respect to several phenotypes [5,6], enabling the discovery of gene functions. Figure 1-1 shows the distribution of the single knock-out strains’ fitnesses for the 6000 genes in

Saccharomyces cerevisiae. The single mutant fitnesses of the non-essential genes is skewed towards a fitness of 1, that is the individual knock-out of a large majority of genes does not affect cell's growth. This observation does not mean that a large majority of non-essential genes are not required by the cell; rather the cell has evolved a complex system of redundancy at several levels to compensate the losses in the genome. In fact, depending on the condition, different sets of genes are essential and 97% of the genes have been found essential in one condition or the other [5]. To better understand the role and importance of these non-essential genes, one approach is to study the combination knock-outs, popularly known as genetic interactions.

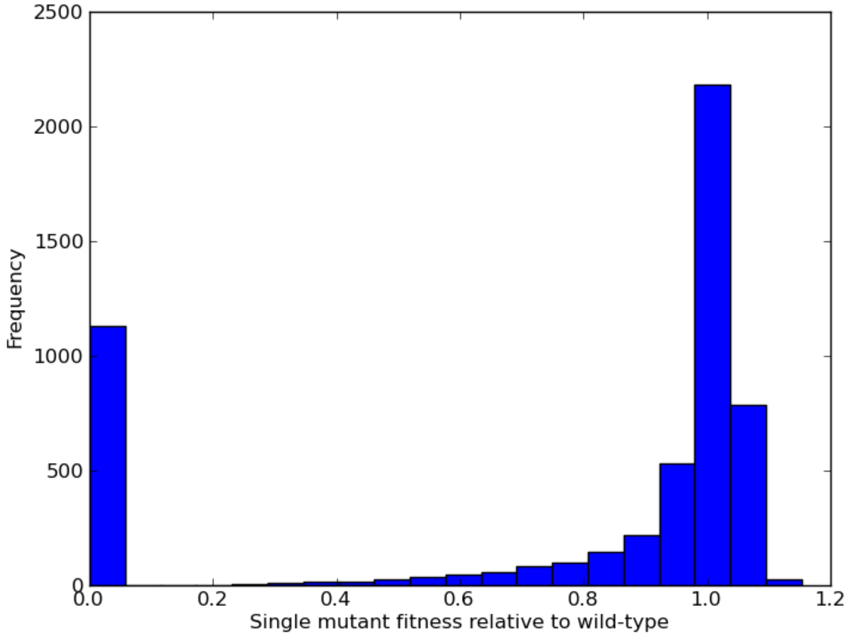


Figure 1.1 Histogram of *S. cerevisiae* single mutant fitnesses

A genetic interaction is said to exist between two genes if a double gene knock-out strain's phenotype is different what is expected given the two single gene knock-out phenotypes [7]. For example, an extreme case of genetic interaction is synthetic lethality when knocking out two non-essential genes causes cell death. Typically, fitness is a popular choice of phenotype for measuring genetic interaction. Genetic interactions are quantitatively measured by screening and finding the fitness of the two single knock-out strains (f_A, f_B) and the double knock-out strain (f_{AB}). Assuming a multiplicative model for genetic interactions, the extrapolated double knock-out fitness is f_Af_B. The difference between the observed and the extrapolated fitnesses (f_{AB} - f_Af_B), is a measure of genetic interaction.

Depending on the sign of the genetic interaction score, growth-based-genetic interactions are categorized as positive or negative interactions. A negative genetic interaction between two genes means that the growth of the double knock-out of the two genes is less than what is expected from single gene knock-out growths. On the other hand if the double gene knock-out grows faster than what is expected then the genetic interaction is referred to as positive interaction. Both signs of genetic interactions have been extensively studied for its utility in gene function prediction, overlap with protein interaction network, and the organization of the interactions with respect to biological processes. Based on the earlier observations, a model was proposed that mainly negative genetic interactions span between the processes and positive interactions occur mainly between the processes [8]. However, for genetic interactions between non-essential genes, both positive and negative genetic interactions are more likely to occur between genes sharing the same function and between genes that have protein interactions between themselves compared to random gene-pairs. Further, both negative and positive genetic interactions have been observed to span across the functional processes [9]. These observations suggest that the model for the genetic interactions is not as simple as the one proposed in [8]; nevertheless, both types of interactions are useful in understanding the gene function.

Genetic interactions have now been screened on a high-throughput scale in several organisms including *S. cerevisiae* [10], *S. pombe* [11], *E. coli* [12], mouse cell lines [13] and human cell lines [14]. In this dissertation, we have used genetic interaction datasets for yeasts, *S. cerevisiae* and *S. pombe*. In these yeasts, a knock-out strain (query) is mated with a library of knock-out strains (arrays). Each of these crosses which creates 4 possible haplotypes, out of which one is a double mutant. The desired double knock-out strain can be selected using the design that the knock-out strains have a marker gene in place of the missing gene; therefore, the double knock-out strain can be selected for by selecting for the two markers. However, the genetic interaction experiment design in higher organisms is slightly different from lower organisms. In higher model organisms, constructing knock-out cell lines is difficult; however, shRNAs could be used to inhibit the expression of a particular gene. The genetic interactions screening in higher organisms can be conducted by systematically using pairs of shRNAs to construct double gene knock-downs. The other strategy has been to use a query shRNA on a pool containing a library of single gene knock-down cells [15].

Genetic interactions by themselves have proven useful in discovering similar genes. However, genetic interaction profile similarity, which is a continuous form of measuring similarity of the neighbors in the genetic interaction screen, has proven to be more reliable for gene function prediction [10]. The reason for increased reliability is that profile similarity is that it takes several interactions into consideration to compute the similarity. Because of this reason, genetic interaction profile similarity has been one of the most popular ways to get similar function genes from genetic interactions.

1.1.2 **Protein-protein interactions**

Proteins are the main worker units of the cell. For example, all enzymes, biological catalysts are proteins; the structure in the cell have proteins as one of the main constituents (example: microtubules); mechanical action is made possible by proteins (example: muscle contract through actin and myosin); and so on. Often proteins organize by binding themselves or other proteins to form larger structures for performing sophisticated functions. The binding relations between the proteins are referred to as protein-protein interactions. High-throughput protein interaction screens have generated protein interaction networks on a genome scale and have been instrumental in revealing the molecular organization of the cell. In fact, many of the protein complex definitions and annotations are based on cliques in the protein interaction network [16], and functions of many uncharacterized genes have been discovered using protein interaction networks [16].

There are several methods to screen protein interactions; however, tandem affinity purification-mass spectrometry (TAP-MS) [17] and yeast 2 hybrid (Y2H) [18-20] are the two most prominent and widely used protein interaction screening methods in the field. Briefly, the TAP-MS method involves attaching a protein to a bead and letting other proteins bind to it, which can then be isolated and proteins in it can be identified using

MS. The binding of a protein to a bead is done by coating the bead with Immunoglobulin G (IgG), an antibody, and constructing a library of strains each of which contains the desired query protein fused with a TAP tag that strongly binds to IgG. To discover the interactions, the cells containing a TAP tagged fusion protein are lysed and the contents are allowed to bind with the IgG coated bead. The extract undergoes multiple stages of purification to remove contaminants and release the complexes containing the query protein from the IgG coated bead. The purified protein complexes are then separated using gel electrophoresis and proteins in each of the complex are identified using MS. TAP-MS has been instrumental in large-scale discovery of stable protein complexes [21]; however, one disadvantage of the TAP-MS method is that the discovered interactions may not be specific. For example, several proteins may be found bound to the query protein but they may not be directly interacting with the query protein. Some of the proteins may be indirectly interacting by binding to proteins that are bound to the bait but in practice all bound proteins are considered to be interacting with each other. In contrast, Y2H provides an alternative where interactions between a specific pairs of proteins can be tested for protein interactions. Y2H is an *in vitro* method in the sense that an interaction between two proteins from any organism is expressed and tested in yeast. The DNA corresponding to the pair of proteins, called bait and prey proteins in Y2H terminology, are introduced into the yeast genome along with fragments of a transcription factor for the bait and the prey respectively. The transcription factor is functional only when the bait and the prey combine, and therefore the existence of the transcription factor can be used to detect the protein interaction. However, the main disadvantage of the Y2H method is that the proteins may not be actually interacting in the organism they belong to. Another protein interaction screening method is protein-fragment complementation assay (PCA) [22], which is especially useful for discovering interactions between membrane proteins, which are hard to find using other methods.

Protein interaction networks have inspired approaches to address several fundamental research questions. For example, protein interaction networks have been a testing ground for several models proposed for genome evolution such as the preferential attachment and duplication-mutation-complementation (DMC) models. Preferential attachment means that new proteins added to the genome would more likely interact with hub proteins in protein interaction network while DMC model suggests that the proteins are duplicated from existing proteins in the system such that both the duplicates share all the protein interactions (duplication) but over time, some of the interactions can be randomly lost (mutation) while some interactions can be split between the duplicated genes (complementation). The comparison of properties of the current protein interaction networks with protein interaction network using the growth models suggested that DMC model is more plausible [23]. The protein interaction networks have been also instrumental in revealing interesting properties of disease genes such as many disease genes exist in the periphery of the protein interaction network. This observation suggests that the disease genes are usually not the most critical genes in the genome. Also protein interaction networks have inspired the field of data integration by acting as a scaffold with which many genomics data including microarray expression data can be combined. For example, protein interaction networks were used in the field of biomarker discovery when biomarkers discovery projects across independent cohorts did not yield very different sets of biomarkers but biomarker discovery using protein interaction networks yielded a more consistent set of subnetwork biomarkers across the cohorts [24]. Due to its immense functional utility in many aspects of genomics, protein interaction network have been screened in several organisms including *S. cerevisiae* [16,18-20,22], *E. coli* [25] and *H. sapiens* [26].

1.1.3 **Functional linkage networks**

Functional linkage networks are generated by integrating several diverse kinds of genomic datasets. Often we have multiple genomics datasets, and we would like a single dataset for analysis. For example, consider a situation where we have three different experimental datasets available, e.g. Y2H, TAP-MS and PCA protein interaction networks, and we want to combine and generate a single interaction network. One approach to combine the datasets is to only consider the protein interaction edges that are present in all three networks. This approach will create a highly confident but small network. In statistical terminology, this approach creates a network with high precision (rate of being correct on the covered interactions) but low recall (coverage). Alternatively, the networks can be combined using a union approach, where we will consider any edge that occurred in any of the three networks. Such an approach will create a network that is larger but may not be very accurate as many of the edges may exist only in one network and therefore may not be reliable, that is, a low precision and high recall network. Yet another approach is to create a weighted network based on the frequency of occurrence of edges in the three networks. This seems to take both recall and precision into consideration. This network is as large on the one created by the union approach but at the same time, if we sort the edges by their weight, the most confident edges will be at the top. However, we have given equal weight to all three datasets, but we may have more confidence for some datasets compared to others. We can evaluate the utility of the networks by an independent high confident standard and use the evaluation score as a way to weigh the three networks. Further, the three datasets may not contain a binary edge but may have a confidence score associated with each edge, which can also be considered when calculated the weight for the edge. This is the concept behind the Bayesian integration, which is used to integrate heterogeneous genomics data to create a functional linkage network [27].

Like protein interaction networks, the functional linkage networks are instrumental in gene function discovery [27]. In addition, functional networks are more comprehensive than the protein interaction networks. While genome-scale protein interaction networks are on the order of tens of thousands of interactions, functional linkage networks can span millions of edges, depending on the confidence threshold one requires on the edges in the network. Another advantage of functional linkage networks over protein interaction networks is that functional linkage networks can be constructed even with limited data such as gene expression data. This is especially advantageous for organisms with limited protein interactions data. Functional networks have now been generated in several organisms including baker’s yeast [27], fission yeast [28], bacteria [29], fly [30], mouse [31], and human [32].

1.1.4 Chemical genomics

Chemical genomics is the study of interactions of chemical compounds with genes on the genome scale. There are several types of interactions between chemical compounds and genes; however, in this dissertation, we will be discussing only two kinds: chemical genetic interactions and compound-target interactions.

Chemical genetic interactions are analogous to genetic interactions and refer to the unexpected phenotype of a mutant strain when exposed to a chemical compound. Examples of chemical genetic interactions are sensitivity or resistance of a knock-out strain to a compound relative to the wild-type strain. A few large-scale chemical genetic interaction screens have been conducted; however, to date, major studies have mainly been conducted in the model organism *S. cerevisiae* [5,33]. The chemical genomics experiments in model organisms are typically conducted by growing the knock-out strains with the compound, and the growth of the knock-out strains are measured for the control and the compound conditions. The knock-out strains which grew slower in the compound condition compared to the control condition are sensitive to the compound while the strains that grew faster in the compound condition are resistant to the compound.

In contrast to fitness measurements in the context of most genetic interaction assays which rely on colony size, growth in high-throughput chemical genomics is typically measured using other technology such as microarray or sequencing if a barcoded knock-out collection is available. [5,33]. The reason is that some compounds may not uniformly dissolve in the solid media and the strains need to be grown in liquid media containing the compound. In order to simplify the experiment, all knock-out strains are pooled and grown together in the compound and the control conditions respectively. The abundance of the different strains is measured by counting the unique barcodes in the knock-out strains which are present in the place of knocked-out gene. A unique barcode sequence was originally inserted in the genome of each mutant strain so that Polymerase chain reaction (PCR) experiment, which replicates a DNA sequence with a particular primer DNA sequence, could be conducted selectively on the knock-out strain. However, we can adapt this tool for other technologies such as microarray and the sequencing that allows for large-scale counting of barcodes to determine the abundance of the knock-out strains by counting the number of barcodes in the chemical genomics experiment.

The second type of interaction we discuss in this dissertation is the compound-target interaction which can be thought of as analogous to protein interactions. The interaction refers to a ligand (compound) binding to a molecule, usually a protein, but the target can be a metabolite produced by the protein. One of the major quests in chemical genomics is to discover compound-target interactions and this has been one of the most difficult research challenges in this field. Understanding how a compound acts on live cells, in particular what molecules the compound binds in the living cell and how the binding affects the cell is often referred to as "mechanism of action" of the compound. One way to discover interesting compounds is to search for compounds that are similar to already known drugs in terms of chemical structure or biological activity. However, a limitation of this approach is that prior knowledge for similar compounds is required and thus, completely novel compounds cannot be discovered. Often a target centric approach is used where given a target protein of therapeutic value, compounds are screened for inhibition of the target protein. The search can be conducted using chemical informatics means where the binding of the compound-ligand is computationally inferred. Alternatively, the search can be experimental in nature, where an appropriate assay to check the inhibition of the target is designed and large libraries of compounds are then screened with the assay.

In contrast to the target-centric approach, drug targets can be predicted in a compound-centric way using chemical genetic interactions. The hypothesis is that if a compound targets a gene, the compound effect should mimic the target gene’s knock-out effect (genetic interaction profile) [34]. The compound’s effect on the library of non-essential knock-out strains can be screened using chemical genomics while gene’s knock-out effect on the library of knock-out strains can be screened using genetic interactions. The target of a compound can be predicted by correlating the chemical genetic profile with the genetic interaction and discovering the gene whose genetic interaction profile is closest to the compound’s chemical genetic interaction profile (see Chapter 5 for details).

1.1.5 Methods for Gene Orthology

Gene orthology refers to a map between similar genes across species, which have evolved from the same gene in the common ancestor. The most popular way to discover similar genes across species is by detecting similarity of the protein sequences. One of the early orthology prediction methods was the reciprocal best hit approach [35], in which a candidate gene in one species is searched in the other species’ genome using a fast sequence alignment algorithm such as BLAST [1]. The most similar gene in the other organism is searched in the first organism, and if the original candidate gene is recovered as the most similar gene then the genes are considered orthologous. One drawback of the reciprocal best hit strategy is that the genes may have duplicated in the two species after a speciation event which may cause a cluster of proteins in one organism to be orthologous to a cluster of proteins in the other organism.

Several methods have been developed to incorporate the fact that duplicates of the ancestral gene after the speciation event are orthologous [36]. These methods can be broadly split into two classes: methods that use pair-wise gene-relationship mostly based on protein sequence similarities [37] and methods that construct evolutionary trees for the genes [38]. I have used Inparanoid (version 5 [39] and version 7 [37]) for my work in this dissertation, which is a method that uses pair-wise gene relationships to discover orthologs and it claims that the pair-wise gene relationship methods such as itself are better than tree reconstruction method in general and especially on a global scale [37]. The method builds on the orthologs discovered from the reciprocal best hit strategy by conducting pairwise gene comparison between and across the species and discovering paralogs within the species (inparalogs). The inparalogs are discovered by comparing the within species gene similarities with similarities of the orthologs. For example, if B2, C2 are orthologs from reciprocal best hit strategy across species B and C, the inparalog score for candidate inparalog C3 is : Inparalog score of C3 = Blast[C2:C3]−Blast[C2:B2])/(Blast[C2:C2]−Blast[C2:B2]) , where Blast[P:Q] is average blast score between P and Q in bits (formula taken from Inparanoid 5).

1.1.6 **Microarray expression studies**

Microarray expression experiments measure mRNA concentrations in the living cells for a particular condition, which is reflective of the activity of proteins in the cells. Measuring activities of proteins, the main workers of the cell, is important for understanding how the condition affects the cell; however, measuring protein concentrations is difficult and expensive. Relatively, the concentration of the precursor of the proteins according to the central dogma of genetics, mRNA, can be much more easily measured. The main idea behind measuring mRNA concentration in microarray studies is that complementary RNA strands will bind each other. Based on this idea, complementary strands (probes) for different mRNA are each fixed to different places on a plate and exposed to a sample extract so that mRNAs in the extract hybridize with the respective probes. After the plate is washed to remove the non-hybridized mRNA, the amount of mRNA attached to each probe can be quantified. [40]. In the recent years, with the dramatic reduction of sequencing costs, RNA-seq experiments have become a popular alternative to microarray expression experiments. RNA-seq experiments sequence the entire transcriptome and measure the abundance of the mRNA transcripts by aligning it to the known genes and counting them.

1.2 Dissertation focus

With technological breakthroughs in genomic data generation, computational biology is becoming increasingly important to the understanding of biology. It has in fact become a necessity to develop newer ways of analyses and newer methods for understanding new genomic datasets produced, develop new algorithms to keep up with amount of genomic data generated, and integrate the diverse type of datasets to generate a comprehensive picture of the underlying biological phenomena. The end goal of computational biology is to holistically understand the biological system from the observed data and to apply these insights to various applications such as disease diagnosis the development of new therapies.

In my Ph.D., I have developed approaches and algorithms that make headway in some of the major computational biology challenges. I briefly outline the challenges here, and each chapter is described in further detail in Chapters 2-5.

Integrating enormous amounts of diverse types of data is a critical challenge in computational biology. Active subnetwork discovery is a popular data integration challenge where context-specific gene scores (example: gene expression) are interpreted using a static gene network (example: PPI). Such integration can reveal important and significant functional modules in the network which cannot be discovered by either of the datasets alone. The subnetworks are discovered by searching for high-scoring clusters that are densely connected in the network. The optimal solution to this problem is intractable, so heuristics must be used to perform this integration. Several heuristics have been developed, the first of which was jActiveModules developed by Ideker *et al* [41] and later other methods including MATISSE [42], CEZANNE [43], and heinz [44], were developed. However, there are some major limitations to these methods. Firstly, these methods do not scale to large networks such as functional linkage networks, and secondly, all of the previous methods are single species methods. However, the reality today is that analogous genomic datasets such as gene expression data are being generated in several organisms and at the same time, networks for several organisms are now available, including many at a much higher density. To address this reality, I have developed a method, neXus, that is scalable to existing functional linkage networks and can be extended for cross-species analysis. Another limitation not appreciated by several earlier methods is that subnetworks can be discovered even with random sets of genes, so simply the discovery of a subnetwork does not mean it is significant. The method I have designed takes this fact into account and assesses significance of the subnetworks based on the discovery of similarly sized subnetworks by a random strategy. I applied neXus to the problem of comparative stem cell biology, which is covered in detail in Chapter 2.

The war on cancer has been waged since 1971, and while there have been major successes, the war is still not yet won. One of the reasons for cancer being so impervious is that it is not a uniform disease - cancer is a combinatorial genome problem and the combination of gene mutations leading to cancer across different tissue types is very diverse. This combinatorial problem requires a combinatorial solution. It is known that while mutations grant tumor cells the ability to proliferate out of control, the tumor cells gain susceptibilities that normal cells do not have. Synthetic lethal genetic interactions involving genes that are mutated in tumor cells can be used to discover cancer targets, which when targeted, selectively kill the tumor cells. This idea has been recently utilized for developing a novel treatment for breast cancer and other cancers that harbor BRAC1/2 mutation. BRAC1/2 and PARP are synthetic lethal so targeting PARP can kill these cancer cells. Based on this idea, olaparib, a PARP inhibitor was developed and is currently in phase 2 clinical trials. Like the BRAC1/2-PARP interaction, we expect that there are several additional synthetic lethal interactions involving cancer-associated mutations, each one of them with the potential to lead to a new cancer therapeutic. Ideally, we would like to search for such synthetic lethal interactions in human but genetic interaction data is scarce in human. However, a wealth of genetic interaction data exists in model organisms, especially in baker's yeast, which can be utilized to predict synthetic lethal interactions in human [10]. In chapter 3, I describe an approach we have developed to translate the wealth of yeast genetic interaction data for discovery of cancer targets in human.

Identifying similar genes using genetic interaction profile similarity measures has been a widely used approach for various analyses including validating the genetic interaction data and discovering functions of uncharacterized genes from these data. However, the optimal metric for profile similarity has not yet been studied. In chapter 4, I compare several profile similarity measures for their utility in discovering similar genes according to established knowledge of gene function [45], robustness to noise, robustness to batch effects, stability when only a partial genome is screened. Historically, Pearson correlation has been a popularly used correlation measure, but we found that the dot product is preferable in many situations especially when the data is thresholded to remove noise and when smaller profiles are considered. The dot product similarity measure is a critical component of the gene selection method discussed in chapter 5 and we found that many other similarity measures were not suitable for this setting because they are unstable on smaller profiles.

High-throughput screening technologies have dramatically improved over time, but there remain several research problems where complete screens are not feasible as of now. For example, screening all triple order genetic interactions in yeast would cost around 50 billion dollars, which is almost double the entire NIH annual budget for the year 2013. Likewise, there are other problems where adding a dimension makes complete screening infeasible. The screenable space has to be optimized to make the best use of available resources, both in model organisms and more importantly, as the technologies are established, in higher eukaryotes. I have developed a method to optimize genetic interactions and chemical genetic screening for various scenarios. For example, if the objective is to generate a profile similarity network, we can select and then screen around 10% of the genome such that profile similarity network based on the partial screening is as informative as the complete screen. I have also developed approaches for other scenarios such as covering as many interactions as possible, predicting hubs, and discovering local structures in the data. I discuss these screening strategies in Chapter 5.

Finally, I conclude my dissertation in Chapter 6 and discuss future work.

Chapter 2: A Scalable Approach for Discovering Conserved Active Subnetworks across Species

2.1 Overview

Overlaying differential changes in gene expression on protein interaction networks has proven to be a useful approach to interpreting the cell's dynamic response to a changing environment. Despite successes in finding active subnetworks in the context of a single species, the idea of overlaying lists of differentially expressed genes on networks has not yet been extended to support the analysis of multiple species' interaction networks. To address this problem, we designed a scalable, cross-species network search algorithm, neXus (Network - cross(X)-species - Search), that discovers conserved, active subnetworks based on parallel differential expression studies in multiple species. Our approach leverages functional linkage networks, which provide more comprehensive coverage of functional relationships than physical interaction networks by combining heterogeneous types of genomic data. We applied our cross-species approach to identify conserved modules that are differentially active in stem cells relative to differentiated cells based on parallel gene expression studies and functional linkage networks from mouse and human. We find hundreds of conserved active subnetworks enriched for stem cell-associated functions such as cell cycle, DNA repair, and chromatin modification processes. Using a variation of this approach, we also find a number of species-specific networks, which likely reflect mechanisms of stem cell function that have diverged between mouse and human. We assess the statistical significance of the subnetworks by comparing them with subnetworks discovered on random permutations of the differential expression data. We also describe several case examples that illustrate the utility of comparative analysis of active subnetworks.

The work in this chapter is published in [46] and includes contributions from Shikha Sharma, Catherine M. Verfaillie, Wei-Shou Hu and Chad Myers. Shikha collected the microarray expression data and helped in interpretation of the subnetworks. Catherine also helped in interpreting the subnetworks. Wei-Shou and Chad supervised the project.

2.2 Background

Developments in genomic and proteomic technologies in recent years have given us numerous methods for capturing high resolution snapshots of cellular processes. The end result of a genome-scale experiment is typically a long list of candidate genes that provide a basis for further, more detailed, follow up experiments. For example, gene expression microarrays are a popular approach for identifying differentially expressed genes between two cell types or experimental conditions, and this technology typically yields several hundred to a few thousand differentially expressed genes in a typical comparison [47,48]. While there are sometimes obvious biological processes represented within these lists, developing precise hypotheses from such a long list of candidates can be challenging. Although to varying degrees, this is also true of other genome-scale experiments or screens (e.g. genome wide association studies [49] or genetic interaction screens [10]). In short, the bottleneck in genomic research has quickly moved from the production of high-quality data to interpretation and hypothesis generation.

One powerful approach that has been used to aid in the interpretation of candidate genes lists is integrative analysis with complementary genome-scale data. For example, in a landmark study, Ideker *et al.* addressed the challenge of interpreting lists of significantly differentially expressed genes by overlaying them on a protein-protein interaction network [41]. They found that certain groups of differentially expressed genes tend to cluster together on the interaction network, building confidence that the signature was indeed biologically relevant and suggesting that entire physical modules were differentially expressed together. This approach has since been extended to several other scenarios, all demonstrating the utility of this idea. For example, Rajagopalan *et al.* extended Ideker's method to larger, literature-curated biological networks [50]. Others incorporated co-expression scores to favor selected edges of the protein interaction network [43,44,51,52]. Dittrich *et al.* later formulated the problem as an integer linear programming optimization problem [44]. Recent work has also extended this idea to show that sample classification based on expression profiles can also take advantage of complementary structural information in protein-protein interaction networks [24].

In separate studies, groups have compared and aligned the structure of protein-protein interaction networks across species [53,54]. The basic approach adopted by these methods is to identify subgraphs with conservation at the protein sequence level (nodes) as well as at the physical or functional interaction level (edges). This approach has been used to suggest core pathways that are conserved across species and to build confidence in individual protein-protein interactions based on the co-occurrence in multiple species [53,54]. However, to our knowledge, no one has yet applied this idea to study network-based patterns of expression across species. We propose that just as protein-protein interaction networks can be mined for conserved patterns, differential expression patterns overlaid on biological networks can be aligned to identify conserved patterns of expression, which we call *conserved active subnetworks*.

In this study, we describe a novel approach for identifying conserved active subnetworks in interaction networks across multiple species. Given differential expression measures representing analogous phenotypes in two different species and corresponding interaction networks (for example, protein-protein interaction networks), our approach identifies tightly connected network modules that show a high degree of differential expression, i.e. dense subnetworks, and are conserved in both networks. This is in contrast to previous approaches, which focused on using differential expression or other activity scores to identify dense subnetworks in protein-protein interaction networks for a single species [24,41,43,44,50-52].

In addition to addressing the new question of conservation of network patterns across species, our approach presents a scalable solution to active subnetwork identification, which has typically been restricted to relatively sparse protein-protein interaction networks. Sparse coverage of current protein-protein interaction studies limits the ability to match patterns across species. Recent work in area of genomic data integration helps to address this issue. Several approaches now exist which integrate interaction and other information to infer functional associations between genes, to form functional linkage networks [27,31,32]. Such approaches can incorporate protein-protein and genetic interactions, gene expression, protein localization, phenotype, and sequence data; and have been applied now in many species including yeast, bacteria, worm, fly, plants (Arabidopsis), mouse, and human[27,30-32,55-58]. These networks are often significantly denser than protein-protein interaction networks and include hundreds of thousands or even millions of weighted edges that reflect confidence in gene-gene functional relationships. The power (and challenge) in using functional linkage networks is that they capture a broad range of functional relationships that have relevance for defining network modules: for example, physical interactions between proteins, co-expression, regulatory relationships, or shared mutant phenotypes. This is in contrast to protein-protein interaction networks which focus on physical interactions between proteins, our knowledge of which is relatively limited in many species, particularly higher eukaryotes. A more detailed comparison of functional linkage and protein-protein interaction networks and the implications for their use for active subnetwork discovery is provided as Supplementary Material (see a detailed discussion in Appendix 1, Note 1, "Implications of using functional linkage vs. physical interaction networks for active subnetwork discovery").

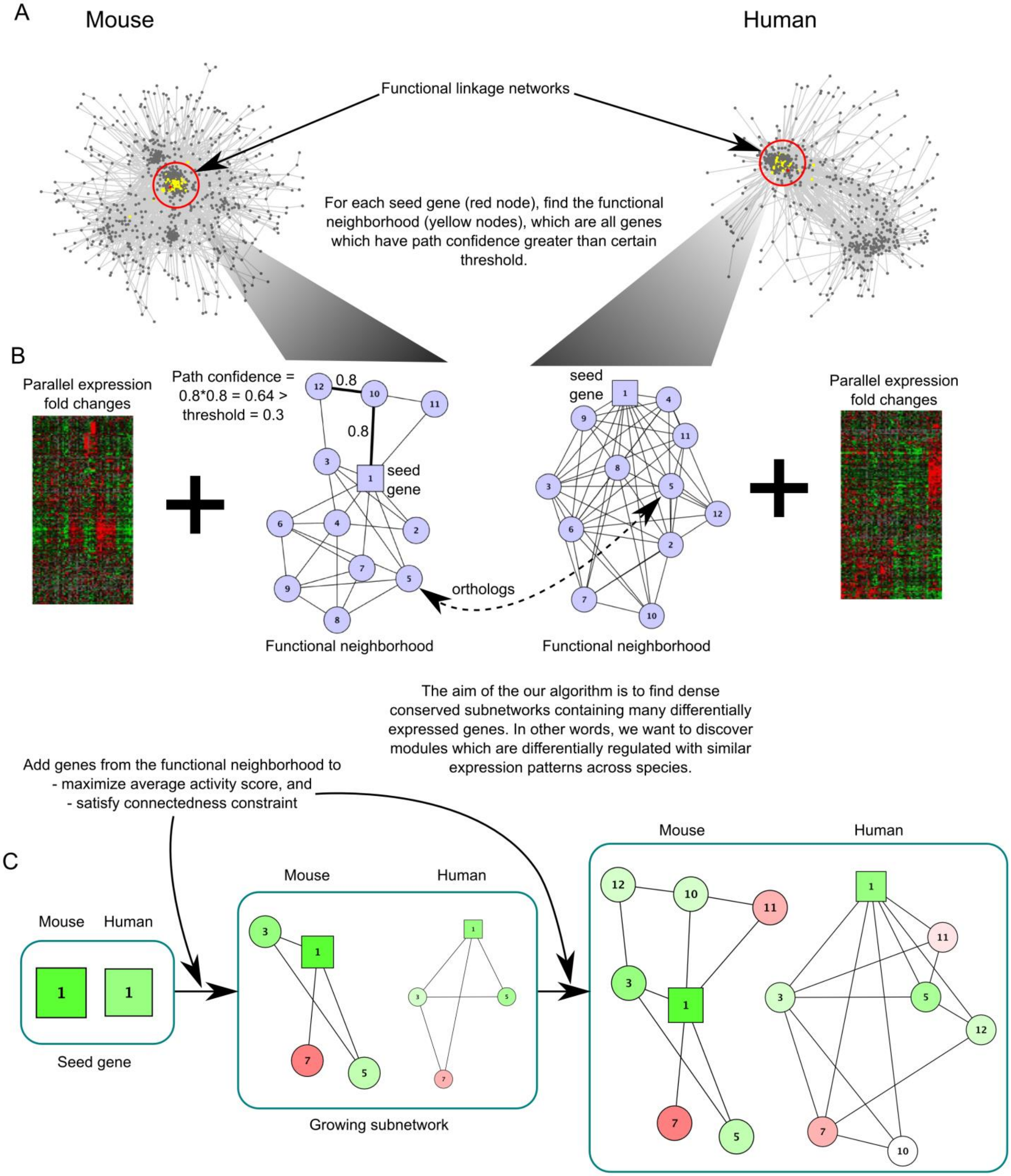
Given their more comprehensive coverage of a broad variety of gene relationships, functional linkage networks should allow for more sensitive discovery of networks that are differentially expressed under various conditions of interest. However, with their broader coverage also come several computational issues. Given the fact that functional linkage networks are orders of magnitude more dense than protein-protein interaction networks, existing algorithms for the discovery of dense subnetworks do not easily scale to this problem. Using functional linkage networks from human and mouse as a basis, we applied our scalable cross-species network discovery approach to identify conserved subnetworks that are differentially active in stem cells relative to differentiated cells based on parallel gene expression studies in mouse and human. We show that these conserved patterns are not likely to have occurred by chance, and that they are enriched for known as well as novel stem cell and differentiation-related processes. Another useful application of our approach is to find functional modules which have diverged or which have been rewired across the two species, which has been previously approached using expression data alone [59]. We designed a variation of our cross-species network search approach to find a number of species-specific networks, which likely reflect differences in the active cellular program between mouse and human pluripotent stem cells. Finally, we demonstrate the usefulness of our algorithm by discussing specific examples of subnetworks discovered, some of which highlight the potentially novel candidate genes involved in the maintenance of stem cell pluripotency.

2.3 Results and Discussion

2.3.1 A method for discovering conserved active subnetworks across species

We developed an algorithm to find conserved active subnetworks across species (Figure 2.1). Our approach requires lists of differentially expressed genes and corresponding fold change values in two different species, assumed to represent analogous conditions. The aim of our approach is to overlay gene activity scores on the respective functional linkage or interaction networks to discover dense subnetworks with a large number of differentially active genes with similar expression patterns in both species. Our approach assumes a set of orthologous clusters for the two species of interest and weighted linkage networks in both species, although it can be also applied to binary interaction networks (e.g. protein-protein interaction networks [60]).

Briefly, subnetworks are simultaneously grown in both species from seed genes by adding nearby genes in the interaction networks that maximize the average activity score of the



For each seed gene (red node), find the functional neighborhood (yellow nodes), which are all genes which have path confidence greater than certain threshold.

The aim of the our algorithm is to find dense conserved subnetworks containing many differentially expressed genes. In other words, we want to discover modules which are differentially regulated with similar expression patterns across species.

Add genes from the functional neighborhood to
- maximize average activity score, and
- satisfy connectedness constraint

Figure 2.1 A method for discovering conserved active subnetworks across species.

(A) The flowchart describes the growth of a subnetwork from a candidate seed gene (red) in the functional linkage network. (B) Genes that are functionally related to the seed are defined as those whose path confidence from the seed gene is above a certain threshold (colored yellow in A), and are considered to be the functional neighborhood of the seed. The aim of the approach is to integrate the expression data with functional linkage networks and discover active conserved subnetworks. (C) The candidate subnetwork initially contains the seed gene and is grown by adding genes iteratively from the functional neighborhood so as to maximize the average expression activity score of the genes in the subnetwork. At all iteration steps, the connectivity constraint must be satisfied before a candidate gene is added. The nodes in the growing subnetworks are genes and the edge-weights are derived from the functional linkage network in either species. The genes are colored green if they are up-regulated in stem cells relative to differentiated cells and red if they are down-regulated in stem cells relative to differentiated cells. The color intensity represents the expression normalized fold change in either direction.

subnetwork while at the same time maintaining a minimum desired clustering coefficient of the genes in the subnetwork (see Materials and Methods for details). Subnetwork growth is stopped when the average activity score reaches a minimum threshold. This process is then repeated with each differentially active gene in either species serving as the seed. The result is a set of highly clustered subnetworks with a high density of matched differential expression in both species (see Materials and Methods for details).

2.3.2 Differential expression analysis of a compendium of human and mouse stem cell expression data

To test our subnetwork discovery method, we compiled a compendium of gene expression data for mouse and human pluripotent stem cells. Briefly, 249 mouse and 132 human expression profiles were obtained from several independent datasets from the Gene Expression Omnibus (GEO) database [61]. Our goal was to identify subnetworks whose activity was associated with the maintenance of stem cell pluripotency in both human and mouse. It has been shown that human embryonic stem (ES) lines across the world are identical in expression of key pluripotency markers like Nanog and Pou5f1, but they can show remarkable differences in expression of other lineage specific markers such as AFP, possibly due to different culture conditions and varying levels of spontaneous differentiation in cultures [62]. Thus, we reasoned that a large compendium of data in both species could support a more robust differential expression analysis, free of any biases from individual studies or cell lines. To group expression profiles at similar stages of differentiation, we used non-negative matrix factorization (NMF) [63], which is an unsupervised clustering method (see Materials and Methods for details). Clusters resulting from NMF clearly separated the expression profiles of undifferentiated, pluripotent cells from those that were in early stages of differentiation or late stages of reprogramming. Differential expression analysis (SAM) was then performed between these two classes of samples to identify a set of genes that change in expression as the pluripotent cells start to exit the self-renewal program during differentiation (see Materials and Methods for details). This clustering and differential expression analysis process was performed independently on the mouse and human expression data. The genes deemed significant by this analysis were labeled with activity scores reflecting normalized fold change values (see Materials and Methods for details) and used as input for our subnetwork discovery approach.

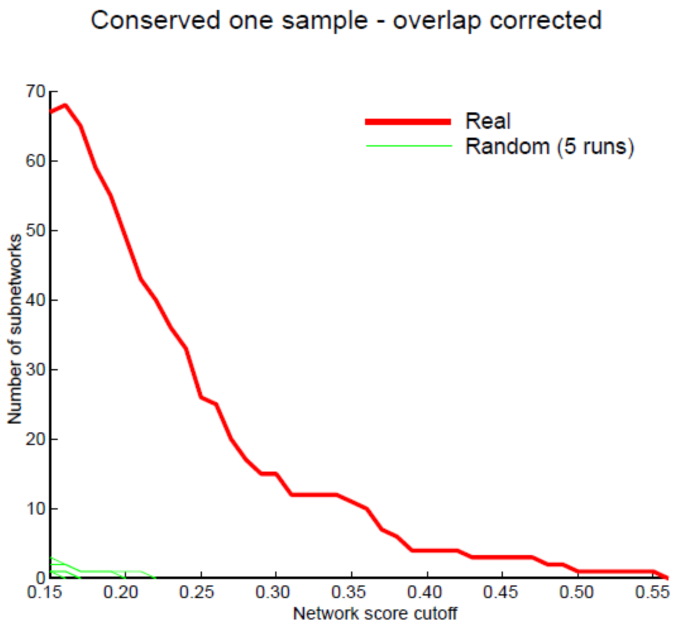


Figure 2.2 neXus applied to a single-dataset differential expression analysis.

neXus was applied to differential expression lists resulting from analysis of one mouse dataset (GSE3653) and one human dataset (GSE9940). For a clustering coefficient constraint of 0.1 on the mouse network and 0.2 on the human network, we plotted the number of distinct subnetworks generated for a range of network score cutoffs. Overlapping subnetworks were removed when their member genes overlapped more than 60% with larger subnetworks. The number of subnetworks obtained given randomized differential expression values for human and mouse across 5 different random instances is also plotted. We observe a similar enrichment over random subnetworks as in the analysis described in the Results section, demonstrating that the approach applies equally well to smaller-scale differential expression analysis.

It is important to note that the method for differential expression analysis (or other means of generating activity scores) is completely independent of the subnetwork discovery algorithm. Our large compendium of stem cell expression data for mouse and human provided an interesting setting for subnetwork discovery, but our approach could also be applied to activity scores derived from more standard, single-dataset differential expression studies, assuming comparable datasets are available for two different species (see Appendix 1, Note 2, "neXus applied to single dataset differential expression study" and Figure 2.2 for an example).

2.3.3 Evaluation of conserved subnetworks

We applied our subnetwork discovery approach to the results of the stem cell differential expression analysis and functional linkage networks from human and mouse. Human and mouse functional linkage networks were obtained from previous work [31,32]. The human network incorporates physical and genetic interactions, sequence information (shared protein domains, transcription factor binding sites), and gene expression profiles [32]. The mouse network incorporates physical interaction data, shared phenotype data, phylogenetic profile information, the yeast functional linkage network where orthologs exist, and gene expression information [31]. These functional networks reflect broad functional relationships between genes or proteins and thus are more general than protein-protein interaction networks (see a detailed discussion in Text S1, Note 1, "Implications of using functional linkage vs. physical interaction networks for active subnetwork discovery”).

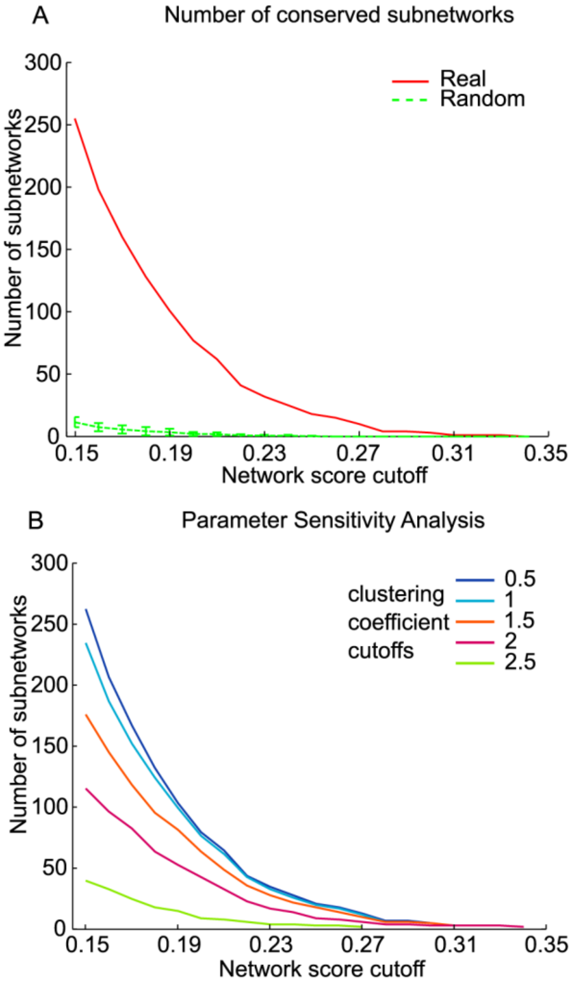


Figure 2.3 **Evaluation of conserved subnetworks.** (A) The cross-species algorithm mines subnetworks in the functional linkage network with a high density of differentially expressed genes. The network score of a subnetwork reflects the average differential activity of all genes in the network. The number of subnetworks identified at a network score threshold is plotted (solid line) and is compared to the number of subnetworks identified after differential expression scores were randomly shuffled (dotted line). The parameters for average clustering coefficient are 0.1 for mouse and 0.2 for human. (B) The number of conserved subnetworks discovered is plotted for a range of connectedness parameters (minimum clustering coefficient). All clustering coefficients noted are relative to the background, single-gene average clustering coefficient, which is 0.08 for mouse and 0.35 for human.

While the input data for these networks are largely independent, physical interaction data for mouse was derived from human interactions (see a detailed discussion in Text S1, Note 3, “Independence of the datasets”).

Conserved active subnetworks between human and mouse were identified by varying the two parameters of the algorithm, the average expression activity (normalized fold change) of the network, and the minimum clustering coefficient. This resulted in between 1 and 255 network(s) from the most conservative to the most lenient parameter settings, respectively. For example, at a network score cutoff of 0.15 (see Materials and

Methods, “Microarray data processing” for fold change normalization), and strict clustering coefficient criteria (> 0.1 for mouse and > 0.2 for human), we found a total of 255 conserved subnetworks involving 607 genes in each of the two species (Figure 2.3A). Increasing the clustering coefficient cutoff or increasing the network score threshold enabled the discovery of fewer, but increasingly confident subnetworks (Figure 2.3B, Figure 2.4).

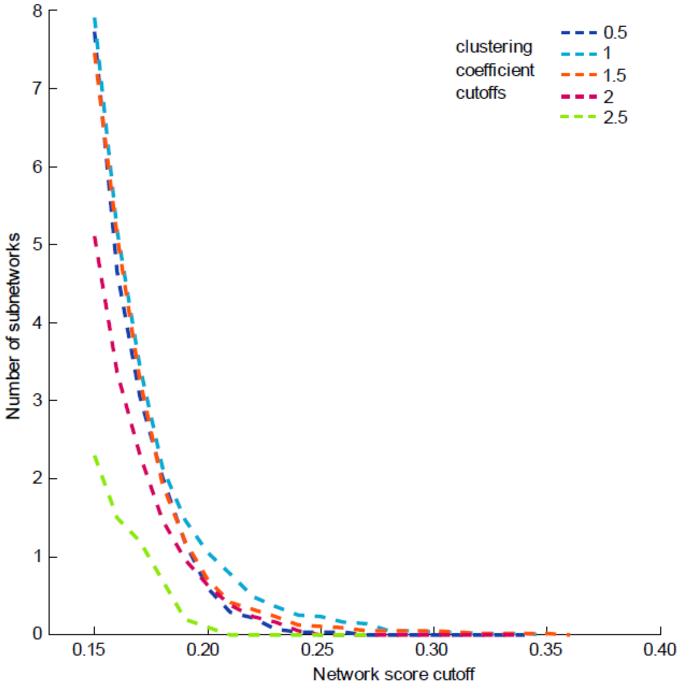


Figure 2.4 Parameter sensitivity analysis to randomized expression data.

The cross-species subnetwork discovery algorithm depends on the setting of two parameters: a network score cutoff and a clustering coefficient constraint. Based on 5 random instances in which the differential expression data were shuffled for both species, this figure shows how the number of random conserved subnetworks discovered varies with changes in both the clustering coefficient and network score parameters. This figure can be compared to the parameter sensitivity analysis of real discovered subnetworks (Fig. 2B). All clustering coefficients noted are relative to the background, single-gene average clustering coefficient, which is 0.08 for mouse functional linkage network and 0.35 for human functional linkage network.

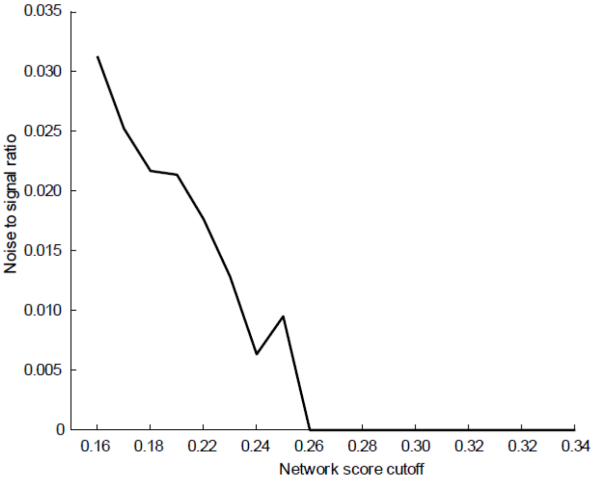
To assess the statistical and biological significance of the networks, we performed a network randomization analysis. Specifically, the expression activity scores in both mouse and human were randomly shuffled five times with respect to the gene labels, and the algorithm was then applied to the shuffled expression profiles. Any conserved patterns of these randomized expression data on the functional linkage network should then represent false positives and not biologically relevant conservation. In all randomization experiments, the functional linkage network structure was retained and only gene activities were shuffled, so that we could specifically estimate the conserved expression patterns arising out of clustering of the active genes by random chance. Importantly, we found that while some subnetworks were discovered in various instances of the randomization experiment, far fewer subnetworks were discovered than for the original expression profiles (Figure 2.3A). For example, at our lenient network score and clustering coefficient cutoffs, we discovered an average of 11.4 subnetworks (standard deviation of 4) across five randomization experiments in contrast to the 255 real subnetworks discovered on the original expression data (Figure 2.3A). Moreover, the average size of the real subnetworks was much larger than the random subnetworks as they contained an average of 22 genes compared to 5.7 genes (standard deviation of 0.6) across the random trials. This comparison clearly suggests that the subnetworks obtained by our cross-species approach are statistically significant, and are not likely to have been discovered by chance. We also found that the signal to noise ratio, which is the ratio of number of real subnetworks to the average number of random subnetworks, improved as we increased the network score cutoff (Figure 2.5) and clustering coefficient cutoffs

(Figure 2.4). This improvement suggests that tuning these parameters is an effective means of isolating high-confidence conserved network signatures for hypotheses generation.

We also evaluated the subnetworks in terms of their functional coverage and relevance. The function enrichment of the genes contained in each subnetwork was measured based on significant overlap with biological processes in the Gene Ontology [45] (see Materials and Methods). A large majority of the subnetworks (235 of 255) were found to be enriched for GO processes, many with suspected involvement in stem cell maintenance and differentiation (Figure 2.6). Furthermore, many subnetworks were monochromatic, that is, they contained genes with concordant changes in expression in either stem cells or differentiated cells. Around a third of the subnetworks were consistently more highly expressed in stem cells while approximately half of them were consistently more highly expressed in differentiated cells. As expected, the

Figure 2.5 Fraction of random to real subnetworks vs. network score cutoff.

For a range of network score cutoffs (average normalized fold change), the crossspecies subnetwork discovery approach was run on the real differential expression values as well as on several random instances, where the differential expression data were shuffled with respect to the gene labels. At each parameter setting, the ratio of the number of subnetworks obtained from the random instances was measured relative to the number of real subnetworks (noise to signal ratio). The parameters used for this experiment are clustering coefficient 0.1 and 0.2 for mouse and human respectively and > 0.15 for network score cutoff.



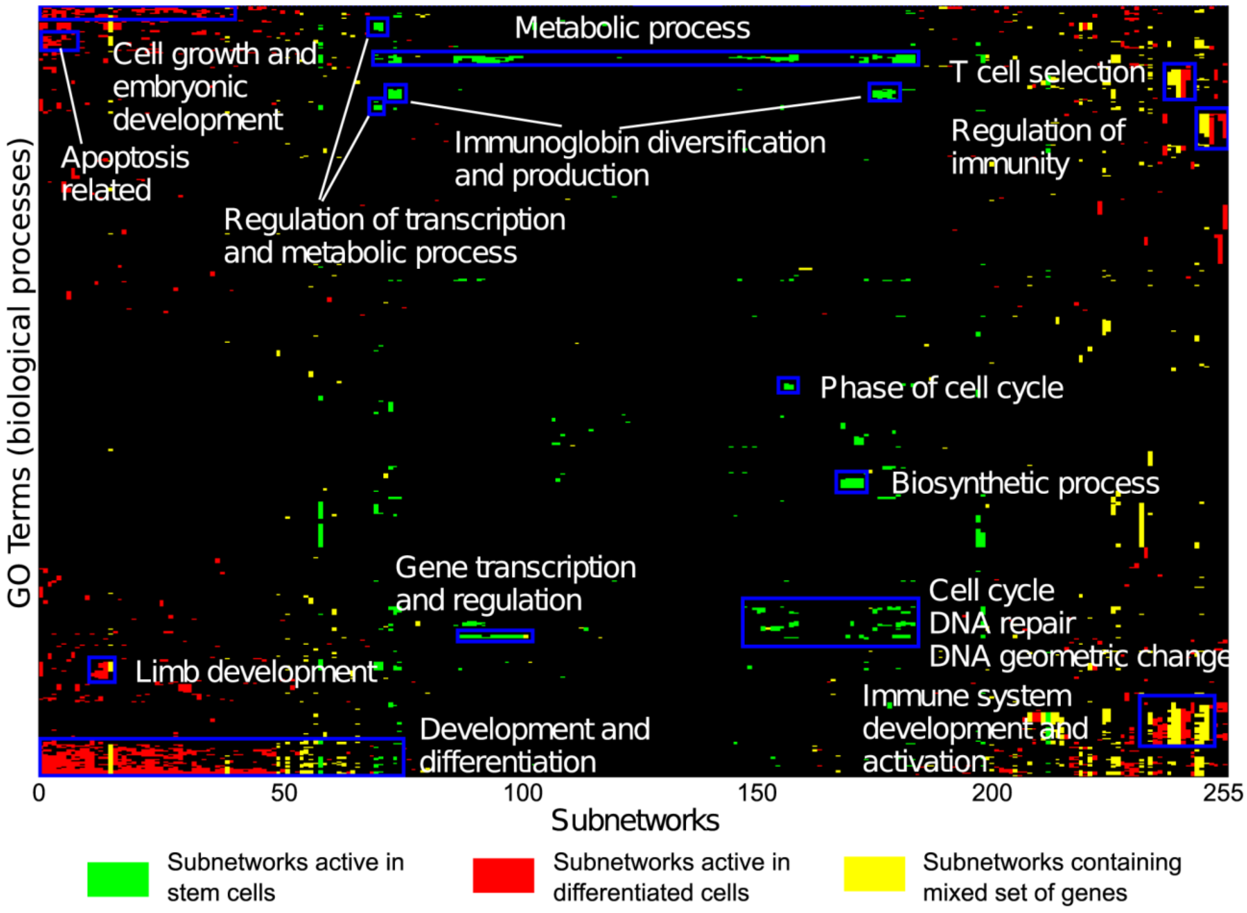


Figure 2.6 Functional summaries of the subnetworks.

The 2D hierarchically clustered matrix of subnetworks' functions highlights functional enrichments based on Gene Ontology annotations (biological process category) for the mouse counterparts of all conserved active subnetworks. A subnetwork column is colored green if the subnetwork contained genes predominantly up-regulated in stem cells, red if the genes in the subnetwork are up-regulated in differentially expressed cells, and yellow, if the subnetwork contains mixed genes, some of which are more highly expressed in stem cells and some in differentiated cells. Enrichment was measured for all GO terms (Bonferroni-corrected $p<0.05$), and the enrichment patterns were clustered to reveal patterns of enrichment across the subnetworks. Enriched GO Terms for individual subnetworks have been uploaded on the subnetworks website and can be browsed at http://csbio.cs.umn.edu/neXus/subnetworks.

monochromatic subnetworks active in stem cells were found to play a role in metabolic processes and regulation, biosynthetic processes, cell cycle, DNA repair, and gene transcription and regulation (Figure 2.6). On the other hand, the monochromatic subnetworks active in diffe rentiated cells were involved in development and differentiation of various cell types, tissues and organs (Figure 2.6). We also noted another interesting class of subnetworks that showed mixed changes in expression, including a combination of up and down-regulated genes, whose patterns matched across species. This class may highlight pathways that require or at least exhibit dramatic imbalances in gene expression to maintain stem cell state.

2.3.4 **Comparison to gene expression overlap**

We compared conserved subnetworks discovered by our approach to gene sets obtained from a simple intersection of orthologs on the human and mouse differentially expressed gene lists. One might suggest that a reasonable approach to finding the core conserved modules underlying stem cell pluripotency is to simply analyze the most extreme differentially expressed genes in both species. We attempted this approach by comparing the top 600 differentially expressed genes from mouse and human, which is comparable to the total number of genes contained across our subnetworks. There was relatively low overlap between the gene sets: of the 600 genes, only 36 are up-regulated in the both species while 34 are down-regulated (Figure 2.7). This level of agreement is higher than the number expected by chance (~15-20),

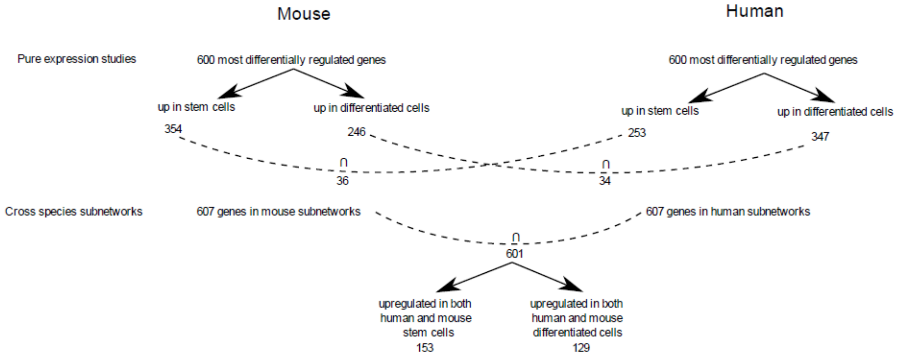
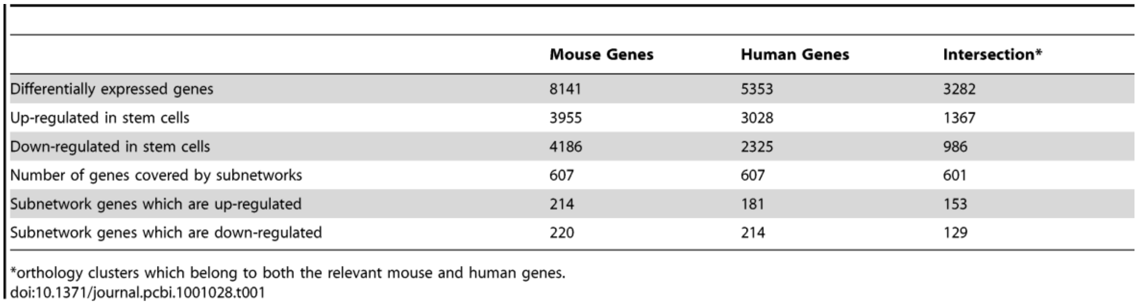

Mouse
Human
Pure expression studies
600 most differentially regulated genes
600 most differentially regulated genes
up in stem cells
up in differentiated cells
up in stem cells
up in differentiated cells
354
246
253
347
∩
36
∩
34
Cross species subnetworks
607 genes in mouse subnetworks
607 genes in human subnetworks
∩
601
upregulated in both human and mouse stem cells
153
upregulated in both human and mouse differentiated cells
129

Figure 2.7 Analysis of ortholog overlap in differential expression lists vs. conserved subnetworks.

To address the question of whether the core conserved modules involved in stem cell pluripotency could be identified by simply comparing the most highly differentially expressed genes in both species, we compared among differentially expressed genes to that obtained from our subnetworks. Specifically, we selected a subset of the significantly differentially expressed genes (based on SAM) that was similar in size to the total number of genes that appear in the human and mouse subnetworks produced by our approach (~ 600 genes). This gene list contained roughly half up- and half down-regulated genes. We then measured the intersection (based on our orthology mapping) between the human and mouse gene lists, which resulted in 36 up-regulated and 34 down-regulated genes in common. Although this overlap is highly statistically significant, it is much lower than the overlap between the mouse and human gene lists in the subnetworks produced by our approach (overlap of 601 as compared to 70). The subnetworks from our approach were obtained with clustering coefficient constraints of 0.1 on the mouse network and 0.2 on the human network and a network score cutoff of 0.15.



	Mouse Genes	Human Genes	Intersection*
Differentially expressed genes	8141	5353	3282
Up-regulated in stem cells	3955	3028	1367
Down-regulated in stem cells	4186	2325	986
Number of genes covered by subnetworks	607	607	601
Subnetwork genes which are up-regulated	214	181	153
Subnetwork genes which are down-regulated	220	214	129

*orthology clusters which belong to both the relevant mouse and human genes.
doi:10.1371/journal.pcbi.1001028.t001

Table 2.1 Gene expression overlap

but certainly not as high as one might expect, suggesting that there are a number of core modules that do not exhibit the most extreme expression changes. The overlap does improve when we consider any genes that show significant changes in expression (FDR 5%): 1367 genes are significantly up-regulated in pluripotent stem cells in both human and mouse while 986 are significantly down-regulated, which reflects an overlap of ~50% (Table 2-1). However, this more lenient cutoff yields thousands of candidate genes to consider, which makes determination of the core conserved modules difficult. Our conserved subnetworks offer a solution to this problem: we find 255 modules containing approximately 600 genes that appear in both the human and mouse subnetworks, including 282 that are differentially expressed and show similar expression patterns. Simultaneous network discovery guided by the combined differential expression data allows us to directly identify the core conserved patterns of expression, even where some of these patterns are subtle but consistent.

We were intrigued by the fact that our conserved subnetworks actually contained a significant fraction of genes (~20%) that showed no evidence of differential expression. By its design (see Materials and Methods, Algorithm), the subnetwork discovery algorithm can include non-differentially expressed genes in identified subnetworks if they connect across highly differentially expressed genes. Briefly, for a given seed gene, the algorithm starts by finding the surrounding functional neighborhood of that seed, which is defined as the set of genes that can be reached within a given path confidence (the product of linkage weights along the path). From this set of genes in the functional neighborhood, the gene that results in the greatest increase in the network activity score is added to the current subnetwork, including any genes required for its connection to the seed. The addition of the corresponding path can potentially bring in non-differentially expressed genes, which may reflect genes that are causally linked to the corresponding subnetwork but whose activity is simply post-transcriptionally regulated [24]. Their activity may be modulated at the protein level which is typical of transduction pathways that control gene expression programs [24]. For example, *TEP1* is not differentially expressed but is found in an active subnetwork with many well-characterized stem cells genes like *POU5F1* (Figure 2.8). TEP1 is involved in telomerase activity [64] and has been shown to be regulated by phosphorylation in breast cancer cells [65]. These examples illustrate the advantages of integrating differential expression data with the broader relationships captured by functional linkage networks in that complete modules can be identified, including genes whose activity is not necessarily transcriptionally regulated.

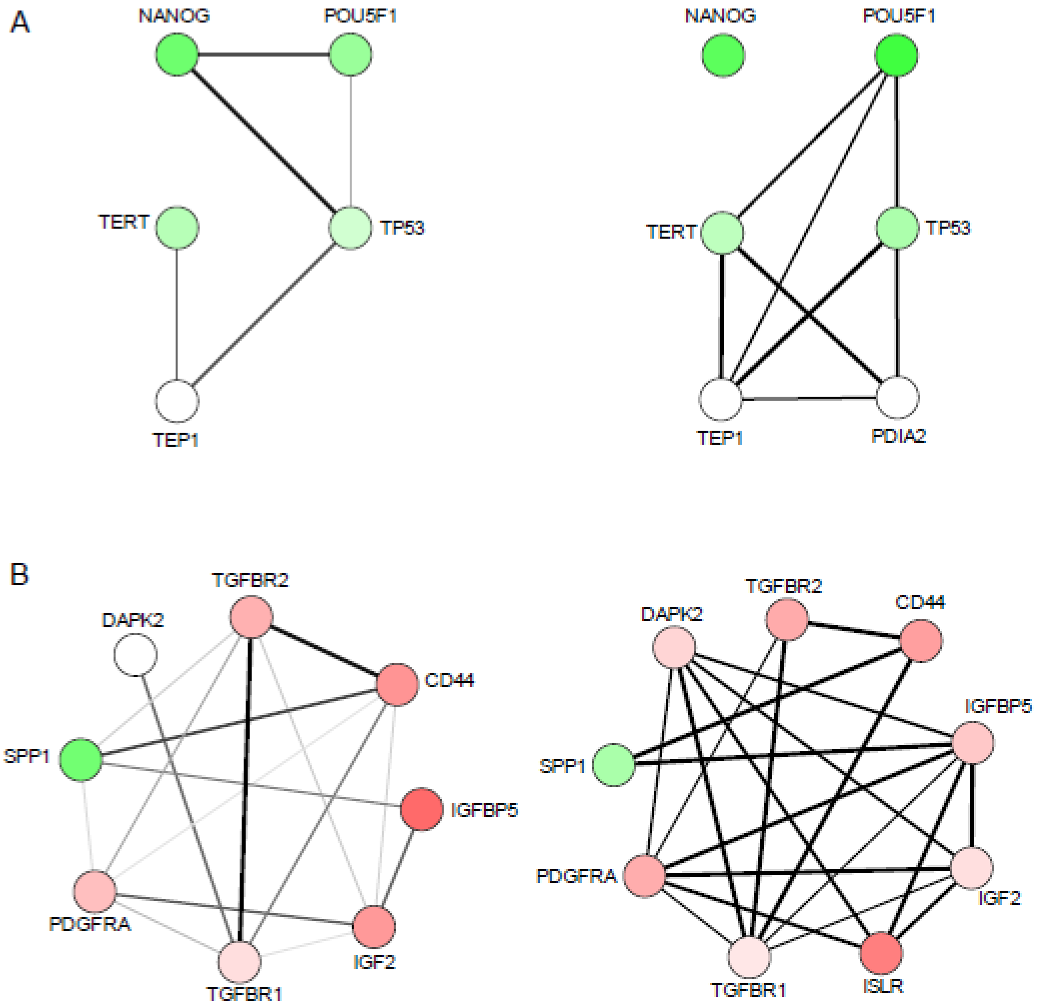


Figure 2.8 Example conserved active subnetworks.

Subnetworks (a-b) are interesting subnetworks discovered by the cross-species network search algorithm on differentially expressed genes between stem cells and differentiated cells. Each subnetwork represents a subgraph of the mouse (left column) and human (right column) functional linkage networks. Nodes are genes, and they are colored green if they are pregulated in stem cells relative to differentiated cells. The intensity of the green or red color of the genes represents the normalized fold change in expression. The edge thicknesses in the subnetworks represent the edge confidence based on the functional linkage networks. The subnetwork (a) shows that TEP1 is not differentially regulated in the subnetwork enriched for transcription factor genes. The subnetwork (b) is an interesting case where both up-regulated and down-regulated genes are found in the subnetwork.

The subnetworks also sometimes contain mixed expression signatures (both up- and down-regulated genes) that are conserved across species, highlighting genes in the same pathway that are antagonistic or genes that exhibit different interactions at various stages of development. For example, one conserved network with mixed expression changes was centered about the important extracellular structural protein ostepontin (also known as secreted phosphoprotein 1, SPP1) (Figure 2.8B). SPP1 is highly up-regulated in both mouse and human stem cells while its surrounding subnetwork is significantly down-regulated in comparison to differentiated cells in both species. Osteopontin is known to be highly expressed in bone and other cell types like smooth muscle cells, endothelial cells and hematopoietic stem cell niches. The subnetwork captures some well-known interactions of SPP1 in these cells. For example, osteopontin has been shown to be a ligand for CD44 in tumor cells [66]. Pou5f1 has been shown to bind to the preimplantation enhancer element of osteopontin, and thus, the expression of the two proteins is highly correlated in early mouse embryonic development [67]. The induction of osteopontin in immortalized mouse embryonic fibroblasts, in response to TGF-β2, has been shown to promote the maintenance of undifferentiated human embryonic stem cells [68]. This is attributed to the presence of a TGF-β responsive element in the osteopontin enhancer. Thus, osteopontin likely plays a pivotal role in the maintenance of both human and mouse embryonic stem cells, and this subnetwork supports this idea. The functional linkages of osteopontin in early embryonic cells have not been fully elucidated yet, but this subnetwork suggests that this gene may play a role in the embryonic context since the other genes in the subnetwork show an opposing expression pattern. These interesting cases would not be readily discovered through a simple comparison of differential expression lists across species.

2.3.5 Comparison to other single-species network discovery methods

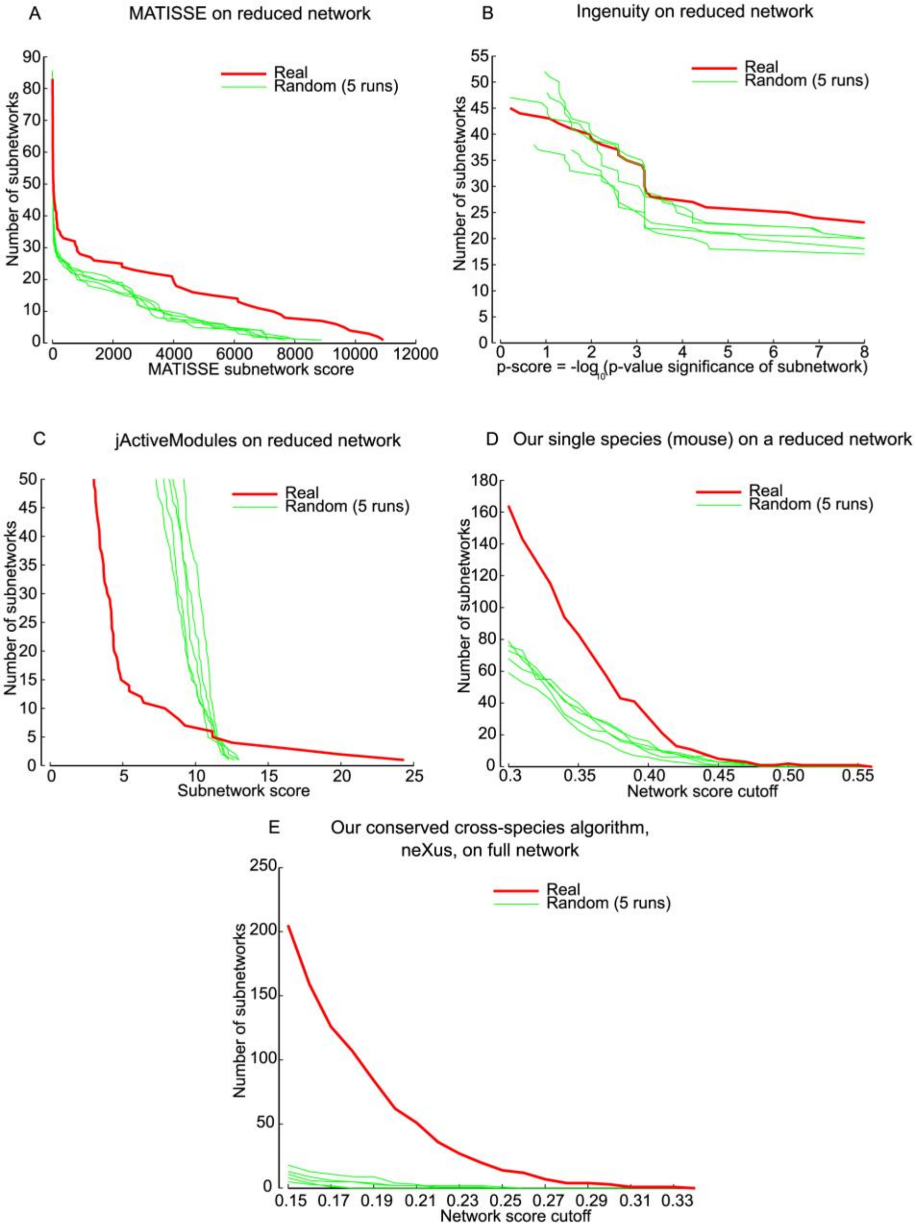
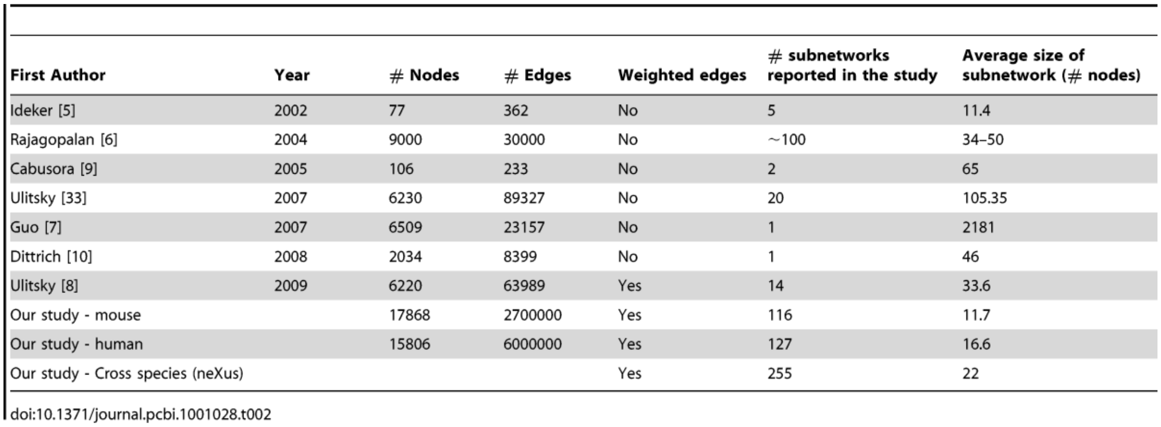


Figure 2.9 Comparison with other methods.

The number of real subnetworks and random subnetworks at various network score cutoffs are plotted for MATISSE (A), Ingenuity (B), jActiveModules (C) and the single-species version of our algorithm (D). The network scores are the metric used by each algorithm to rank the subnetworks. Random subnetworks were obtained by running respective algorithms on the expression data, whose gene labels have been randomly shuffled. Each of the methods uses different forms of the expression data: MATISSE uses expression profiles; jActiveModules uses significance values of the genes; Ingenuity uses focus genes, for which we took any differential expressed gene whose log fold change value was greater (lesser) than 20% of the maximum (minimum) of the most up-regulated (down-regulated) gene; Our method uses fold change scores from the SAM analysis. The scale of the functional linkage network was reduced for all methods shown in (A–D) for a fair comparison. The cross species algorithm on the full network has also been shown for a complete comparison (E).

To our knowledge, our method is the first attempt to interpret differential expression data by integrating with interaction networks across multiple species. Thus, we further assessed the advantages of simultaneous, cross-species network search as compared to active subnetwork discovery in a single species, which has been the focus of previous methods [41,43,44,50,52], and is the principle behind commonly used analysis tools such as Ingenuity Pathway Analysis (Ingenuity® Systems, www.ingenuity.com). Analogous experiments to those performed on our cross-species algorithm were applied to discover active subnetworks in the mouse functional linkage network alone (see Materials and Methods). Most of the existing approaches did not scale to the complete functional linkage network used by our approach (Table 2-2), so we reduced the scale of the mouse functional linkage network by restricting the network to the 50,000 highest weight edges to allow for a direct comparison of our approach to other methods in the single-species context. We implemented MATISSE [42], jActiveModules [41] and Ingenuity (Ingenuity® Systems, www.ingenuity.com) on the mouse data and compared with a single-species version of our approach as well as our cross-species algorithm. For methods that do not incorporate weighted edges, we binarized the reduced network. To allow a direct comparison of the number of subnetworks produced by each approach, subnetworks were sorted in descending order by size and overlapping subnetworks were removed when their overlap with larger networks (in genes) was greater than 60%. To estimate the significance of the subnetworks identified by each algorithm, we randomized the gene labels in the expression data and ran each algorithm five times on randomized expression data. The number and scores of subnetworks produced by each algorithm were compared with the number and scores of the subnetworks generated from the 5 runs on randomized expression data (Figure 2.9).



First Author	Year	# Nodes	# Edges	Weighted edges	# subnetworks reported in the study	Average size of subnetwork (# nodes)
Ideker [5]	2002	77	362	No	5	11.4
Rajagopalan [6]	2004	9000	30000	No	~100	34–50
Cabusora [9]	2005	106	233	No	2	65
Ulitsky [33]	2007	6230	89327	No	20	105.35
Guo [7]	2007	6509	23157	No	1	2181
Dittrich [10]	2008	2034	8399	No	1	46
Ulitsky [8]	2009	6220	63989	Yes	14	33.6
Our study - mouse		17868	2700000	Yes	116	11.7
Our study - human		15806	6000000	Yes	127	16.6
Our study - Cross species (neXus)				Yes	255	22

doi:10.1371/journal.pcbi.1001028.t002

Table 2.2 Comparison to previous approaches

Although our main contribution in this work is the cross-species algorithm, we found a single-species version of our approach performed favorably in comparison to existing approaches (Figure 2.9). Specifically, it produced more subnetworks than other approaches on the real expression data while producing far fewer subnetworks on the randomized data (Figure 2.9).

Surprisingly, we found that 2 of the 3 existing approaches (Ingenuity and jActiveModules) produced as many or more networks on the randomized data as on the real data for most score cutoffs (Figure 2.9B-C). Among the existing methods we evaluated, MATISSE provides the best performance, often reporting 1.5-2 fold more real networks at a given score cutoff than on randomized data (Figure 2.3A). There was significant variation in the size of subnetworks produced across the various approaches, with some producing networks as large as 2000 genes and others producing relatively small subnetworks consisting of less than 10 genes (Figure 2.10). The most useful number and size of networks will, of course, depend on the application, but one particularly unique feature of our implementation is that subnetworks are captured at all stages of their growth, thus giving the user to control of the tradeoff between size and significance of the subnetwork in consideration (see Web Interface section).

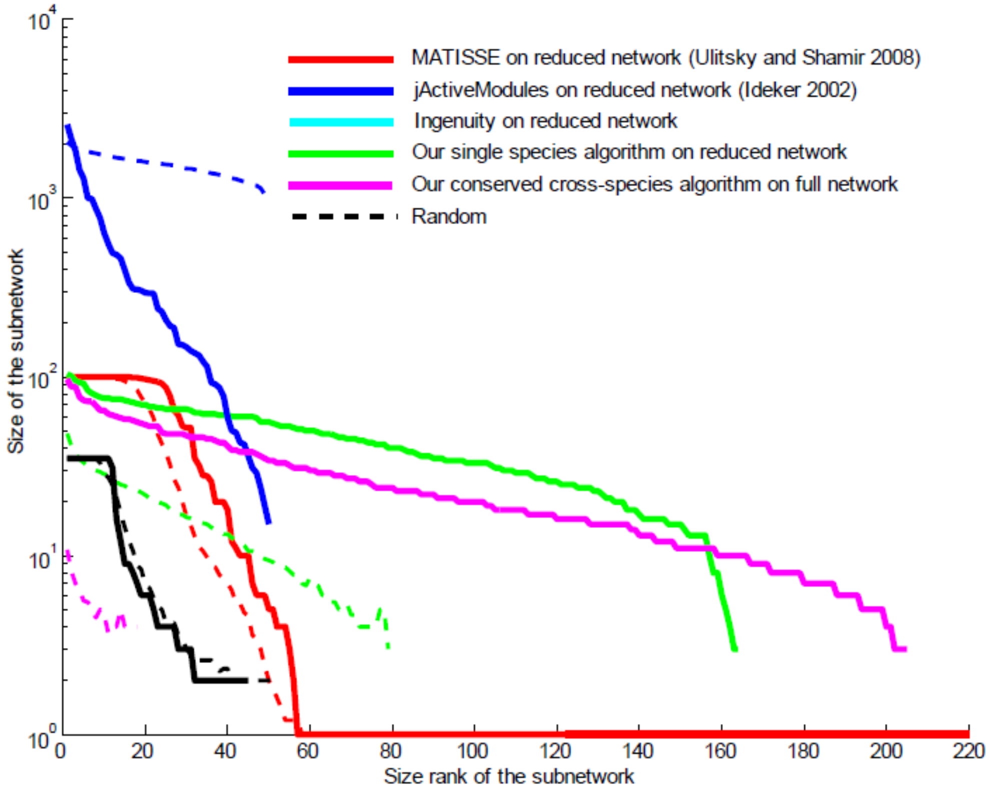


Figure 2.10 Cumulative size distribution of subnetworks generated by existing methods.

All methods were run on the mouse reduced functional linkage networks (50,000 highest weight edges). For each method, the subnet works were sorted in term of the sizes and the sizes were plotted against their rank in the sorted list. The greater the difference between the real and random curve, the greater the confidence we can have in the biological significance of the real subnetworks. To display the utility of our cross species approach, we ran the approach (clustering coefficient parameters > 0.1 and > 0.2 for mouse and human, respectively and network score > 0.15) on the full functional linkage networks which are also shown for comparison.

Perhaps the most striking result of our comparison was our finding that any single species approach, including our own, performed much worse than our cross-species subnetwork discovery algorithm. For example, in the single-species setting for mouse, we were able to find 164 subnetworks while discovering an average of 71 (standard deviation of 7.8) subnetworks in our randomization experiments under the same setting (mouse, clustering coefficient threshold = 0.1, network score cutoff = 0.3), suggesting an enrichment of approximately 2.5-fold (Figure 2.9D). Using the cross-species approach,

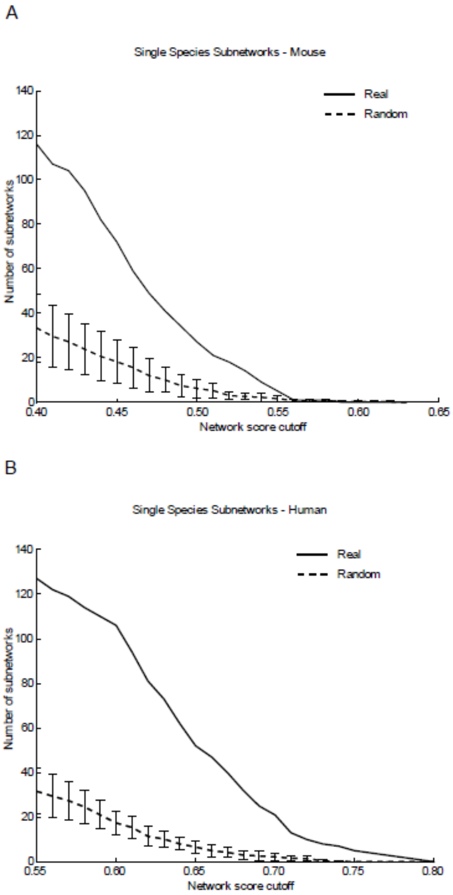


Figure 2.11 Evaluation of single species approach.

The figures show the comparison of number of real subnetworksto average of random subnetworks over multiple experiments (5), when the single species variant of the network search algorithm was applied to the human and mouse expression data and functional linkage networks. The number of subnetworks identified at increasingly network score criteria is indicated when the algorithm was applied independently to (A) mouse (clustering coefficient criterion > 0.2) and (B) human (clustering coefficient criterion > 0.5).

we found 234 subnetworks while discovering an average of 9.8 (standard deviation of 4.16) in our randomization experiments (parameter setting: mouse and human clustering coefficient thresholds = 0.1 and 0.2, network score cutoff = 0.15), which represents a 20-fold enrichment (Figure 2.3B). Thus, not only did we discover more candidate networks in the cross-species setting, but the networks we found were of higher statistical confidence. Similar results were obtained when we applied our single-species approach to the complete functional linkage network (Figure 2.11).

The improvement in sensitivity and specificity by the cross-species approach is a particularly interesting result because it suggests that simultaneous cross-species network discovery can serve as an effective means of improving the signal-to-noise ratio in network discovery even if one is not necessarily interested in asking questions about conservation across species. More pessimistically, this result suggests that separating biologically relevant active subnetworks from random networks based on a single functional linkage network is a challenging problem.

The enhanced performance of the cross-species approach can be attributed to the fact that coordinated expression changes can be reasonably clustered in both species' functional linkage networks. Due to the small-world nature of functional linkage networks (or protein-protein interaction networks)[69], given a large set of genes, subnetworks involving partitions of this set can often be readily found even if these genes do not necessarily play a specific role together. The coherent grouping of genes across species eliminates random aggregation of active genes, and thus, the cross-species approach is able to relax both the network score and clustering coefficient stringency criteria, while still maintaining statistical confidence in the networks. Indeed, when our approach was applied independently to mouse and human data, we found little intersection among the two species' subnetworks: of the genes covered by human (305 orthologous clusters) and mouse subnetworks (261 orthologous clusters), only 21 were overlapping. In contrast, the cross-species approach discovers around 250 subnetworks covering 607 genes in both mouse and human (Table 2-1). We obtained a similar result when comparing to subnetworks derived from another approach, MATISSE, applied to the human and mouse data (see Appendix 1, Note 4, "Comparison of the overlap of mouse and human subnetworks discovered through MATISSE and neXus"). Thus, in addition to the underlying biological question of conservation of expression signatures, cross-species analysis can serve as an effective noise filter, which is critical for discovering clustered patterns of expression changes in a dense interaction network.

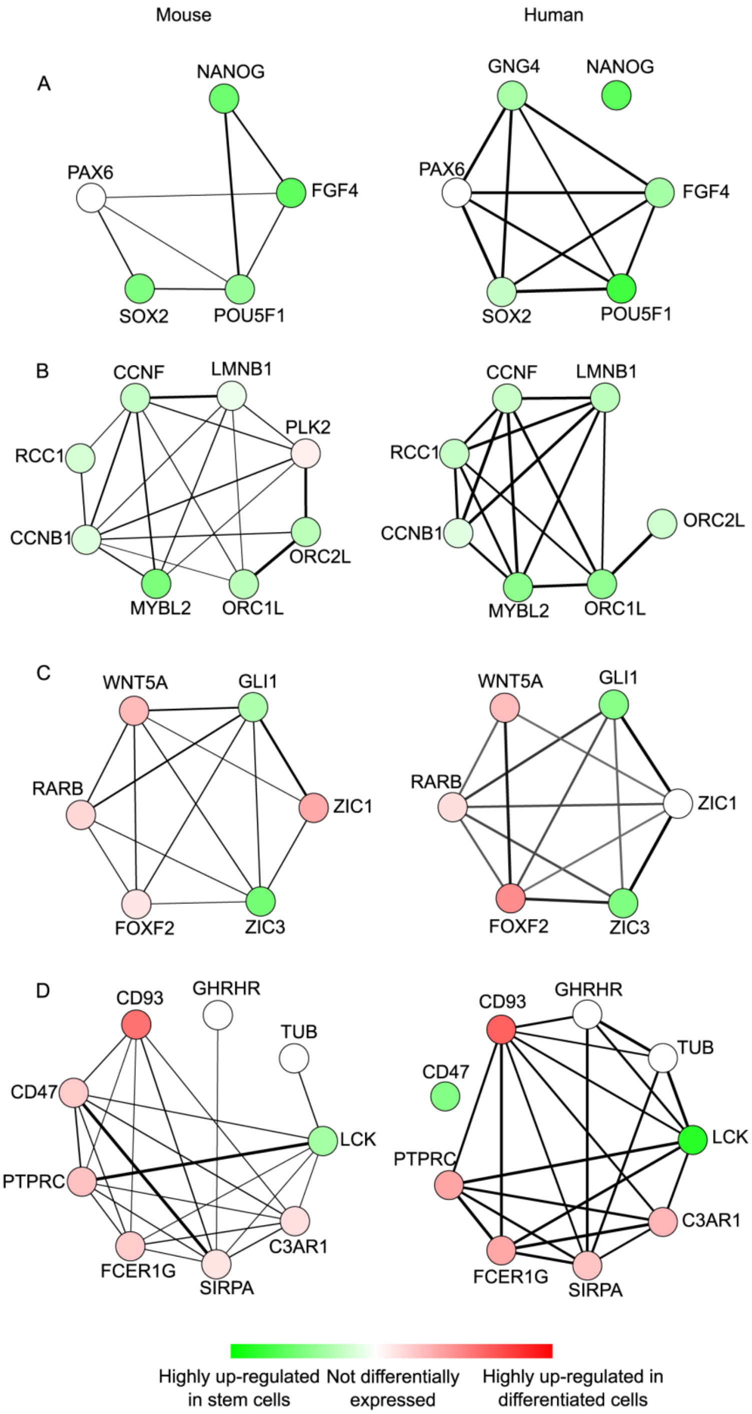
The difficulty in identifying subnetworks from a list of genes within a single species has important implications for how the statistical significance of such networks should be assessed. This problem often arises in practice during the interpretation of candidate gene lists. For example, analysis tools such as Ingenuity Pathway Analysis (Ingenuity® Systems, www.ingenuity.com) are now being widely used based on the single-species discovery method we evaluated above. The significance of networks identified by such approaches are typically assessed by comparing the network score after optimization to scores that obtained by randomly sampling a similarly sized set of genes. However, as demonstrated above, high-scoring networks are often obtained when search algorithms are applied to randomly selected candidate genes. Put simply, in many protein interaction networks, random lists of genes are much easier to connect than one might expect. Our results suggest that significance should instead be estimated by applying the network search process (with the same parameters) to several random candidate genes lists, and evaluating the actual scores in the context of the resulting random score distribution.

2.3.6 Discussion of specific examples

Using the cross-species network discovery algorithm, we are able to find subnetworks reflecting conserved functional modules between mouse and human pluripotent stem cells. We found many of these subnetworks to be monochromatically active in stem cells or differentiated cells. This was not a prerequisite for network discovery, but reflects that the majority of genes supporting a local process are regulated in the same direction. Monochromatic subnetworks up-regulated in stem cells were our primary focus because these reflected potential candidate processes that are necessary for maintaining a pluripotent, self-renewing stem cell state. One of the most significant conserved subnetworks of this type captures the core pluripotency circuit in embryonic stem cells (Figure 2.12A). This network recovers associations between important transcription factors such as *POU5F1, NANOG, SOX2* and *FGF4,* all of which have been shown to form an important transcriptional circuit in embryonic stem (ES) cells, consisting of feed-forward and autoregulatory loops [70]. Chromatin immunoprecipitation experiments have shown that these three proteins exhibit a significant overlap in their binding sites in the genome [70,71]. The subnetwork links FGF4 to the core signaling circuitry formed by POU5F1, SOX2, and NANOG. FGF4 has been shown to be expressed in the peri-implantation mouse embryo [72] and the SOX2/POU5F1 complex has been shown to activate transcription of *FGF4* by binding to an enhancer element [73]. The role of this module has also been studied quite extensively in early embryonic development. FGF4 null mutants in mouse are embryonic lethal due to defective primitive endoderm [74]. The cells of the mouse inner cell mass (ICM) show a reciprocal expression pattern of FGF4 (ligand) and FGFR2 (receptor). It has been shown that the FGF4 secreted by the epiblast precursor cells is crucial to the differentiation and maintenance of cells of the trophectoderm and extraembryonic endoderm lineages [75,76]. Human ESCs show a striking resemblance to mouse epiblast-derived stem cells in terms of morphology and maintenance culture conditions, amongst other characteristics [77,78]. Thus, this network highlights a core, conserved module active in the pluripotent cells of both the species, irrespective of the downstream effects on cell signaling and morphology. *FGF4* stimulation of ERK1/2 signaling in mouse ES cells has been shown to facilitate lineage commitment [79]. In human ES cells, FGF signaling promotes self-renewal by directly affecting the expression of *NANOG* [80,81] as well as suppressing expression of genes responsible for reversion to an ICM-like state [82].

Figure 2.12 **Examples of conserved subnetworks.**

Subnetworks (A–D) are examples of interesting conserved subnetworks discovered by the cross-species network search algorithm on differentially expressed genes between stem cells and differentiated cells. Each subnetwork represents a subgraph of mouse (left column) and human (right column) functional linkage networks, respectively. Nodes are genes and they are colored green if the gene is up-regulated in stem cells when compared to differentiated cells and red if down-regulated in stem cells relative to differentiated cells. The intensity of green or red color of the genes represents the normalized fold change of the expression. The edge thickness in the subnetworks represents the edge confidence based on the functional linkage networks. The subnetwork (A) shows a conserved subnetwork which contains important stem cell transcription factors. The subnetwork (B) highlights cell cycle related pathway genes. The subnetworks (C, D) are mixed subnetworks, as they contain both up-regulated and down-regulated genes. The genes are functionally related but their mode of function is antagonistic in nature.



Another highly significant subnetwork discovered by our approach pertains to the control of cell cycle progression in ES cells (Figure 2.12B). Both human and mouse ES cells have a very short G1 phase which can be attributed to the constitutively active CDK2/6 [83,84]. CCNB1 and MYBL2 are two important cell cycle regulators that are expressed at high levels in undifferentiated ES cells and their expression decreases rapidly upon induction of differentiation [85]. This happens even before loss of the important regulator proteins such as POU5F1 or NANOG can be detected. The conserved subnetwork highlights the role of these two genes in the maintenance of cell cycle progression in ES cells. Knockdown of *MYBL2* has been shown to induce polyploidy/aneuploidy in ES cells and *CCNB1* is a known target of *MYBL2* [86]. *B-MYB* is also crucial for inner cell mass development in mice embryos [87]. The role of *CCNF* in embryonic stem cells has not been explored but yeast two hybrid assays have shown that the NLS domain of CCNF can regulate nuclear localization of CCNB1 [88].

Many conserved subnetworks also included genes that are up-regulated during the initiation of differentiation. This supports the idea that the maintenance of ES cell phenotype requires the suppression of differentiation-associated gene expression as well. One interesting example of this phenomenon was highlighted in a third subnetwork discovered by our approach, which was centered on the protein ZIC3 (Figure 2.12C). *ZIC3* has been shown to be required for maintaining pluripotency of mouse embryonic stem cells by suppressing endoderm specification [89] while *GLI1* has an important effect on embryonic stem cell proliferation [90]. These two proteins are known to work in coordination for transcriptional activation or repression [91]. Both of these genes code for DNA binding zinc finger proteins and they share and recognize highly conserved zinc finger domains. The down-regulated genes in the subnetwork, namely, *WNT5A*, *FOXF2* and *RARB*, play important roles in the differentiation of embryonic stem cells [92-94]. It is interesting to observe that these genes have GLI binding sites in their promoter region or cis-regulatory domains, which suggests that GLI1 and ZIC3 could potentially regulate their expression in ES cells [95,96]. Also, GLI proteins participate in regulation of Hedgehog signaling, of which RARB and FOXF2 are members, and GLI is also known to regulate the members of WNT family [97]. These functional interactions and coordinated expression strongly suggest ZIC3 and GLI1 might be responsible for suppressing the expression of genes such as *FOXF2*, *WNT5A* and *RARB*.

This network in particular provides an illustrative example of how subnetwork discovery can provide novel testable experimental hypotheses. This hypothesis could be explored experimentally through RNAi knockdown of *ZIC3* and *GLI1* in embryonic stem cells to check for resultant changes in expression of the other genes in the network. Lim *et al*. [89] conducted RNAi knockdown of *ZIC3* in human and mouse ESCs and saw enhanced expression of endodermal transcripts like *SOX17* and *PDGFRA*. Further experiments could also be used to check for direct binding of *ZIC3* and *GLI1* to the promoter regions of the differentiation-associated genes. The subnetwork also highlights the striking observation that the gene *ZIC1*, despite sharing 69% homology with *ZIC3*, does not show the same trend in expression in either mouse or human pluripotent stem cells. While *ZIC2* and *ZIC3* have been suggested to have partially overlapping or redundant roles in suppressing endoderm in embryonic stem cells, the role of *ZIC1* in this context has been not been explored much. Further overexpression studies of this gene could be used to elucidate its exact role in this network.

Another interesting subnetwork found by our approach was centered around the seed gene SIRPA. The only gene in the whole subnetwork that is found to be up-regulated in mouse and human pluripotent stem cells is *LCK* (Figure 2.12D). *LCK* is one of the eight SRC family kinase genes, which are known to play crucial roles in regulating signals from a variety of cell receptors, affecting a variety of cellular processes such as differentiation, growth and cell shape [98]. Members of this family, namely *Hck* and *Lck*, have been implicated in the maintenance of self-renewal of murine embryonic stem cells [46]. *Cyes*, along with *Hck*, have been shown to be regulated by LIF in mouse embryonic stem cells and the expression of their active mutants allows the maintenance of these cells at lower concentrations of LIF [99]. Other studies have also reported the evolutionarily conserved transcriptional co-expression of *LCK* in human and mouse embryonic stem cells based on transcriptomic studies [100]. LCK has also been shown to induce STAT3 phosphorylation and this is believed to cause transformation of cells having constitutive LCK activity [80]. All of the other genes in the sub-network are down-regulated in ES cells, which may be due to the fact that the expression of *SFK*s is generally associated to lineage-restricted patterns in the adult, such as, the expression of *LCK* in T lymphocytes.

While the hypotheses suggested by the discovered subnetworks ultimately require experimental follow-up, these examples illustrate that the networks capture many of the well-characterized processes supporting stem cell pluripotency as well as implicating some novel players. In general, the process of active subnetwork discovery can play an important role in interpreting differential expression or other genome-wide data. Active subnetworks, and in particular those that are conserved across species, provide evidence that a whole process or pathway is up/down-regulated, which is more definitive than the type of information provided by a differential expression list, for example. A single highly differentially expressed gene is less compelling than an entire functional module with evidence of differential expression. Furthermore, because the underlying functional linkage networks are based on large collections of genomic data, our approach can potentially identify functional modules that are not yet characterized, but that play a critical role under the conditions being studied.

2.3.7 Discovery of species specific subnetworks

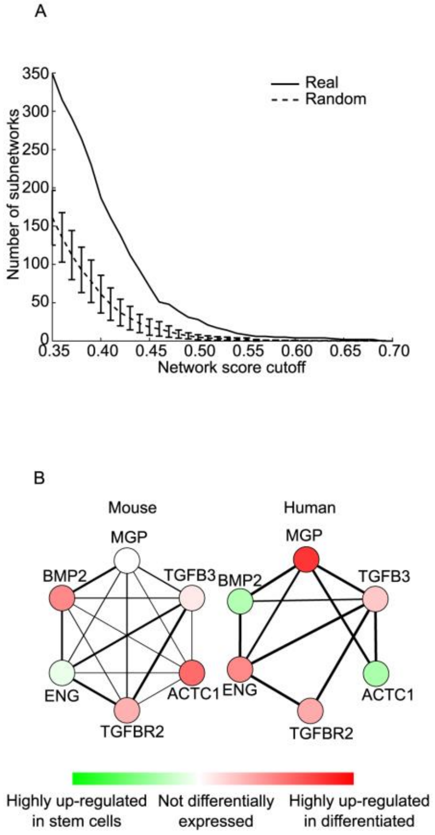


Figure 2.13 Species specific subnetwork.

(A) The number of species-specific subnetworks discovered is plotted versus the network score cutoffs and compared with the number of subnetworks generated by applying the same approach after randomly shuffling gene labels in the expression data. Species-specific networks represent subnetworks with highly divergent patterns across species. (B) An example species-specific subnetwork that highlights the difference in expression of BMP2 pathway related subnetwork in human and mouse. The subnetwork nodes are genes, whose color represent whether are they are active in stem-cells (green) or differentiated cells (red) and intensity of the color represent the degree of expression activity. The thickness of edges of the subnetwork represents the edge confidence based on the functional-linkage network.

We modified the cross-species network discovery algorithm to discover subnetworks that are markedly different in the expression patterns between the two species (see Materials and Methods, "Score of a Subnetwork"). These subnetworks represent tightly interconnected groups of genes or proteins that are active only in one of the species or where the expression changes are in opposite directions, highlighting places where pluripotent stem cell signaling differs between human and mouse. Through randomization experiments similar to the conserved subnetwork identification approach (see Materials and Methods) we found that we were able to find such non-conserved network signatures approximately twice as frequently as on randomized expression profiles (Figure 2.13A). We note that this is a substantially lower signal-to-noise ratio than for the conserved subnetwork discovery approach, for which we achieved approximately 20-fold improvement over random, suggesting that statistically significant species-specific active subnetworks are harder to discover. This is not surprising given that the relatively frequent appearance of random subnetworks in a single species (Figure 2.9D), which cannot be easily classified as statistical artifacts or biologically relevant changes across species. The species-specific network discovery problem is not able to take advantage of the noise filtering property of the conserved network search described above.

Nevertheless, we find interesting subnetworks which highlight differences between gene expression in mouse and human stem cells. For example, one species-specific subnetwork (Figure 2.13B) recapitulates the well-known difference in BMP signaling between human and mouse embryonic stem cells. Mouse embryonic stem cells require BMP2/BMP4 to induce the expression of Inhibitor of differentiation (Id) genes via Smad pathway for self-renewal [101]. Thus, exogenous addition of LIF and BMP4/2 is required to maintain mouse ES cells in culture without differentiation. On the other hand, human ES cells cultured in unconditioned medium exhibit high levels of BMP signaling which causes the cells to differentiate. Mouse epiblast stem cells, like human ES cells, differentiate to trophoectoderm upon BMP4 induction [77]. This needs to be suppressed through an antagonist such as noggin to maintain these cells in an undifferentiated, self-renewing state [81]. The other genes in the subnetwork that show opposite trends in differential expression between human and mouse ES cells are MGP, ACTC1 and ENG. Endoglin (Eng) is an accessory receptor for several TGF-β growth factors, including BMP2, and has been shown to be crucial for embryonic hematopoiesis [102]. Matrix GLA protein (MGP) is a small matrix protein that has been shown to have a direct interaction with BMP2 and has been shown to modulate BMP signaling [103]. The potentially disparate role of these genes in mouse and human ES cells can be explored further.

2.3.8 Web Interface

To facilitate public access to active cross-species subnetworks identified by our approach, we developed a web-based interface for convenient browsing of conserved and species-specific stem cell expression signatures (http://csbio.cs.umn.edu/neXus/subnetworks, download subnetworks in raw text from http://csbio.cs.umn.edu/neXus). Subnetworks are listed according to their corresponding network seed gene, and when a seed gene is selected, the following information is displayed: the conserved active human and mouse subnetworks, significance of the identified subnetwork based on a comparison to network randomization, expression fold changes and name details of mouse and human genes, and the function enrichments of the genes in respective to human and mouse subnetworks based on the Gene Ontology [45]. The subnetwork generation was automated using neato, a Graphviz graph plotting tool with spring model layouts [104]. The Cytoscape version of the subnetworks are also available on the website, which are linked using Cytoscape Webstart [105]. The gene names in the subnetwork are linked to gene information at the Mouse Genome Informatics (MGI) database [106] and GeneCards [107] for mouse and human genes, respectively. Another useful feature of our web-interface is that subnetworks can be interactively expanded based on the cross-species discovery algorithm, which allows for real-time analysis of additional candidate genes that are closely associated with the network of interest. As networks are expanded, a statistical significance score is calculated after each iteration, which allows the user to estimate the potential biological relevance of the network as it is expanded.

2.3.9 Conclusion

We have described a scalable approach for discovering conserved active subnetworks across species. Starting from candidate gene lists reflecting parallel differential expression studies in two different species, we are able to search for dense subnetworks with conserved patterns of differential expression. In contrast to previous active subnetwork discovery algorithms, our approach not only extends this idea across species, but also enables application of the approach to functional linkage networks as opposed to sparse protein-protein interaction networks. Functional linkage networks integrate information from a diverse collection of genomic and/or proteomic studies (including protein-protein interactions), and thus offer the potential for more sensitive discovery of active subnetworks, including those which involve previously uncharacterized genes.

We applied our approach to a differential expression study between pluripotent mouse and human stem cells versus their differentiated cell types to produce several hundred subnetworks that reflect conserved changes between mouse and human. Network search across species produced specific hypotheses about conserved and differentiated mechanisms of stem cell maintenance, and importantly, demonstrated that such an approach can be an effective means of filtering noise from the active subnetwork discovery problem. We found that identifying statistically significant active subnetworks independently within a single species may be a harder problem than previously appreciated, and we suggest the cross-species approach as one solution to this problem.

Despite the success of our approach, there are a number of promising directions for further improvement and broader application of the method. While the approach was successfully applied to relatively dense functional linkage networks for mouse and human, it is a computationally challenging problem, and the algorithm cannot be applied in real-time as it still requires several days to run. Strategies for improving the efficiency of conserved network discovery and more formal selection criteria for the parameters associated with our approach are both useful future directions. Furthermore, the approach can be readily extended to discover conserved subnetworks across more than just two species, which will make another fruitful direction as we begin to accumulate functional genomic data across a broad variety of other model organisms. Finally, although our study focused on the interpretation of candidate gene lists derived from differential expression analyses, the algorithm is general and can be readily applied to interpret lists arising from other genomic screens, including, for example, genome-wide association studies.

2.4 Materials and Methods

2.4.1 Microarray data processing

249 mouse microarray data samples were obtained from 20 GEO datasets. All the samples had been hybridized to the Affymetrix mouse chip MOE 430 2.0. 132 human microarray data samples were obtained from 12 GEO datasets. All the samples had been hybridized to the Affymetrix human chip HGU 133 plus 2.0. The raw data was normalized using the MAS 5.0 algorithm **[108]** and the average chip intensity was scaled to 500. The probes set IDs with detection p-values higher than 0.4 were termed absent and were filtered out for further analysis, along with the probe set IDs with average intensity lower than 50. Non-negative matrix factorization (NMF) was used to identify major biological classes in the data in both species independently **[63]**. The algorithm factorizes the expression matrix A into two matrices, W and H. If the expression matrix is of size *N X M*, the algorithm computes an approximation $A \approx WH$, where W and H have sizes *N X k* and *k X M*, respectively **[109]**. Here, k represents the number of clusters that the samples can be divided into. Each of the k columns of matrix W defines a metagene

in such a way that the entry w_{ij} represents the coefficient if gene i in metagene j. Each of the M columns of matrix H depicts the metagene expression profile in different samples such that the entry h_{ij} represents the expression level of metagene i in sample j. The accuracy of the classification is evaluated by the value of the cophonetic coefficient. NMF was used to cluster the samples into biologically meaningful sets. As an example, for k = 6, the mouse samples were clustered into the classes that represented the different levels of pluripotency of the stem cells. The cophonetic coefficient for this classification was 0.978. Similar classification could be achieved for k = 5 in the human gene expression data (cophonetic coefficient of 0.977). As mentioned earlier, the matrix W detected the metagenes representing every cluster of similar samples in the data and, the matrix H gave the expression profile of every sample in the particular metagene. The expression profile of the various samples in the metagene corresponding to the cluster of pluripotent stem cells was used to divide the samples into two major classes, on the basis of the values of the entry h_{ij}. Class 1 included the pluripotent ES cells and induced pluripotent stem cells while class 2 represented samples that were in the process of early differentiation or late reprogramming. Differential expression analysis was performed between these two biological classes using Significance Analysis of Microarrays **[47]**. The results of this differential expression analysis were used as the starting point for subnetwork discovery. The differential expression criteria were set at false discovery rate less than 5%. The results of this differential expression analysis yielded fold changes for significantly differentially expressed genes which was log normalized for both up-regulated and down-regulated genes, separately. The log-ratios were rescaled to ranges from -1 to +1, where -1 represented the gene which is most down-regulated and +1 represented the most up-regulated gene. The majority of the genes were not significantly differentially expressed; the log-ratio of these genes was set to zero. The normalized expression fold change data can be downloaded from http://csbio.cs.umn.edu/neXus.

2.4.2 **Functional linkage networks**

We used the mouse functional linkage network previously published in [31] with all edges below 0.10 confidence set to zero, which resulted in around 2.7 million weighted edges among 17868 genes. We obtained the human functional linkage network from [32] (the "global network") and trimmed the network to the highest 6 million weighted edges, which corresponded to a minimum edge weight of 0.58 and covered 15806 genes.

2.4.3 **Algorithm**

The algorithm identifies functional modules enriched for active genes in both species under consideration. Conserved active modules are found based on two criteria: (1) a high degree of clustering in both species' functional linkage networks, and (2) a high average normalized differential expression fold-change (network score) sharing the same sign across species. Because the search space is exponential, a greedy heuristic is applied to expand subnetworks from candidate seed genes. Each candidate network is grown until it fails to meet one of the constraints. This algorithm is implemented in Python and the source code can be downloaded from the supplementary website (http://csbio.cs.umn.edu/neXus) (see Box 1 for pseudocode). Each component of the algorithm is described in more detail below.

2.4.4 **Score of a Subnetwork**

The network score of a cross-species subnetwork is the average activity scores (described below) of the genes in the two species' subnetworks given that they obey the following constraints: first, the subnetworks satisfy a connectedness constraint on their respective functional linkage network; second, the network score of the subnetwork is above a threshold. In all other cases, the score of the subnetwork is zero. The first condition guarantees that the genes in the subnetwork are interconnected in each species' functional linkage network, which suggests the corresponding set of genes represents a functional module. By enforcing this constraint on both species, conserved modules are selectively chosen. The second constraint guarantees that the subnetwork exhibits a high degree of differential expression, which reflects a coherent response to the phenotype or conditions under consideration.

The connectedness of a subnetwork is quantified by the average weighted clustering coefficient of the subnetwork, which is the ratio of existing connections between the neighbors to the total pairs of neighbors possible. The clustering coefficient for node k is given by $\rho_k = \frac{\sum_{i,j,k=\Delta} 1}{{}^{n}_{2}C}$, where i, j, k = Δ means nodes i, j, k form a triangle in the graph, and n is total number of neighbors of node k. For a weighted network, the clustering coefficient can be modified to $\rho_k = \frac{\sum_{i,j,k=\Delta} w_{ij}}{{}^{n}_{2}C}$ [110], where w_{ij} is the weight of the edge ij. Average (weighted) clustering coefficient is the average of the (weighted) clustering coefficients of all the nodes in the graph, which is given by $\rho = \frac{\sum_{k=1}^{n} \rho_k}{n}$. The network score of the subnetwork is the average of the activity scores across all genes in the subnetwork. For single species subnetwork discovery, the normalized fold change of the gene was used as the activity score. For the conserved cross-species approach, the magnitude of the activity scores of genes were calculated as the geometric mean of magnitudes of normalized fold changes of the genes across the two species. The gene activity scores were assigned the same sign as the product of the signs of the normalized fold changes. This means that if the gene was up-regulated or down-regulated in the same direction in both the species, the gene activity score was positive, while genes showing the opposite direction of differential expression were assigned a negative sign. For the species-specific approach, the absolute difference in the normalized fold changes was used as the gene activity score.

2.4.5 Growing Subnetworks

Subnetworks are grown greedily to optimize the subnetwork score, starting from each gene as a seed. The genes are added from a pool of genes in functional proximity to the seed gene, which are defined by any genes within a minimum path confidence, i.e. the product of all weighted edge confidences in the path, from the seed gene. This pool of genes is discovered using a modification of the depth first search algorithm. Nodes are picked starting from the seed gene, in depth-first fashion, and if the confidence of the path of the searched gene from the seed gene exceeds a threshold (mouse > 0.3, human > 0.8), it is selected. Subnetworks are grown iteratively by selecting the single gene from the functional neighborhood pool at each stage that maximizes the subnetwork activity score. For each gene in the pool, this score is calculated by adding that gene *in addition to* any genes that are included in its highest confidence path to the current subnetwork. This stage allows interesting non-differentially expressed genes to be added to the subnetwork when they bridge highly differentially expressed genes. Growth of each subnetwork is constrained by two parameters: a minimum network activity score and a minimum clustering coefficient constraint. The first restricts the subnetworks from incorporating too many low-activity genes, while the second ensures that the subnetwork remains highly clustered— genes can only be added if the subnetwork still meets both criteria as described above. Subnetwork growth is stopped when either the clustering coefficient constraint or the minimum network score constraint is not satisfied. This process is repeated for all differentially expressed genes (non-zero activity score).

For the cross-species network discovery approach, the networks are simultaneously grown in parallel. As described above, the activity score is based on the geometric mean of two or more orthologs' normalized differential expression scores, so selected orthologs are added to the respective subnetworks at each step.

2.4.6 **Orthology**

All genes for both human and mouse were mapped to Inparanoid clusters [111]. The clusters contain mapping of genes across species. For human to mouse or vice versa, the majority of ortholog mappings are one to one. However, some of the mappings are many to one, one to many, or many to many. To reduce ambiguity, during comparison, all genes were associated with their corresponding orthologous clusters. The mapping of the functional linkage network from gene space into orthologous cluster space was non-trivial as the interactions of paralogs, genes from the same species in the same orthologous cluster, had to be merged. For a cluster with multiple orthologs, the average of all genes' interactions was assigned as the cluster interaction. For this process, the lack of an edge in the functional linkage network was considered to be a zero weight edge. The outputs of the algorithm, the discovered subnetworks, are reported in the orthologous cluster space.

2.4.7 **Randomization**

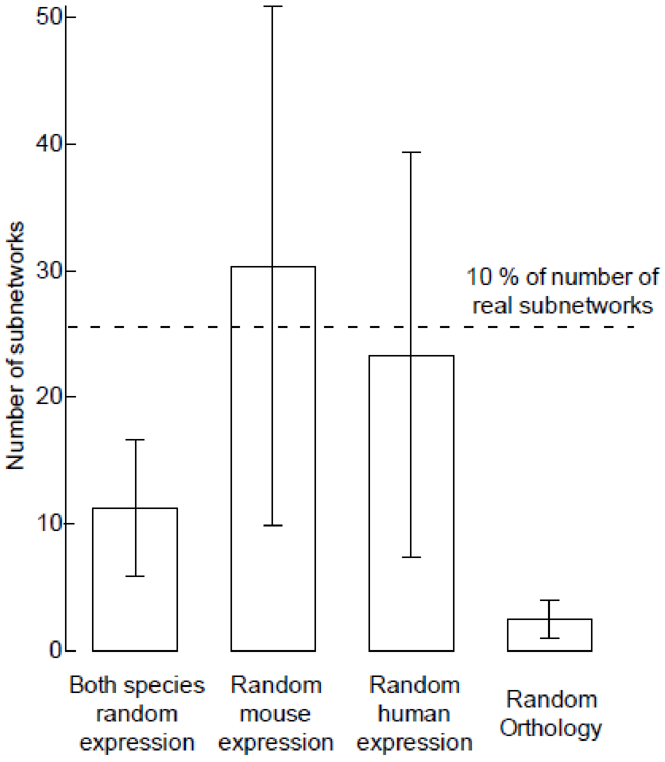


Figure 2.14 Subnetwork evaluation based on alternative randomization schemes.
In addition to the randomization scheme described in the Results section, which involves shuffling the differential expression values in both species, we evaluated three other schemes as well: randomizing differential expression values in only mouse, randomizing differential expression values in only human, and randomizing the orthology links between mouse and human. The figure plots the average number of subnetworks discovered across 5 random instances for each scheme with the dotted line providing a reference corresponding to 10% of the subnetworks identified on the real data. At the same parameters at which we discover 255 real subnetworks

(clustering coefficient parameters > 0.1 and > 0.2 for mouse and human, respectively and network score > 0.15), we found an average of ~11 with our original randomization approach, an average of ~30 with the mouse-only randomization, an average of ~24 with the human-only randomization, and an average of ~3 with the orthology randomization. Even by the most conservative randomization scheme, our approach finds ~10- fold more real networks than random.

To estimate the significance of the obtained subnetworks, randomization experiments were carried out. For both species, the differential expression values were shuffled independently relative to the gene names to remove any connection between them. Fold change values were only shuffled among genes present in the functional linkage network, while the functional linkage network was kept the same. The network discovery algorithm was then run on the shuffled expression data to discover any conserved subnetworks. This entire process was repeated several times to establish a mean and standard deviation for the number of conserved subnetworks identified by chance, which was used to assign confidence values for the real subnetworks. Alternative randomizations schemes provided similar results, and they are described in more detail in Appendix 1 (Appendix 1, Note 5, “Other Randomizations”, and Figure 2.14).

2.4.8 **Functional coverage of the subnetworks**

Gene Ontology [45] enrichment analysis was conducted on each of the subnetworks discovered by our approach using terms from the “biological process” ontology. Significance was assessed using the hypergeometric distribution was used to assess significance [112] and terms with a p-value of less than 0.05 after Bonferroni multiple hypothesis correction were deemed significant. The GO term enrichment analysis results are summarized as a hierarchically clustered matrix with subnetworks as columns and GO terms as rows, where colored elements represent significant enrichment (Figure 2.6). To distinguish monochromatic subnetworks active in stem cells from the subnetworks active in differentiated cells, we colored the subnetworks green and red, respectively. If the number of genes up-regulated in stem cells is more than twice the number of genes up-regulated in differentiated cells, then the subnetwork is considered active in stem cells and the column corresponding to the subnetwork in the functional matrix is colored green. On the other hand, if the number of genes in the subnetwork up-regulated in differentiated cells is more than twice the number of the genes up-regulated in stem cells, then the subnetwork is active in differentiated cells and is colored red. All the other cases where neither the gene up-regulated in stem cell nor the gene up-regulated in differentiated clearly dominates, the subnetworks are colored yellow in the functional matrix.

Box 1: Pseudocode for neXus algorithm

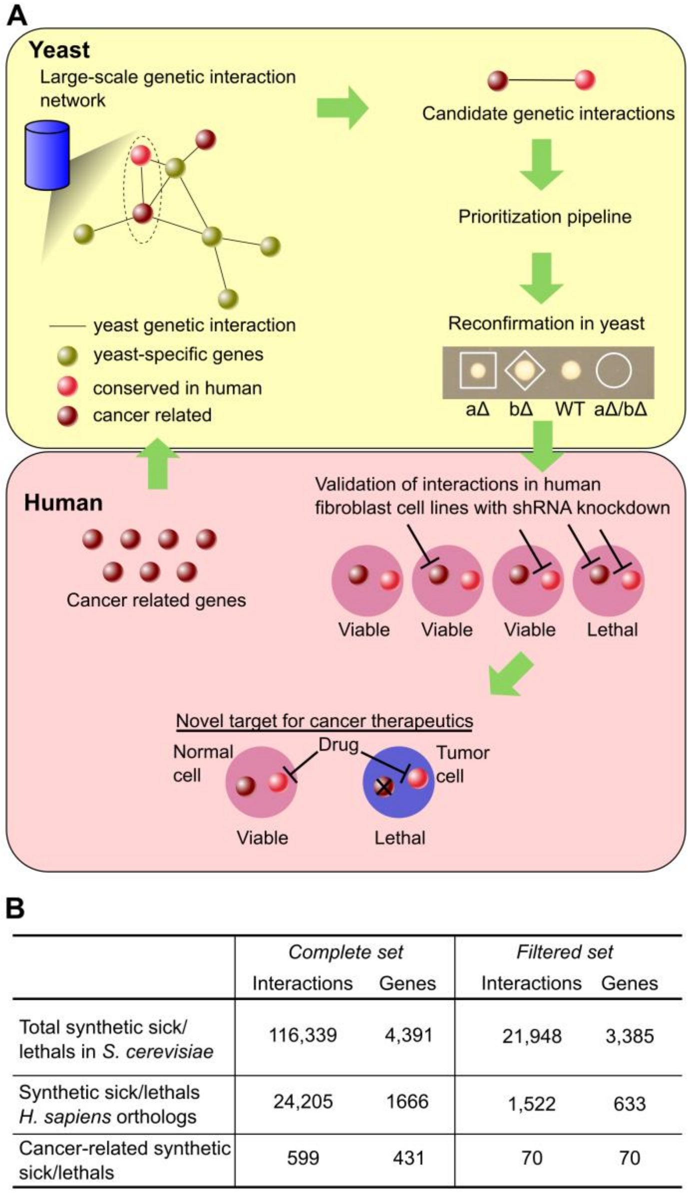
```
# assuming global mouseDifferentialGenes, humanDifferentialGenes, mouseFN,
humanFN
function subnetworks()
for seed $\in$ mouseDifferentialGenes $\cap$ humanDifferentialGenes
    mouseGenesInConsideration
        =DepthFirstSearch++(seed,mouseFN)
    humanGenesInConsideration
        =DepthFirstSearch++(seed,humanFN)
    genesInConsideration = mouseGenesInConsideration $\cap$
            humanGenesInConsideration
    growingSubnetwork = [seed] # list with single gene
    while growingSubnetwork can be grown
        addBestGene(growingSubnetwork, genesInConsideration)
        store subnetwork
   return stored subnetworks

function DepthFirstSearch++(gene, seed, functionalNetwork, threshold)
    for gene $\in$ functionalNetwork
        if $\exists$ path between gene and seed in the functionalNetwork, such
        that the product of edge weights in the path exceed threshold, then
        include the gene. Also store the best path.
   return included genes

function addBestGene(growingSubnetwork, genesInConsideration)
    return gene in genesInConsideration \ growingSubnetwork such that
    score(growingSubnetwork + gene) is the maximum

function score(subnetwork)
    if clustering coefficient of subgraphs of subnetwork in mouseFN and
    humanFN is not within constraints
        return 0
    return average of score(gene) of all genes in subnetwork

function score(gene) # the scoring is simple foldchange[gene] for single species
    experiment
    return sign(mousefoldchange[gene]*humanfoldchange[gene])* sqrt(
    mousefoldchange[gene]*humanfoldchange[gene] )
```

Chapter 3: A comparative genomic approach for identifying synthetic lethal interactions in human cancer

3.1 Overview

Synthetic lethal interactions enable a novel approach for discovering specific genetic vulnerabilities in cancer cells that can be exploited for the development of therapeutics. Despite successes in model organisms such as yeast, discovering synthetic lethal interactions on a large scale in human cells remains a significant challenge. We describe a comparative genomic strategy for identifying cancer relevant synthetic lethal interactions whereby candidate interactions are prioritized based on genetic interaction data available in yeast, followed by targeted testing of candidate interactions in human cell lines. As a proof of principle, we describe two novel synthetic lethal interactions in human cells discovered by this approach, one between the tumor suppressor gene *SMARCB1* and *PSMA4*, and another between alveolar soft-part sarcoma-associated *ASPSCR1* and *PSMC2*. These results suggest therapeutic targets for cancers harboring mutations in *SMARCB1* or *ASPSCR1*, and highlight the potential of a targeted, cross-species strategy for identifying synthetic lethal interactions relevant to human cancer.

This work has been submitted to Cancer Research journal and is currently in review process. Michael Asiedu, Mitchell Klebig, Shari Sutor, Elena Kuzmin, Justin Nelson, Jeff Piotrowski, Seung Ho Shin, Minoru Yoshida, Michael Costanzo, Charles Boone, Dennis A. Wigle and Chad L. Myers contributed to this work. Michael Asiedu, Mitch and Shari conducted the human cell line experiments. Elena conducted the yeast tetrad experiments to confirm the yeast interactions. Michael, Dennis and Chad supervised the project.


A
Yeast
Large-scale genetic interaction network
Candidate genetic interactions
Prioritization pipeline
Reconfirmation in yeast
yeast genetic interaction
yeast-specific genes
conserved in human
cancer related
aΔ
bΔ
WT
aΔ/bΔ
Human
Cancer related genes
Validation of interactions in human fibroblast cell lines with shRNA knockdown
Viable
Viable
Viable
Lethal
Novel target for cancer therapeutics
Normal cell
Drug
Tumor cell
Viable
Lethal
B

	Complete set		*Filtered set*	
	Interactions	Genes	Interactions	Genes
Total synthetic sick/ lethals in *S. cerevisiae*	116,339	4,391	21,948	3,385
Synthetic sick/lethals *H. sapiens* orthologs	24,205	1666	1,522	633
Cancer-related synthetic sick/lethals	599	431	70	70

Figure 3.1 Comparative genomic approach for discovering cancer related synthetic sick/lethal interactions in human.

(A) Flowchart describing steps to use the wealth of synthetic sick/lethal interactions available in yeast and knowledge of genes commonly mutated in cancer (Sanger Institute Cancer Gene Census) for discovery of novel cancer drug targets in human. (B) Summary of yeast synthetic sick/lethal interaction network statistics and mapping of interactions between human orthologs. The "Complete set" contains all significant synthetic sick or lethal interaction pairs at an intermediate confidence cutoff as described in [10] ($\varepsilon < -0.08$; p-value < 0.05), and human totals include any genes with human orthologs. The "Filtered set" contains only high confidence interactions ($\varepsilon < -0.2$; p-value < 0.05) or interactions replicated in two independent experiments, and human totals include only gene pairs with one-to-one orthologs (see Methods – Processing yeast genetic interaction data).

3.2 Background

Synthetic lethality is an exciting new avenue to disrupt cancer cells for targeted treatment. Two genes are said to be synthetic lethal if mutations in both genes cause cell death but a mutation in either of them alone is not lethal. In applying synthetic lethality to the discovery of cancer drugs, the goal would be to identify a target gene that when mutated or chemically inhibited, kills cells that harbor a specific cancer-related alteration, but spares otherwise identical cells lacking the cancer-related alteration [113]. This concept has recently been exploited in the development of *PARP* inhibitors as novel chemotherapeutics for breast cancer. While *PARP* is not an essential gene in normal cells, *BRCA* mutant cells are dependent on *PARP* for their survival. The described efficacy of an oral *PARP* inhibitor, olaparib (AZD2281), in early phase clinical trials for treating *BRCA* mutant tumors, is a remarkable success story for translational cancer therapeutics [114]. Importantly, strategies based on synthetic lethal interactions enable drug targeting of cancer-specific alterations in tumor suppressors which might otherwise be undruggable. Several recent studies have reported large-scale assays based on RNA-interference (RNAi) technology to discover synthetic lethal interactions with common cancer mutations, including *BRCA1/2* and *RAS* genes [15,115-120]. These studies typically target cells with a well-defined genetic background using a library of short hairpin RNAs (shRNAs) to identify combinations that result in cell death or growth inhibition. While such approaches have the potential to rapidly discover genetic interactions at a full genome scale, a number of technological challenges remain to be solved, and the number of independently validated interactions produced by these efforts has been relatively limited to date [121].

One complementary strategy to whole-genome screens in cancer cell lines is motivated by the wealth of potentially relevant interaction data in model organisms. Publication of the first eukaryotic genome-scale genetic interaction map in yeast (*Saccharomyces cerevisiae*) [10], where approximately 30% of all possible gene pairs were tested for interactions, provides a unique opportunity for discovering potentially therapeutic synthetic lethal interactions. For example, putative synthetic lethal interactions in human could be inferred based on yeast synthetic lethal interactions between conserved genes in yeast and human. These predicted pairs of human genes provide a rich database of possible candidates for further study in the context of human disease. In fact, several interactions related to chromosome stability have already been mapped from yeast to worm to human [122], suggesting that such a strategy has the potential to yield promising new drug targets. In combination with the exponentially accumulating volume of data regarding the landscape of genomic alterations in human cancer, such an approach has the potential to become increasingly powerful going forward.

We describe a combined computational and experimental approach whereby yeast interactions between human orthologs are filtered by cancer association and interaction strength in yeast, and candidates from the prioritized list are then validated in human cell lines. Using this approach, we discovered two previously unknown synthetic sick interactions, one between *SMARCB1* (yeast *SNF5*) and *PSMA4* (yeast *PRE9*), and another between *ASPSCR1* (yeast *UBX4*) and *PSMC2* (yeast *RPT1*). The predicted synthetic sick/lethal interactions between these genes were validated with shRNA double knock-down in multiple cell lines and single knock-down of *PSMA4* in two cancer cell lines containing endogenous *SMARCB1* mutations. These interactions suggest potentially new therapeutic targets for *SMARCB1* and *ASPSCR1* mutated cancers, and more broadly, illustrate the potential of this cross-species approach.

3.3 Materials and Methods

3.3.1.1 Cell culture and shRNAs

IMR90, 293TN, A-204, G-401 and 293 cell lines were obtained from American Type Culture Collection (ATCC). IMR90, 293TN and 293 cells were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% Fetal Bovine Serum (FBS), Penicillin, and Streptomycin. A-204 and G-401 cells were cultured in McCoy's 5A media containing 10% Fetal Bovine Serum (FBS), Penicillin, and Streptomycin. Bacterial stocks of control and validated gene-specific shRNA expressing vectors including *PSMA4* and *SMARCB1* shRNAs were selected from the RNAi consortium database and purchased from Sigma-Aldrich (St Louis, MO).

3.3.1.2 Preparation of viral particles

Bacterial stocks of validated shRNAs clones were amplified and DNA extracted using the HiSpeed Plasmid purification kit (Qiagen, Valencia, CA). 293TN cells were then transfected with shRNA vector clones mixed with viral package vectors pMD2 and psPAX2 using Lipofectamine 2000 transfection reagent (Invitrogen). After 48 hours, culture media containing viral particles were mixed with polybrene and centrifuged at 10,000 rpm to precipitate and concentrate the viral particles.

3.3.1.3 RNAi mediated gene knockdown

IMR90 cells were seeded into 96-well plates and transduced with pre-determined pairs of shRNAs to generate four conditions with 6 replicates each: control shRNAs, control shRNA/*PSMA4*, control shRNA/*SMARCB1* and *PSMA4*/*SMARCB1*. A similar set-up of four conditions were used for with other pairs of interactions tested. Cells from each treatment were cultured for 8 and 10 days, and the number of viable cells determined by the CellTiter-Glo Luminescence Cell Viability Assay (Promega, Madison, WI). This assay determines the number of viable cells in culture based on the amount of ATP produced by the living cells and is designed for use with multiwell plate formats and high-throughput screening (HTS) for cell proliferation and cytotoxicity assays. The addition of the assay reagent results in cell lysis and generation of a luminescent signal proportional to the amount of ATP present, which is directly proportional to the number of living cells present in each well. The intensity of the luminescent signal was measured in relative luminescence units (RLU) using the Beckman Coulter DTX 880 multimode plate reader.

3.3.1.4 Immunoblotting

Cell lysates from 293, A-204, G-401 and IMR90 control cells, and from shRNA infected cells were extracted after 5 days incubation and quantified using Bio-Rad Protein Assay reagent. An equal amount of protein (50ug) was subjected to SDS-PAGE, transferred onto a PVDF membrane and blocked with non-fat milk. The membranes were then incubated in primary antibody overnight at 4ºC and then with anti-mouse (1:6000) or anti-rabbit (1:1000) secondary antibody at room temperature for 1 hr. Primary antibodies rabbit anti-SNF5 (SMARCB1) and rabbit anti-TUG (ASPSCR1) were purchased from Cell Signaling whereas rabbit anti-PSMC2, 20S Proteosome α-4 (PSMA4) and anti-β-actin-HRP were obtained from Santa Cruz biotechnology. Protein expression was detected using enhanced chemoluminiscence (ECL) substrate (Pierce).

3.3.1.5 Estimation of the significance of genetic interactions in human shRNA experiments

Growth rate of a single or double shRNA knockdown relative to the empty shRNA vector control were calculated using

$$f_A = \frac{RLUCount_A}{RLUCount_{Control}}$$

Where f_A is relative growth rate for a single or double knock down experiment (A, B or AB), and $RLUCount_A$ is the intensity of the luminescent signal measured in relative luminescence units (RLU). Since f_A is a ratio of two quantities that has error associated with it, error for f_A is given by

$$\sigma_A = \sqrt{\left(\frac{\Delta RLUCount_A}{RLUCount_A}\right)^2 + \left(\frac{\Delta RLUCount_{Control}}{RLUCount_{Control}}\right)^2} * f_A$$

where $\Delta RLUCount_A$ is the standard deviation in the n (= 6 for our experiment) observations of $RLUCount_A$.

Expected double mutant fitness and error associated with it is given by

$$f'_{AB} = f_A * f_B$$

$$\sigma'_{AB} = \sqrt{\left(\frac{\sigma_A}{f_A}\right)^2 + \left(\frac{\sigma_B}{f_B}\right)^2} * f'_{AB}$$

In order to assess the significance of the interaction, we assumed a normal distribution for f'_{AB} and f_{AB}, and compared f'_{AB} with f_{AB} using Welch t-test[123]. The significance of the difference between f_{AB} and f'_{AB} can be calculated using one tailed t-test, which requires the t-test score t [123] and degree of freedom ν given by the Welch-Satterthwaite equation[124], as follows:

$$t = \frac{f'_{AB} - f_{AB}}{\sqrt{\frac{\sigma'^{\,2}_{AB}}{n'_{AB}} + \frac{\sigma^2_{AB}}{n_{AB}}}}$$

$$\nu = \frac{\left(\frac{\sigma'^{\,2}_{AB}}{n'_{AB}} + \frac{\sigma^2_{AB}}{n_{AB}}\right)^2}{\frac{\sigma'^{\,4}_{AB}}{n'^{\,2}_{AB}(n'_{AB}-1)} + \frac{\sigma^4_{AB}}{n^2_{AB}(n_{AB}-1)}}$$

Here $n_{AB} = n'_{AB} = 6$ because we have 6 replicate observations for control, single and double knock-down experiments.

3.3.1.6 Orthology mapping

InParanoid7 [37] was used to map yeast genes to human genes. Only 1:1 orthologs were used for our study (Supplementary Table S1).

3.3.1.7 Collection and processing of yeast genetic interaction data

Yeast genetic interaction data was taken from Costanzo *et al*. 2010 [10], which reported data for interactions between 1711 query genes and 3885 array genes. We applied a p-value cutoff < 0.05 on all interactions. Furthermore, we applied an interaction cutoff in two ways: first, we considered stringent negative genetic interactions (□ < -0.2), and second, we allowed intermediate interactions (□ < -0.08), which were reported in reciprocal screens. Specifically, in the Costanzo *et al*. network, query genes were screened against the entire non-essential deletion array, and in some cases, genes present on the array were also screened as queries. For these cases, an interaction between genes A and B was tested in both screens: A (query) x B (array) and B (query) x A (array). In such cases, we applied an intermediate cutoff because an interaction appearing in both of these screens is of high confidence.

11 new SGA screens were also used to generate candidate gene pairs, including screens for the following queries (human/yeast orthologs): *XPC/RAD4, VTI1A/VTI1, NOP56/NOP56, POLD2/POL31, MLH1/MLH1, XPO1/CRM1, UBA3/UBA3, ERCC4/RAD1, XPA/RAD14, PSMC2/RPT1, PSMB1/PRE7*. A screen involving a temperature sensitive (TS) allele of yeast *RPT1* (human *PSMC2*) was the basis for testing the human interaction *PSMC2-ASPSCR1*, so the yeast interaction data supporting that inference are included here (Supplementary Table 2). A genome-wide screen for the *RPT1* TS allele's genetic interactions was conducted as described in Baryshnikova et al, 2010 [125]. Briefly, a rpt1-1 mutant strain marked with a nourseothricin (NatMX4) resistance cassette and harboring the SGA haploid specific markers and reporter [125] was mated to an array of ~4000 viable *S. cerevisiae* deletion mutants. Nourseothricin- and geneticin-resistant heterozygous diploid mutants were selected and sporulated and MATa rpt1-1 double mutants were subsequently selected [125]. To confirm the SGA results, all gene deletions were constructed in a SSL204 MATa strain and crossed with an

isogenic rpt1-1 MATα strain. Diploid cells were sporulated at 25°C and dissected. Plates were incubated for 3-5 days at either 25°C or 30°C.

3.3.1.8 Yeast tetrad dissection

Confirmations by tetrad analyses were performed as described in Amberg *et al* 2006 [126].

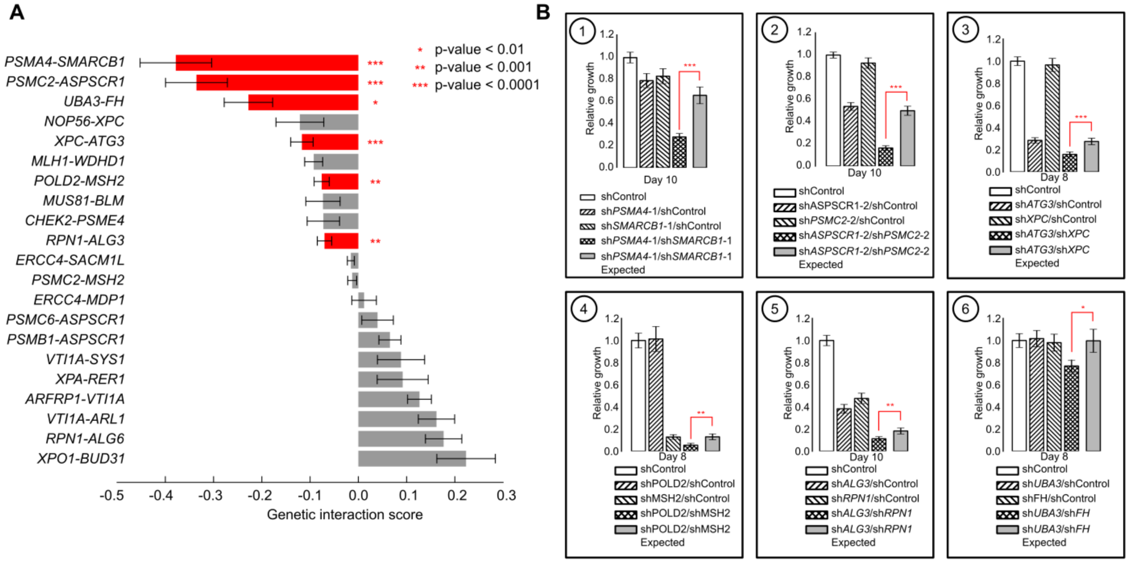


Figure 3.2 Interaction testing of 21 selected candidate synthetic sick/lethal interactions in human fibroblast cell lines.
(A) The interaction scores for all human interactions tested. The interaction score is the difference between observed and expected growth rate based on a multiplicative model. The significant negative genetic interactions are colored red and strength of the significance is denoted by the number of asterisks, according to the legend shown. (B) Results for each significant interaction tested. The number of days the fibroblast cells were grown in the presence of shRNAs is indicated in each plot. The error bars for both (A) and (B) represent twice the width of the standard error in the interaction scores and growth rates.

3.4 Results

To discover cancer-associated genetic interactions in human cells, we first selected a set of highly significant interactions between yeast genes from the large network of synthetic genetic interactions that has recently been mapped in yeast. A recent study reported testing genetic interactions for 5.4 million yeast gene pairs, consisting of instances where two non-essential genes were deleted in combination, or a temperature-sensitive mutation of an essential gene was used along with a deletion of a non-essential gene [10]. In total, approximately 116,000 pairs were reported as having a detectable synthetic sick or lethal interaction, of which around 24,000 interactions connect two genes that both have human orthologs (Figure 3.1). More than 500 of these latter interactions involve at least one gene that has been previously associated with mutations in cancer (Sanger Institute Cancer Gene Census [127]; Figure 3.1B), suggesting a large number of candidate pairs can be generated by this approach (Supplementary Table S3).

To narrow the candidate list for testing in human cells, we first applied a very stringent cutoff on interactions in yeast, either requiring a high-magnitude effect, high-confidence interaction to be reported ($\varepsilon < -0.2$, $p < 0.05$) or selecting gene-pairs for which interactions were reproduced in two reciprocal screens (see Methods for details). Furthermore, we restricted our search to genes with one-to-one orthologs in human to increase the likelihood of functional conservation between yeast and human, and to avoid potentially buffering effects of paralog functional redundancy [128]. Applying these relatively stringent criteria, we obtained 1522 putative synthetic sick/lethal interactions between human orthologs of yeast genes, of which 70 interactions involved a gene that has been previously implicated in some form of human cancer (Figure 3.1B). In addition to these published interactions, we applied the same criteria to 11 previously unpublished yeast screens involving human orthologs (see Methods and Supplementary Table S2). Candidate interaction pairs involving cancer-associated mutations (Sanger Institute Cancer Gene Census) were ranked based on the strength of the yeast interactions and were selected in order up to a maximum of 3 interactions per gene. In total, 21 pairs of genes representing mutations associated with a diverse set of cancers were selected for further experiments in human cell lines (Figure 3.2).

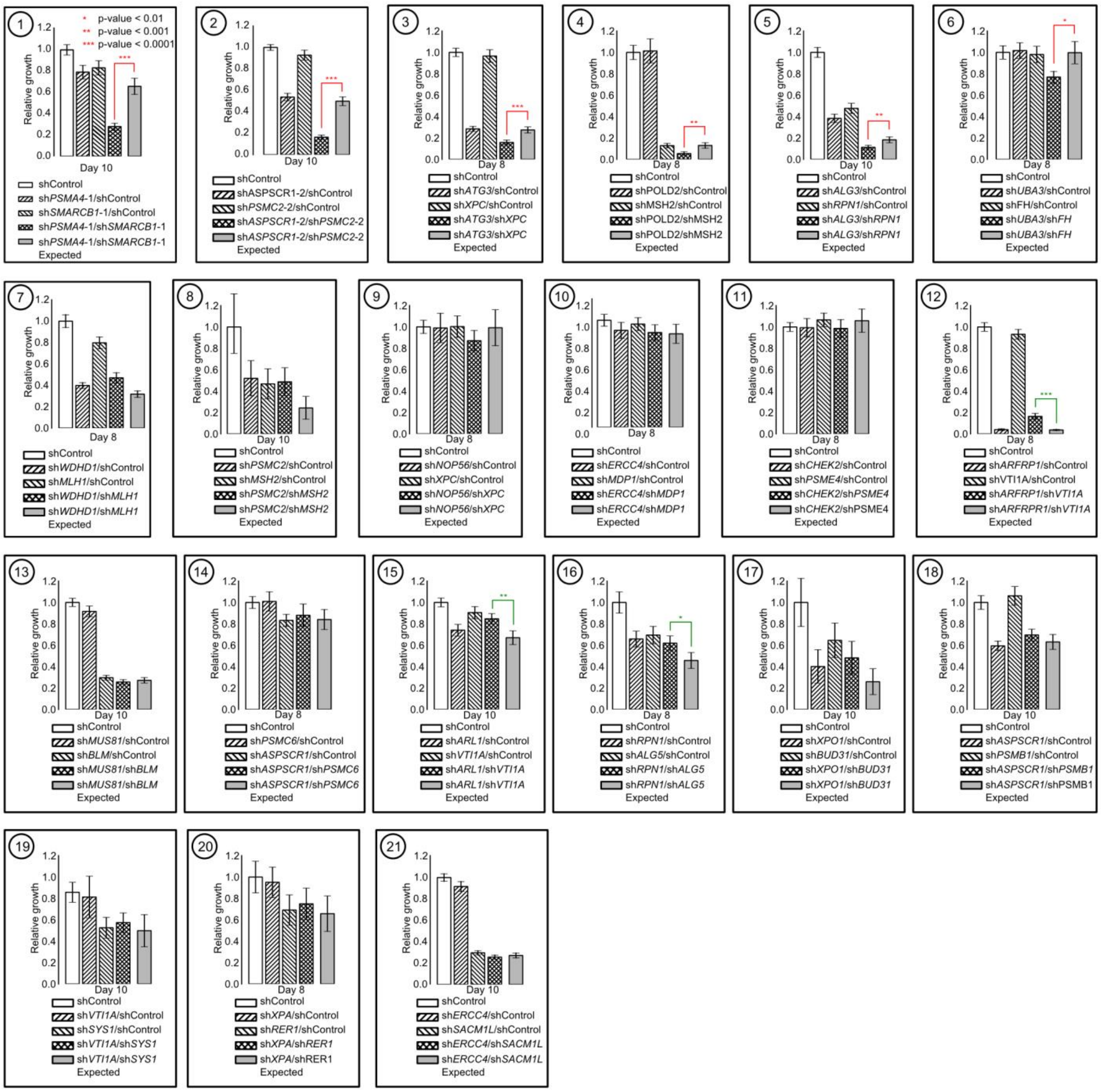


Figure 3.3 Genetic interaction test results for all interactions tested for this study.
Each box contains the growth rate for single and double knock-downs for genes in the interaction we tested. Further, the double knock-down prediction from single knock-downs is also a bar in the plot. Statistically significant differences are indicated with 1, 2 or 3 asterisks: more asterisks refer to more significant synthetic sick/lethal interaction, as explained in the legend shown in panel. The red and green color of the asterisks refer to negative and positive genetic interactions respectively. The number of days the IMR90 fibroblast cells were grown in the presence of the shRNAs is indicated in each plot.

The candidate synthetic sick or lethal pairs derived from the yeast genetic interaction network were screened in normal human IMR90 fibroblast cells using an RNAi approach. IMR90 cells were chosen because the cell line was established from the lungs of a 16-week female fetus and have the advantage of early passage and a low likelihood of accumulated genetic alterations. This stable genetic background allowed us to assess the validity of candidate interactions with the lowest possibility of unknown, confounding genetic alterations. We screened the selected 21 pairs of potential interactions using a CellTiter-Glo luminescence viability assay. We found evidence for significant synthetic sick or lethal interactions for 6 of the 21 tested pairs (see Figure 3.3 for data, Figure 3.2 for significant interactions). We focused further validation efforts on the strongest 2 of the 6 significant interactions: *SMARCB1*/*PSMA4* and *ASPSCR1*/*PSMC2* (Panels 1 and 2 for Figure 3.2B).

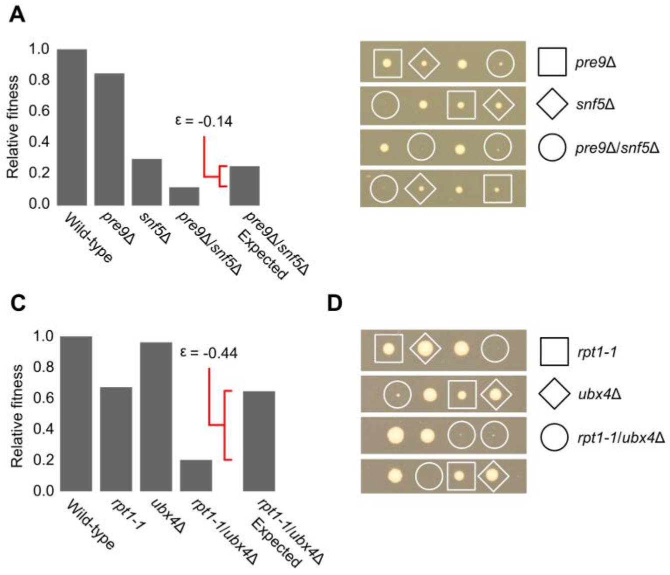


Figure 3.4 Yeast interaction confirmation
(A) Fitnesses of the single and double mutants relative to wild-type for the *SNF5-PRE9* interaction. The interaction score (ε) was estimated by comparing the observed double mutant fitness with the fitness expected based on the single mutant fitnesses. (B) Confirmation of the synthetic sick/lethal interaction using tetrad dissection analysis for the *SNF5-PRE9* double mutant. Each tetrad is oriented horizontally and represents four

meiotic progeny of a heterozygous double mutant between *pre9*Δ::natMX4/*PRE9* and *snf5*Δ::kanMX4/*SNF5*. Four representative tetrads are shown. The genes knocked-out are identified by the presence of the natMX and kanMX markers, respectively. The identified double knock-out spore colonies are enclosed in circles while single gene knock-out strains are enclosed in squares or diamonds, and wild type strains are not enclosed. (C) and (D) present similar data for query mutant *rpt1-1*, a temperature-sensitive conditional mutant of *RPT1*, and

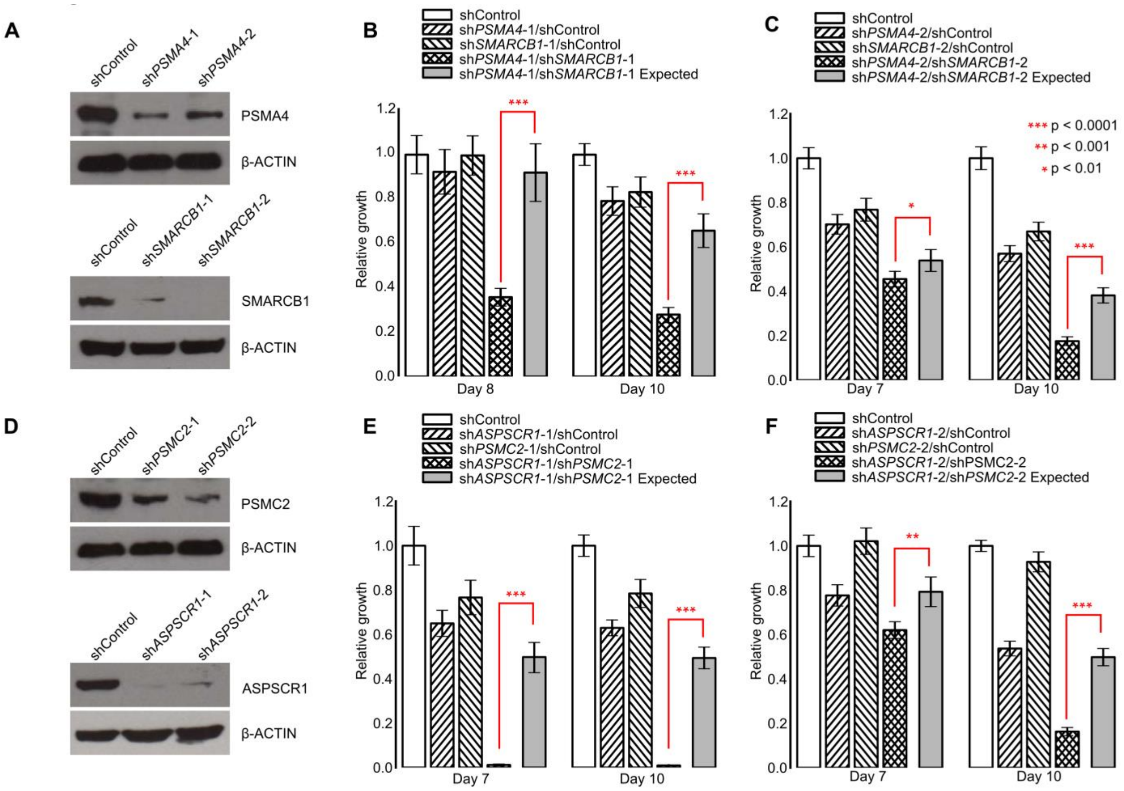


Figure 3.5 Validation of two candidate synthetic sick/lethal interactions in human fibroblast cell lines.

(A) A Western blot of IMR90 cells transduced with PSMA4 and SMARCB1 shRNA virus showing down-regulation of respective protein expression. (B and C) Cell viability analyses of PSMA4 and SMARCB1 interaction using two different clones which showed decreased cell survival in cells depleted of both PSMA4 and SMARCB1 compared to cells expressing shRNAs of individual genes and compared to the expected effect from depletion of both genes. (D) Immunoblotting showing knockdown of PSMC2 and ASPSCR1 expression in IMR90 cells treated with viral particles encoding PSMC2 or ASPSCR1 shRNA. (E,F) Synthetic lethal interaction effect of PSMC2 and ASPSCR1 in IMR90 cells as shown by significantly decreased survival in cells expressing shRNAs of both genes compared to the expected effect from depletion of both genes.

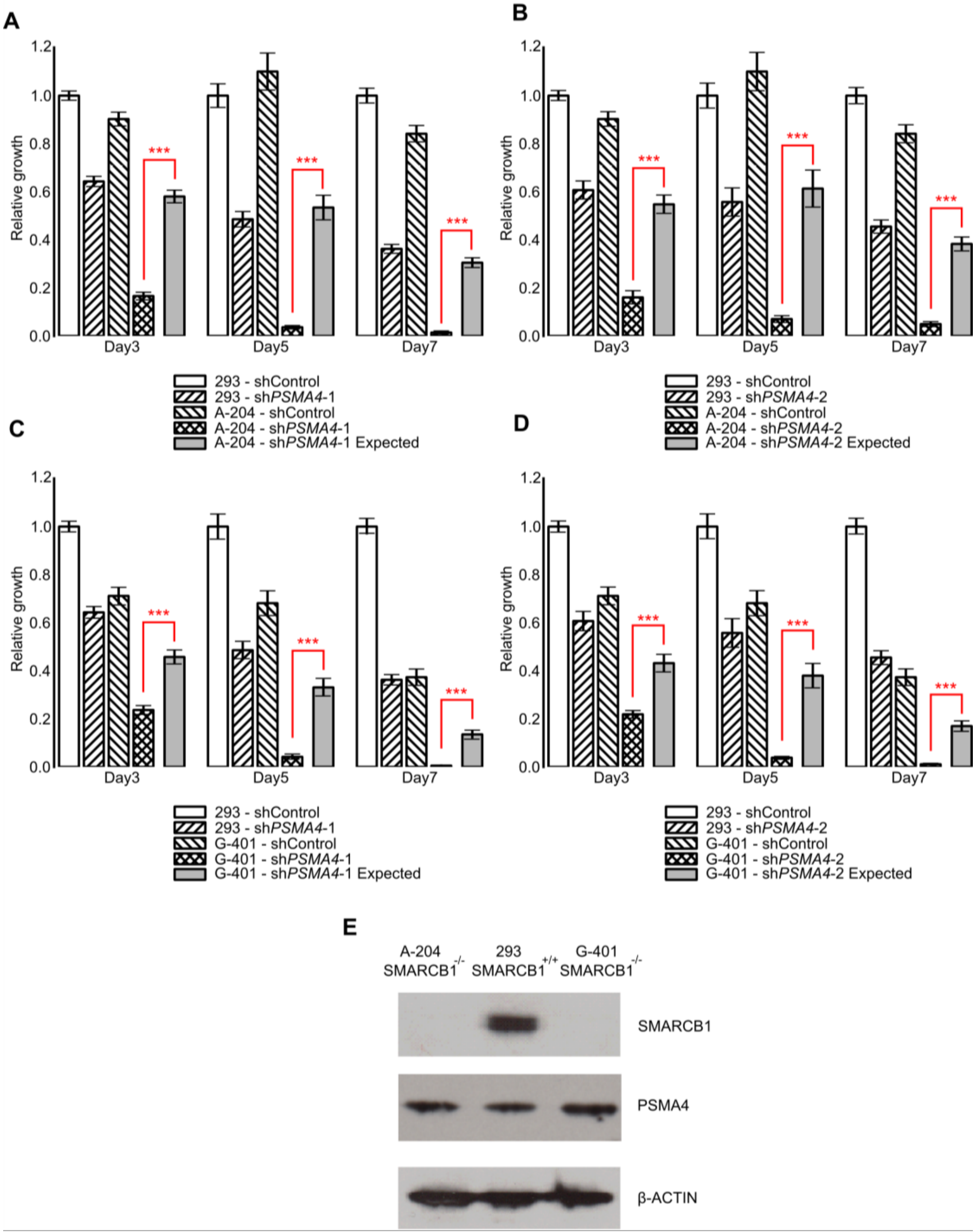


Figure 3.6 Validation of the PSMA4-SMARCB1 synthetic lethal interaction in cancer cell lines harboring SMARCB1 loss-of-function mutations.

(A) Cell viability analyses of cell lines with (A-204) or without (293) endogenous SMARCB1 mutation, grown with or without PSMA4 shRNA knock-down (shPSMA4-1), demonstrate the therapeutic potential for this cancer associated synthetic lethal interaction. (B) The experiment in (A) is repeated with a different *PSMA4* shRNA construct (shPSMA4-2). (C,D) The experiment in A,B is repeated with a different SMARCB1 deficient cell line, G-401. (E) A Western blot showing the complete absence of SMARCB1 protein in cell lines, A-204 and G-401, which have endogenous null SMARCB1 mutations, and normal expression of SMARCB1 in the control 293 cell line. The endogenous PSMA4 and β-actin protein levels detected serve as loading controls.

To further validate these two interactions, we first retested them in yeast cells by dissecting tetrads (Figure 3.4), which indeed confirmed a strong synthetic sick effect between the pairs of yeast orthologs, *SNF5* / *PRE9* (human *SMARCB1*/*PSMA4*) and *UBX4* / *RPT1* (human *ASPSCR1*/*PSMC2*) (Figure 3.4). In human cells, we repeated the same viability assay and additionally performed knock-downs with independent targeting shRNAs for both pairs of genes. After simultaneous depletion of the targeted gene pairs, the number of cells that survived was significantly reduced in all cases (Figure 3.5 B, C, E, F). Importantly, the extent of survival was significantly lower than the expected survival of double knock-downs estimated from the single shRNA effects (Figure 3.5B of *PSMA4*/*SMARCB1*; Welch t-test [123] score = 8.11 at Day 8, 8.90 at Day 10; Figure 3.5E for ASPSCR1-PSMC2; Welch t-test score = 14.86 at Day 7 and 20.95 at Day 10; pval < 0.0001 in all cases). Expected double knock-down effects were calculated assuming a multiplicative null model, which has been widely used in the genetic interaction community [7] (see Methods for details). Similar results were observed when different shRNA clones for *PSMA4/SMARCB1* and *ASPSCR1/PSMC2* knock-down were used (Figure 3.5 C,F). We also confirmed the effectiveness of shRNA silencing of the targeted genes by conducting protein expression analyses using Western blots (Figure 3.5 A, D), which showed greatly reduced protein levels in the shRNA-infected cells.

The discovery of cancer related synthetic lethal interactions can directly impact therapeutic potential, as the synthetic lethal interactor of a cancer related gene can be targeted selectively to kill cancer cells. To test the clinical relevance of the *PSMA4* and

SMARCB1 interactions, we identified an epithelial muscle rhabdosarcoma cell line (A-204) and a renal rhabdoid sarcoma cell line (G-401), each harboring *SMARCB1* mutations, and used embryonic kidney HEK-293 cells expressing wild type *SMARCB1* as a control (Figure 3.6). We observed that *PSMA4* knock-down almost completely kills the cell lines harboring *SMARCB1* mutations, and that this observation was exaggerated versus controls when following the cells to later time points (Day 7, Figure 3.6A-D). We also demonstrated the complete absence of SMARCB1 protein in cell lines A-204 and G-401 by Western blotting (Figure 3.6E). A-204 carries a TC deletion of codons 181 and 182 in exon 5 whereas G-401 harbors a homozygous deletion of exons 1-9 [129] In both cell lines harboring *SMARCB1* mutation, the decrease in growth is greater than expected by the multiplicative combination of the individual *SMARCB1* mutation and *PSMA4* knock-down effects as estimated from the control cell line (Figure 3.6; p value < $2.5*10^{-6}$ for all days, all replicates and both cell lines).

3.5 Discussion

We describe an experimental pipeline where we prioritized synthetic genetic interactions from the global map of yeast interactions to test candidate synthetic sick/lethal pairs involving cancer-associated mutations in human cells. We propose this general approach, involving computational prioritization followed by experimental validation, as a complementary strategy to large-scale RNAi screens that are in progress by several other groups.

Based on the synthetic sick/lethal interaction we discovered between *SMARCB1* and *PSMA4*, we hypothesize that targeting *PSMA4* in therapeutic approaches could selectively inhibit the growth of cancer cells harboring *SMARCB1* mutations. Human *PSMA4* is a proteasome subunit component expressed across numerous tissues. *PSMA4* mRNA levels are increased in lung tumors compared with normal lung tissues, and down-regulation of *PSMA4* expression in lung cancer cell lines decreases proteasome activity and induces apoptosis [130]**.** Human *SMARCB1* is a core component of the BAF ATP-dependent chromatin-remodeling complex, known to play important roles in cell proliferation and differentiation, and inhibition of tumor formation. Deletions in *SMARCB1* are associated with epitheliod sarcomas [131], and are a known cause of rhabdoid tumor predisposition syndrome (RTPS), a highly malignant group of neoplasms that usually occur in early childhood [132,133]. No described direct protein interaction exists between *PSMA4* and *SMARCB1*. Although the clinical implications of this synthetic lethal interaction await further study, one potential application could be in the use of existing proteasome inhibitors such as bortezomib for the treatment of tumors harboring *SMARCB1* mutations. Interestingly, interactions between other SWI/SNF subunits and the proteasome were also observed in yeast [10], suggesting the possibility that perturbations in multiple combinations of subunits across these complexes could have the same effect. Whether interactions exist in human between other genes encoding the SWI/SNF complex and the proteasome remains to be determined, but this merits further study since mutations in other subunits of SWI/SNF have been observed in many other types of cancer [134,135].

The direct clinical implications of the *ASPSCR1-PSMC2* synthetic sick interaction are less clear, but this case also merits further study. *ASPSCR1* is a relatively uncharacterized gene that has been associated with alveolar soft-part sarcoma (ASPS), a rare class of tumors that typically occur in younger patients. [136]. Most cases of this cancer are associated with an unbalanced translocation der(17)t(X;17) (p11;q25) that results in an ASPSCR1-TFE3 fusion protein. The fusion protein appears to act as an aberrant transcription factor, inducing unregulated transcription of TFE3-regulated genes [136,137]. This fusion truncates one *ASPSCR1* allele, leaving the other allele intact in most cases [136,137]. How this genetic interaction could be leveraged for therapeutic purposes awaits further investigation, but one possibility is the potential combined effect of reduced expression of *ASPSCR1* in conjunction with proteasome inhibition.

Interestingly, the two strongest synthetic sick/lethal interactions we observed involved components of the proteasome, even though we tested a variety of genes from multiple pathways that were produced by our approach. These data suggest the proteasome may be a rich target for synthetic lethal approaches in human cancer therapy, and indeed, successful cancer treatment involving proteasome inhibitors has been reported recently in a number of different contexts [138]. The availability of several approved proteasome inhibitors may make such interactions between cancer-associated mutations and the proteasome immediately translatable to several clinical settings. These results also highlight the potential for discovering interactions within core biologic pathways with strong yeast/human homology. Importantly however, one of the limitations of our approach is its dependence on genes and proteins with such homology, and an inability to reflect many known oncogenic pathways where yeast/human homology does not exist. We also note that a recent study identified both *PSMA4* and *PSMC2* as 2 of a set of 56 genes (and the only proteasomal components) for which gene knock-down inhibited the growth of cells with partial copy number loss in the same gene [139]. Our independent finding of synthetic sick/lethal interactions for these same proteasomal subunits is intriguing and suggests that perturbations of these subunits have a relatively unique effect on proteasome function that may not be replicated by manipulation of its other components.

The appeal of a targeted approach for identifying synthetic sick/lethal interaction candidates is strengthened by the fact that there are currently large numbers of tumor genome sequencing efforts in progress which will produce new, potentially lengthy, lists of mutations associated with various types of cancers. As we gain richer knowledge of the spectrum of mutations present in cancer, we can continue to directly screen the most promising candidate synthetic sick/lethal interactions involving these genes. The identification of specific cancer subtypes harboring specific mutations may provide therapeutic opportunities for synthetic lethal approaches that are not currently appreciated. Furthermore, in future studies, we intend to leverage data beyond sequence-similarity and literature-derived functional information to prioritize interactions for testing across species. For example, the large collections of functional genomic data in both yeast and human could allow for a more robust and unbiased assessment of the likelihood of functional conservation of genes and conserved synthetic lethal interactions between them. Our initial results highlight the feasibility of this comparative genomic approach, and suggest its potential utility for rapid translation of novel sequence variants into new therapeutic targets. We believe this approach has the potential to provide a dramatic increase in the number of therapeutic targets beyond those currently available for drug development.

Chapter 4: Comparison of Profile Similarity Measures for Genetic Interaction Networks

4.1 Overview

Analysis of genetic interaction networks often involves identifying genes with similar profiles, which is typically indicative of a common function. While several profile similarity measures have been applied in this context, they have never been systematically benchmarked. We compared a diverse set of correlation measures, including measures commonly used by the genetic interaction community as well as several other candidate measures, by assessing their utility in extracting functional information from genetic interaction data. We find that the dot product, one of the simplest vector operations, outperforms most other measures over a large range of gene pairs. More generally, linear similarity measures such as the dot product, Pearson correlation or cosine similarity perform better than set overlap measures such as Jaccard coefficient. Similarity measures that involve L_2-normalization of the profiles tend to perform better for the top-most similar pairs but perform less favorably when a larger set of gene pairs is considered or when the genetic interaction data is thresholded. Such measures are also less robust to the presence of noise and batch effects in the genetic interaction data. Overall, the dot product measure performs consistently among the best measures under a variety of different conditions and genetic interaction datasets.

This work has been published in [140]. Benjamin VanderSluis helped with batch effect analyses and added a dot product similarity metric plugin to Cluster 3.0. Chad supervised the project.

4.2 Background

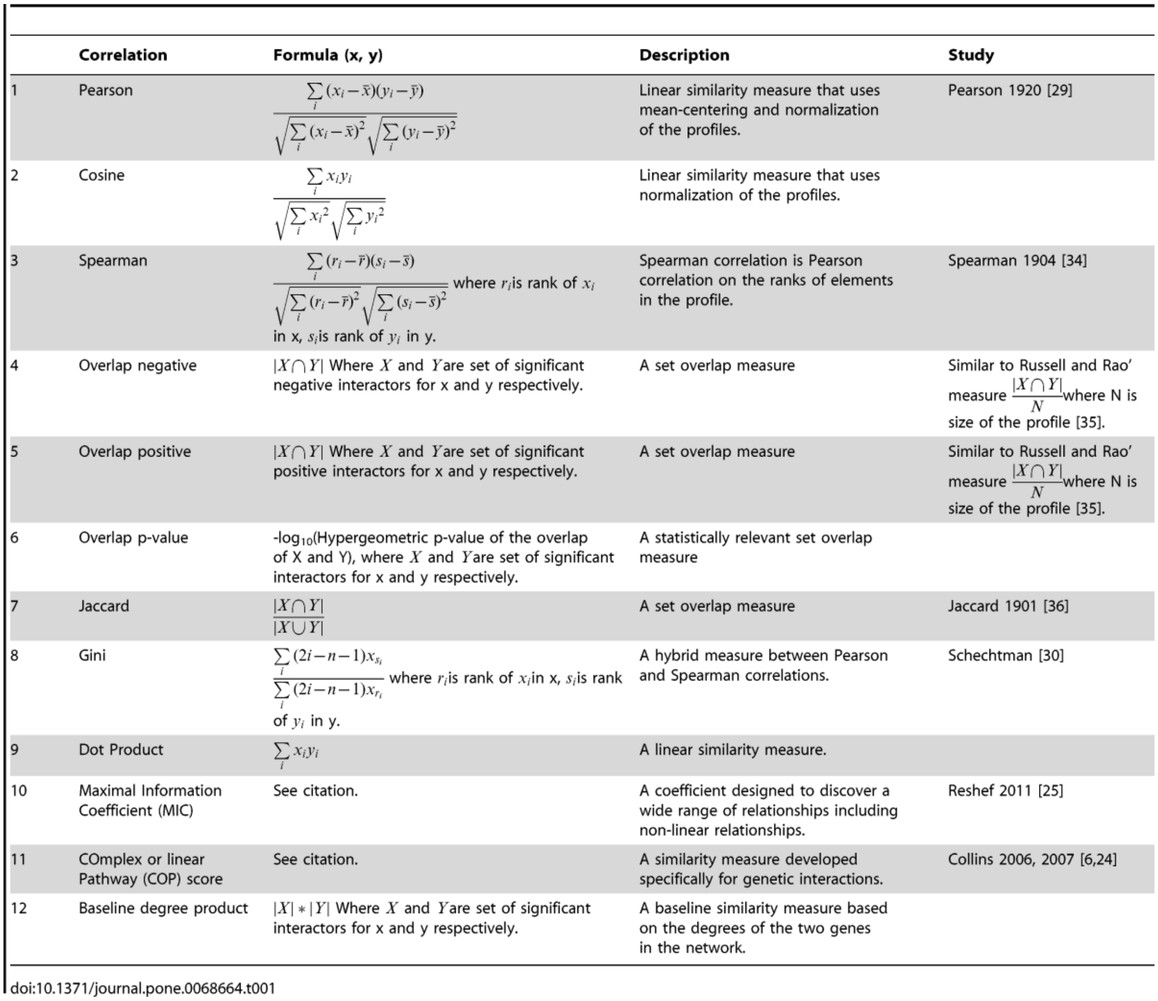
Similarity measures are among the most important operations used in analyzing genomic data. One of the most widely used analysis paradigms, guilt-by-association, requires measuring similarity between gene pairs or other objects of interest based on a high-dimensional set of features. Guilt-by-association has proven particularly important for the analysis of genetic interactions because similarity of genetic interaction neighbors of two genes is often easier to interpret than direct interactions between genes [9,10,141]. A genetic interaction is a measure of how surprising a double gene knock-out phenotype is, compared to the phenotype expected from known single gene knock-out phenotypes [7]. Using this definition, genetic interactions can be quantitatively measured [142], but often occur between genes that lack an obvious close functional relationship [9]. Because of the difficulty in understanding and interpreting individual genetic interactions, they are frequently studied in the context of other interactions in the dataset. For example, the complete list of a gene’s interactions, commonly referred to as a profile, can be compared with other profiles, and similarities between the gene profiles are indicative of functional similarity [10,142].

Understanding profile similarity measures for genetic interactions and other related datasets has become increasingly critical in the recent years because several large-scale or whole-genome genetic interaction studies have been published in several organisms including baker's yeast (*Saccharomyces cerevisiae*) [10,143], fission yeast (*Schizosaccharomyces pombe*) [11,144,145], bacteria (*Escherichia Coli)* [146], fly (*Drosophila melanogaster*)[147], worm (*Caenorhabditis elegans*)[148] and human (*Homo sapiens*)[117]. Many of these high-throughput studies use profile correlation for validation and interpretation of the generated genetic interaction data, or discovery of new gene functions. Despite its importance, no systematic study has been conducted to compare the efficacies of different profile similarity measures on genetic interaction data using biological relevance as a benchmark. Nevertheless, there are studies in other fields where correlation measures have been compared either from a theoretical standpoint [149,150] or for specific applications such as co-citation data analysis [151,152], co-localization detection in visualization [153], chemical fingerprint searches [154], microarray gene expression profile analysis [155], and cluster analysis [156,157]. Most of these applications either involve datasets that are binary in nature, such as the co-citation study, or involve datasets that consist of only positive continuous elements such as pixel scores. In contrast, genetic interaction data are continuous and signed. Other data that are continuous and signed, such as gene expression data, are fundamentally different in that genetic interaction networks tend to contain much sparser signal and the different signs in genetic interaction data have very different biological interpretations [10,142]. The unique properties of genetic interaction data warrant a separate and systematic study of the correlation measures.

In this study, we systematically benchmarked several similarity measures on two of the largest genetic interaction datasets [10,11]. Further, we investigated how measurement noise and batch effects, which commonly affect genetic interaction assays [142], impact the performance of these similarity measures. We also address the effect of thresholding continuous genetic interaction data, a common practice in this field [9,10,142,158], on the performance of the different similarity measures.

4.3 Results and Discussion

We compared a diverse set of profile similarity measures that cover a range of classes including linear similarity measures, set-overlap measures, rank based similarity measures, hybrid measures, and L_2 normalization based measures (see Table 1). We also included a similarity score specifically developed for genetic interactions (COmplex or linear Pathway, COP score;) [143,159] and a recently developed similarity score that works well for discovering several non-linear relationships (Maximal Information Coefficient, MIC) [160]. For a baseline comparison, we use the product of degrees of the two profiles [161].



	Correlation	Formula (x, y)	Description	Study
1	Pearson	$\frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2}\sqrt{\sum_i (y_i-\bar{y})^2}}$	Linear similarity measure that uses mean-centering and normalization of the profiles.	Pearson 1920 [29]
2	Cosine	$\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$	Linear similarity measure that uses normalization of the profiles.	
3	Spearman	$\frac{\sum_i (r_i-\bar{r})(s_i-\bar{s})}{\sqrt{\sum_i (r_i-\bar{r})^2}\sqrt{\sum_i (s_i-\bar{s})^2}}$ where $r_i$is rank of x_i in x, $s_i$is rank of y_i in y.	Spearman correlation is Pearson correlation on the ranks of elements in the profile.	Spearman 1904 [34]
4	Overlap negative	$\lvert X \cap Y \rvert$ Where X and $Y$are set of significant negative interactors for x and y respectively.	A set overlap measure	Similar to Russell and Rao' measure $\frac{\lvert X \cap Y \rvert}{N}$where N is size of the profile [35].
5	Overlap positive	$\lvert X \cap Y \rvert$ Where X and $Y$are set of significant positive interactors for x and y respectively.	A set overlap measure	Similar to Russell and Rao' measure $\frac{\lvert X \cap Y \rvert}{N}$where N is size of the profile [35].
6	Overlap p-value	-log$_{10}$(Hypergeometric p-value of the overlap of X and Y), where X and $Y$are set of significant interactors for x and y respectively.	A statistically relevant set overlap measure	
7	Jaccard	$\frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$	A set overlap measure	Jaccard 1901 [36]
8	Gini	$\frac{\sum_i (2i-n-1)x_{s_i}}{\sum_i (2i-n-1)x_{r_i}}$ where $r_i$is rank of $x_i$in x, $s_i$is rank of y_i in y.	A hybrid measure between Pearson and Spearman correlations.	Schechtman [30]
9	Dot Product	$\sum_i x_i y_i$	A linear similarity measure.	
10	Maximal Information Coefficient (MIC)	See citation.	A coefficient designed to discover a wide range of relationships including non-linear relationships.	Reshef 2011 [25]
11	COmplex or linear Pathway (COP) score	See citation.	A similarity measure developed specifically for genetic interactions.	Collins 2006, 2007 [6,24]
12	Baseline degree product	$\lvert X \rvert * \lvert Y \rvert$ Where X and $Y$are set of significant interactors for x and y respectively.	A baseline similarity measure based on the degrees of the two genes in the network.	

doi:10.1371/journal.pone.0068664.t001

Table 4.1 Similarity measures evaluated in this study.

We used two genetic interaction datasets from two different species and laboratories: Costanzo *et al*. from *S. cerevisiae* [10] and Ryan *et al.* [143] from *S. pombe* to evaluate these similarity measures.

Our primary basis for evaluation of the similarity measures was their ability to discover known functional relationships between genes. More specifically, each metric was used to rank a set of gene pairs based on their interaction profile similarities, and the highest-scoring gene pairs were compared with a functional co-annotation standard developed using the Gene Ontology (GO standard; see Methods: Creating GO standard)

[45,162]. The similarity measures were evaluated using precision-recall analysis, which is a continuous way of evaluating top similarities at different similarity cutoffs with the GO standard. Precision refers to the percentage of similarities at a given threshold that correctly associate functionally related gene pairs, while recall refers to the proportion of all functionally related gene pairs among the set of top similarities. Similar strategies have been used previously in evaluating other kinds of genomic data or prediction methods (e.g. see [162]).

4.4 Comparison of different correlation measures on genetic interaction data

Overall, the dot product was the most consistent top performer in terms of its precision-recall characteristics on both the *S. pombe* and *S. cerevisiae* genetic interaction datasets for both query (rows) and array (columns) genes' correlations. Pearson correlation [163], which is the most widely used profile similarity measure for genetic interactions, performed well for low recall (Figure 4.1). Cosine correlation, also known as un-centered Pearson correlation, shows precision-recall performance close to that of Pearson correlation (Figure 4.1). This similarity in performance is expected as the means of genetic interaction profiles are very close to zero, which makes the two metrics equivalent. High precision at low recall appears to be a general property of similarity measures that normalize profiles, such as Pearson, cosine, Spearman and Gini correlations. Another aspect common to these correlations is that their performance drops sharply for higher recall (Figure 4.1). In contrast, metrics that do not normalize genes' profiles, such as overlap-negative and the dot product, do not experience a similar dramatic drop in performances at high recall. Gini correlation [164], which is considered a hybrid between Pearson and Spearman correlation, performed similarly to or slightly better than Spearman but its precision was worse than Pearson correlation's precision at any choice of recall (Figure 4.1).

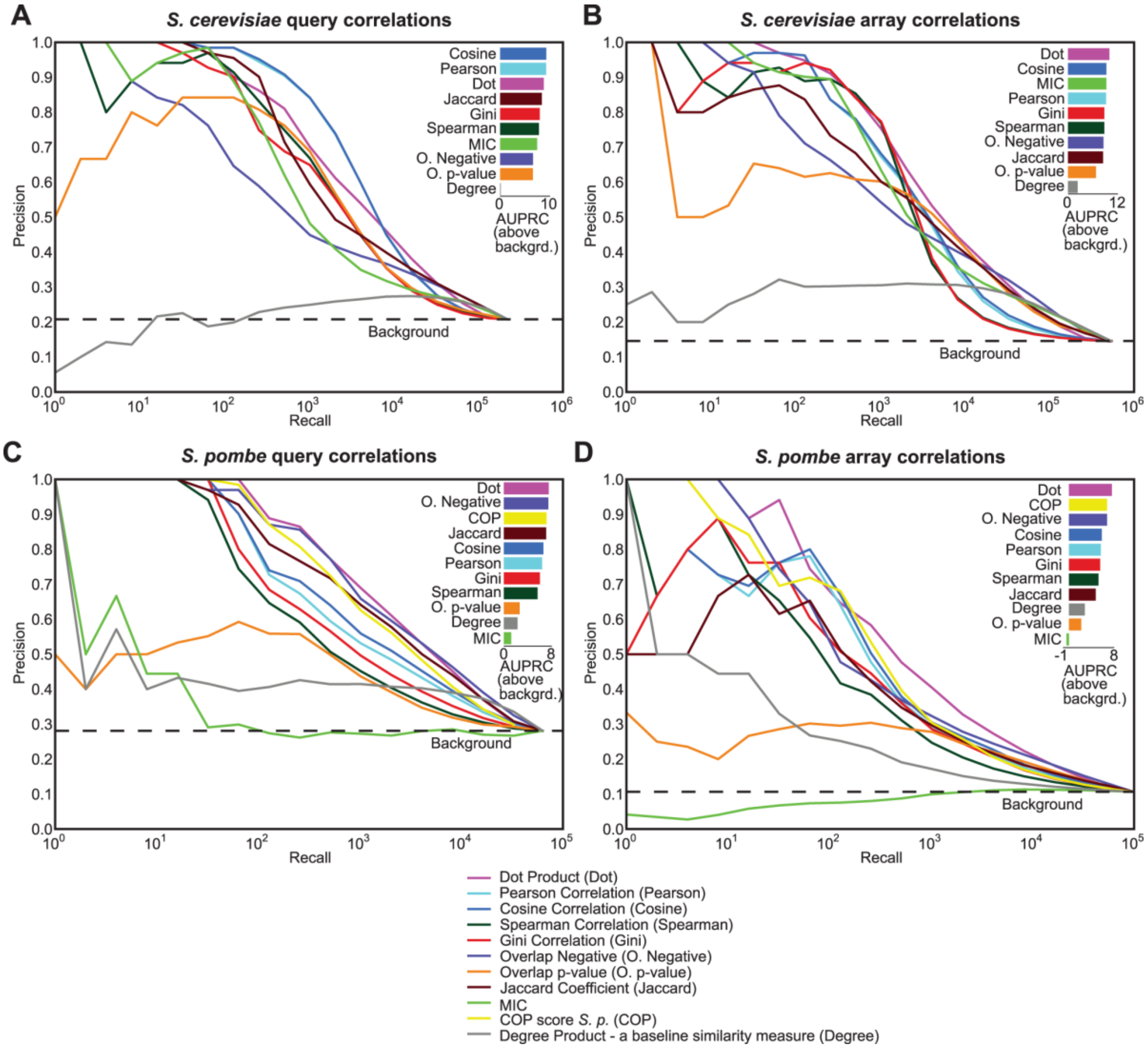


Figure 4.1 Comparison of similarity measures applied to genetic interaction datasets.
Gene pair correlations derived from each similarity measure were benchmarked against a Gene Ontology-based standard using precision-recall statistics. The comparison was conducted on (A) *S. cerevisiae* genetic interaction data (Costanzo *et al.* 2010) - query genes' similarities, (B) *S. cerevisiae* genetic interaction data - array genes' similarities, (C) *S. pombe* genetic interaction data (Ryan *et al.* 2012) - query genes' similarities, and (D) *S. pombe* genetic interaction data – array genes' similarities. The horizontal dotted line shows the background precision expected from randomized ranking of gene pairs. The bar plot on the upper right corner in each section shows the area under the precision-recall curve (AUPRC) above the background for each similarity measure. The

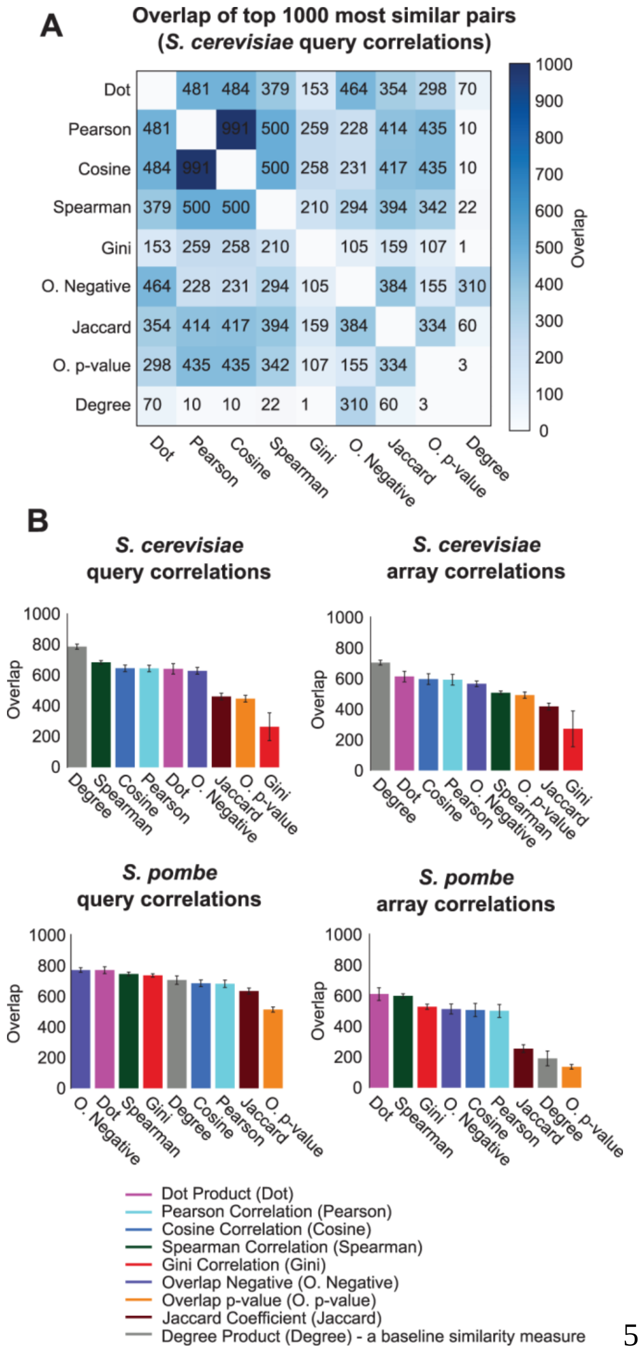
area was calculated by summation of the areas of trapezoids at increments of 2^n ($\log_2$units). The bars are sorted by their respective areas above background.

MIC, a general correlation measure theoretically capable of discovering a wide variety of relationships, performed non-randomly but worse than most other similarity measures on the *S. cerevisiae* data. On the *S. pombe* data, the MIC performed slightly worse than the background (random) expectation, which is surprising. This low performance may be because MIC might be emphasizing technical artifacts in the data (e.g. batch effects [142]) not related to the biological signal. Another notable disadvantage of MIC is that it is computationally very expensive to compute all pairwise correlations on genetic interaction matrices using this measure. For example, the MIC tool takes almost one month on one processor (2.3GHz, 16 GB RAM) to compute all pairwise correlations on the Costanzo *et al.* 2010 genetic interaction matrix, compared to approximately 1 second to compute the dot product on the same data. Therefore, it was not possible to consider MIC in subsequent analyses.

The COP score is a correlation measure developed specifically for clustering and understanding genetic interaction data. The original COP score was based on a combination of log-logistic probabilities of Pearson correlation of the profiles and direct interaction between the genes. The parameters for the log-logistic function were determined from a gold standard of protein complexes [143,159]. One limitation of the original COP score is that it requires interactions between all pairs of genes, so it can be applied only to symmetric genetic interaction matrices, such as the one published in Collins *et al.* 2007 [143]. In the Ryan *et al. S.pombe* study [11], the authors addressed this limitation by using only the Pearson correlation component of the COP score. We used the *S. pombe* similarity matrix published in Ryan *et al.*, 2012 [11], which was calculated using the modified COP score, for the evaluations here. We find that the modified COP score performs better than the Pearson correlation but is slightly worse than the dot product (Figure 4.1C, D).

There is a possibility that the top profile similarities are driven solely by the degrees of the gene profiles [161], trivially favoring high similarities between pairs of hub genes. To address this concern, we included the product of the negative genetic interaction degrees of two genes as a baseline similarity measure between hubs. Our evaluations show that this measure is slightly better than background but is much worse than most similarity measures for both *S. cerevisiae* and *S. pombe* data (Figure 4.1). This result confirms that the similarity measures are discovering relationships between genes that cannot be trivially predicted using genetic interaction degree alone.

Are profile similarity measures discovering the same or different top most similar gene-pairs? To answer this question, we compared the ranked lists of gene pairs obtained from different similarity measures. We find that the top gene-pairs are similar across similarity measures yet they are not completely identical. For example, the maximum overlap between top 1000 most similar gene-pairs for any two similarity measures (except Pearson-cosine) is at most 500 (Figure 4.2A). This observation is surprising considering that several similarity measures exhibit similar precision-recall performance (Figure 4.1). This observation suggests that all similarity measures are not identical even though their precision-recall performances may be and there is considerable diversity between the various measures.



	Dot	Pearson	Cosine	Spearman	Gini	O. Negative	Jaccard	O. p-value	Degree
Dot		481	484	379	153	464	354	298	70
Pearson	481		991	500	259	228	414	435	10
Cosine	484	991		500	258	231	417	435	10
Spearman	379	500	500		210	294	394	342	22
Gini	153	259	258	210		105	159	107	1
O. Negative	464	228	231	294	105		384	155	310
Jaccard	354	414	417	394	159	384		334	60
O. p-value	298	435	435	342	107	155	334		3
Degree	70	10	10	22	1	310	60	3	

50% of the genetic

Figure 4.2 Comparison of similarity measures based on highly similar gene pairs and stability on partial data.

(A) The overlap of the top 1000 similar gene pairs for each similarity measure were compared against each other on the query gene side of the S. cerevisiae genetic interaction data (B) The stability of the similarity measures was assessed by comparing the overlap of top 1000 similar gene pairs computed using 10 different random selections of 50% of the data in each profile.

Stability of a similarity measure when partial profiles are used is important as often genetic interactions profiles are incomplete. For example, for the *S. cerevisiae* genetic interaction screens, the query side is incomplete as screens are still ongoing, and for the *S. pombe* screens, both query and array sides are incomplete. Gene pairs assessed as similar by the similarity measures on partial profiles would ideally not dramatically change as these profiles are completed. To quantitatively assess the stability of the similarity measures, we computed gene similarities using partial profiles generated by randomly selecting interaction data. For example, 50% of the array genes were randomly selected to compute similarities between query genes. This random selection and computation of similarities was conducted 10 times, and the top 1000 similar gene pairs were compared across different selections. We observe that linear similarity measures, Dot product, Cosine and Pearson are the most stable similarity measures (Figure 4.2B). Spearman correlation and overlap negative are similar in performance. Other similarity measures are either not consistent across all datasets or are consistent only for a few cases. For example, Overlap p-value and Jaccard coefficient are the least stable similarity measures across different datasets. Gini correlation is an example of a similarity measure that is consistent only for *S. pombe* genetic interactions but is the least stable similarity measure for *S. cerevisiae* dataset.

4.5 Thresholding effects

Thresholding of the genetic interaction data by considering only interactions greater than a certain score and setting the rest to zero is a common way of removing weaker and noisier interactions [10,142]. Although thresholding focuses the analysis on the stronger signal, it may have undesirable effects depending on which profile similarity measure is used. To check the effect of thresholding on similarity measures, we systematically evaluated the similarity measures for several cases of thresholding on the query genes' correlation of Costanzo *et al.* data (Figure 4.3; refer to Figure 4.1A for no thresholding case).

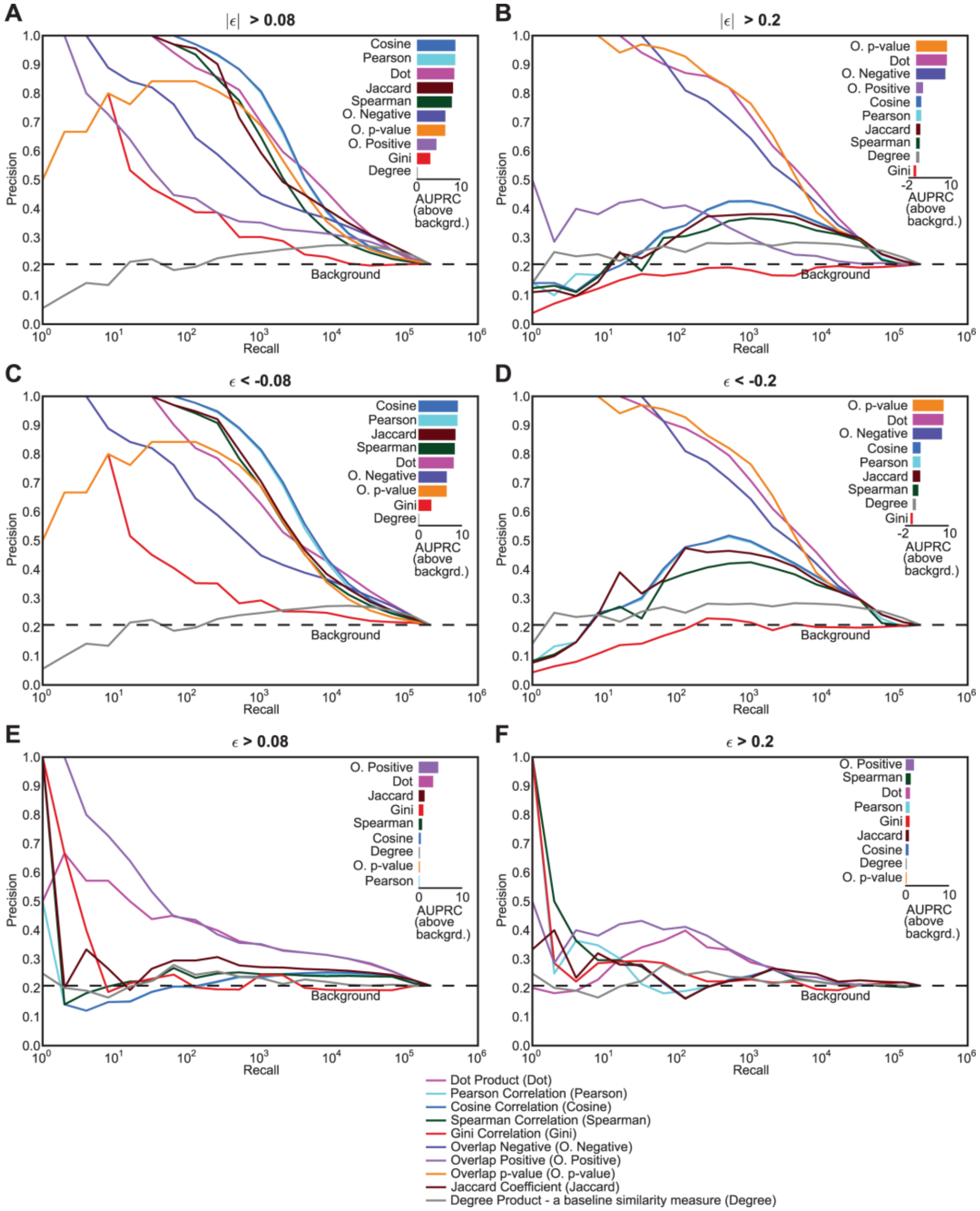


Figure 4.3 Role of thresholding genetic interaction data in the performance of similarity measures. The precision-recall plots were compared on the query side of the *S. cerevisiae* genetic interaction data at several thresholds (A) ε<−0.08 - only negative genetic interactions at intermediate threshold, (B) ε<−0.2 - only negative

genetic interactions at a stringent threshold, (C) ε >0.08 - only positive genetic interactions at an intermediate threshold, (D) ε >0.2 - only positive genetic interactions at a stringent threshold, (E) |ε| >0.08, negative and positive interaction at an intermediate threshold, and (F) |ε| >0.2, negative and positive interaction at a stringent threshold. The bar plot on the upper right corner in each section shows the area under the precision-recall curve (AUPRC) above the background for each similarity measure. The area was calculated by summation of the areas of trapezoids at in increments of 2^n ($\log_2$ units). The bars are sorted by their respective areas above background.

When absolute thresholding was used (intermediate: |ε| > 0.08; stringent: |ε| > 0.2; Figure 4.3A,B) we find that the precision-recall performance of the similarity measures that normalize gene profiles (Pearson, cosine, Spearman and Gini correlations) decreases dramatically after stringent thresholding (Figure 4.3 A,B). On the other hand, non-normalizing correlation measures (for example, dot product and overlap-negative) were robust to thresholding (Figure 4.3). In fact, overlap-negative and overlap p-value's performances improve with the stringency of the threshold cutoff. When only negative interactions were considered (intermediate: ε < -0.08; stringent: ε < -0.2; Figure 4.3 C,D), the performance was more or less similar to absolute thresholding for most similarity measures. However, when only positive interactions were considered (intermediate: ε > 0.08; stringent: ε > 0.2), we found that only the overlap and the dot product showed performances better than random expectation (Figure 4.3E,F). This suggests that negative genetic interactions are the main driver for most correlation measures.

The drop in the performance of some similarity measures on thresholded data raises the question of whether there is any functional signal in the weak interactions, and whether that signal is contributing to better performance of these measures when not thresholded. To address this question, we reassigned the weak interactions (|ε| < 0.2) by randomly shuffling their values among themselves. After randomizing, we were surprised to observe that the performance of Pearson correlation increased to the near original performance (Figure 4.4A). In contrast, the dot product is neither affected by the thresholding nor by the random shuffling of the weak interactions (Figure 4.4B) suggesting that high dot product similarity is not driven by weak interactions. Both of these observations clearly suggest that the weak interactions have little or no biological information in them.

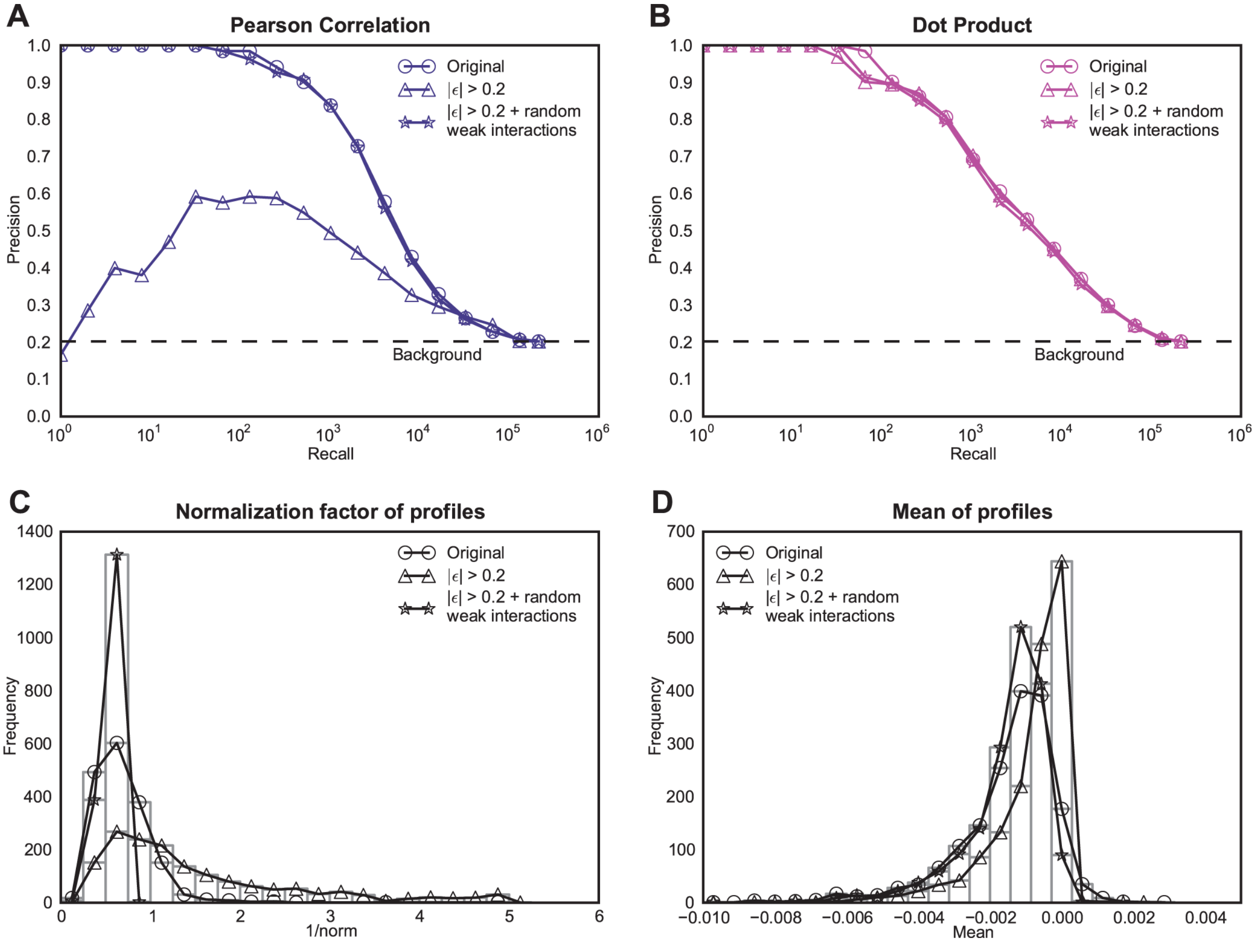


Figure 4.4 Investigation of Pearson correlation relative to the dot product for thresholded genetic interaction data.

In each of the panels, three instances of genetic interaction data have been used: original data, original data with all interactions whose absolute value was less that 0.2 set to zero, and original data where all interactions whose interaction value is less than 0.2 reorganized randomly. The three data instances are investigated using (A) the precision-recall performance of Pearson correlation on each instance, (B) the performance of dot product on the same three instances, (C) a histogram of normalization factor (1/norm) of the profiles for the three instances, and (D) a histogram of the mean of profiles for the three instances.

To find out what caused the sensitivity of the Pearson correlation to thresholding, we investigated two differences between Pearson correlation and dot product: mean centering of the profiles and L_2 normalization of the profiles. Pearson correlation mean centers the profiles, which may affect the strength of all interactions and disrupt the performance. However, we rule out this possibility because we observed that the mean of a genetic interaction profile is normally very close to zero (Mean of means across gene profiles = 0.002 == 1% of the 0.2 threshold used; Figure 4.4D). Furthermore, cosine correlation, which does not mean-center the profiles, exhibits similar precision-recall performance to Pearson correlation (Figure 4.1, 4.2). Therefore, the second difference, L_2 normalization of the profiles, is the main factor in Pearson correlation's sensitivity to thresholding. When Pearson correlation was used on the thresholded data, the normalization factor in the Pearson correlation ($1/L_2$ norm of the profile) greatly varied compared to the normalization factor in the non-thresholded data (Figure 4.4C). This variance in the normalization factor means that the interactions in the hub profiles, which have larger norms, are multiplied by smaller normalization factors, and on the other hand, non-hubs, because of smaller norms, are multiplied by a larger normalization factor. This difference in the normalization leads to a relative magnification of the interaction values in non-hub profiles, making them falsely similar to many other profiles. When the interactions are not thresholded or when random noise is added to the data, the weak interactions or noise serve as filler that equalizes the normalization factor across all profiles (Figure 4.4C). This equalization of the normalization factor effectively protects the Pearson correlation from low degree genes exhibiting spurious high correlations. This equalization of the normalization factor also makes it behave much like the dot product, which assumes a uniform normalization factor (normalization factor = 1; see [149] for a discussion on linear measures and their normalization factors).

4.6 Simulated noise

To assess the robustness of the similarity measures to noise in the genetic interaction measurements, we added three kinds of noise to the *S. cerevisiae* data: false negatives, false positives, and additive Gaussian noise. For false negatives, we assigned 95% of the interactions in the data to zero which is equivalent to removing interactions from the data; for false positives, we sampled real interactions and placed 10 times their number in place of non-interactions; and for additive noise, we added randomly sampled Gaussian noise with 0 mean and 0.08 standard deviation to the *S. cerevisiae* data (0.08 is the intermediate interaction cutoff for Costanzo *et al.* data). Query genes' correlations from the *S. cerevisiae* datasets were evaluated after simulated noise was added.

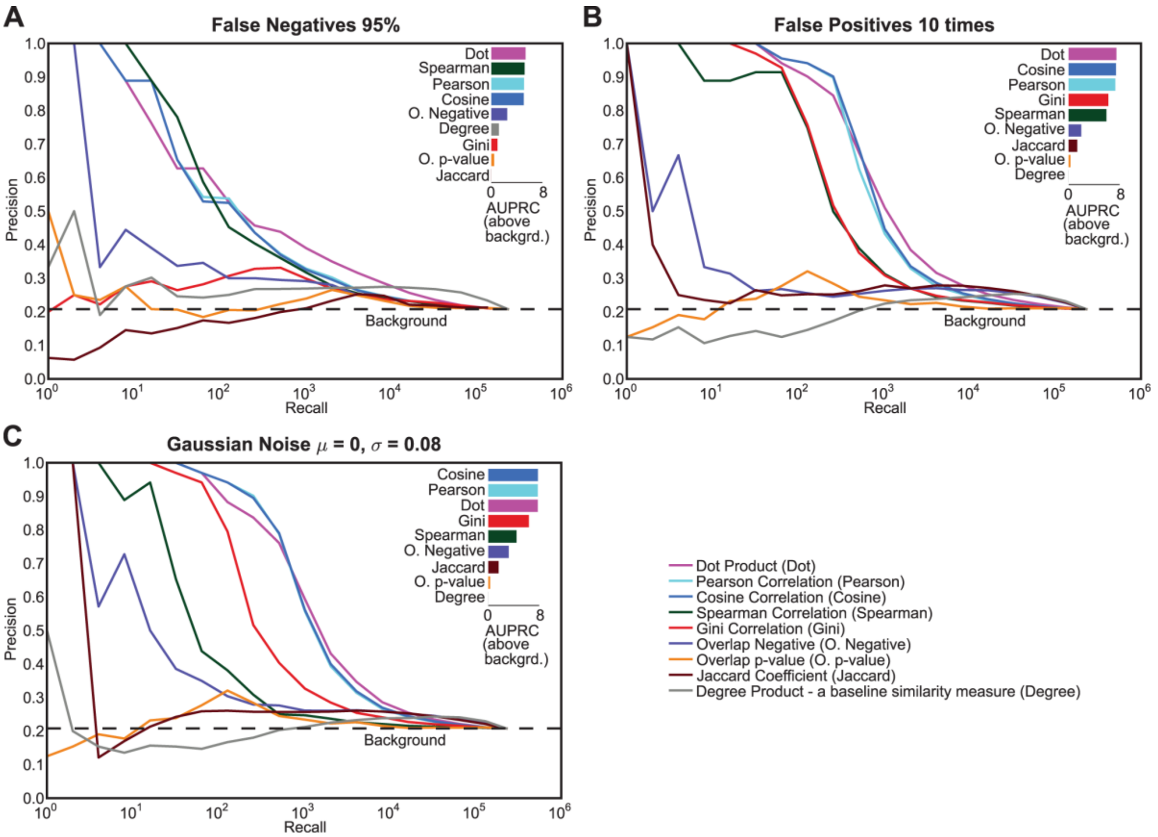


Figure 4.5 Role of noise in the genetic interaction data on similarity measure performance.
In each panel, simulated noise was added to the *S. cerevisiae* genetic interaction data, and query correlations were used for comparing the similarity measures. The simulated noise conditions are (A) false negatives –95% of the significant interactions whose absolute value of interaction is greater than 0.08 were randomly set to 0, (B) false positives – values were randomly sampled from the set of genetic interactions whose absolute interaction value were greater than 0.08 and were randomly substituted in place of randomly selected non-interactions. This random sampling was repeated until 10 times the number of significant interactions were added as false positives in the original data, and (C) Gaussian noise - random values from a Gaussian distribution of mean 0 and standard deviation 0.08 were added to all values (interactions and non-interactions) in the dataset. The bar plot on the upper right corner in each section shows the area under the precision-recall curve (AUPRC) above the background for each similarity measure. The area was calculated by summation of the areas of trapezoids at in increments of 2^n ($\log_2$ units). The bars are sorted by their respective areas above background.

As expected, all types of noise adversely affect the precision-recall performance of all similarity measures (Figure 4.5). However, the dot product measure appears to be the most robust similarity measure to all three noise conditions (Figure 4.5). Surprisingly, overlap-negative, which is a binary version of the dot product on negative genetic interactions, performs poorly in the noise conditions. The two other linear similarity measures, Pearson and cosine correlation, perform well at low recall but the precision-recall performance drastically decreases at higher recall (Figure 4.5). In contrast, Jaccard coefficient and overlap p-value are the most affected in all three noise conditions; their performances drop dramatically from near best performance on the original dataset to almost random performance in the noise conditions. Gini correlation is also strongly influenced by noise, but seems to be most sensitive to false negatives as it performs reasonably well for the false positives and Gaussian noise conditions.

4.7 Batch effects

High-throughput genetic interactions are adversely affected by batch effects, traces of which may still remain in spite of normalization methods for removing them [142]. To determine how batch effects influence the profile similarity measures, we simulated batch effects in the *S. cerevisiae* genetic interaction data by randomly creating batches of 5 query genes and to each profile in the batch we added a common bias for each gene sampled from a Gaussian distribution with mean, $\mu = 0$ and standard deviation, $\sigma = 0.02$ ($\mu = 0$, $\sigma = 1$ for Ryan *et al.* 2012), and further, we added Guassian noise ($\mu = 0$, $\sigma = 0.02$ or 1) to the entire dataset. We simulated more serious batch effects by doubling the magnitude of both the batch signal and the standard deviation of the Gaussian noise added to the entire dataset. Query correlations were analyzed using precision-recall performances to understand the impact of batch effects.

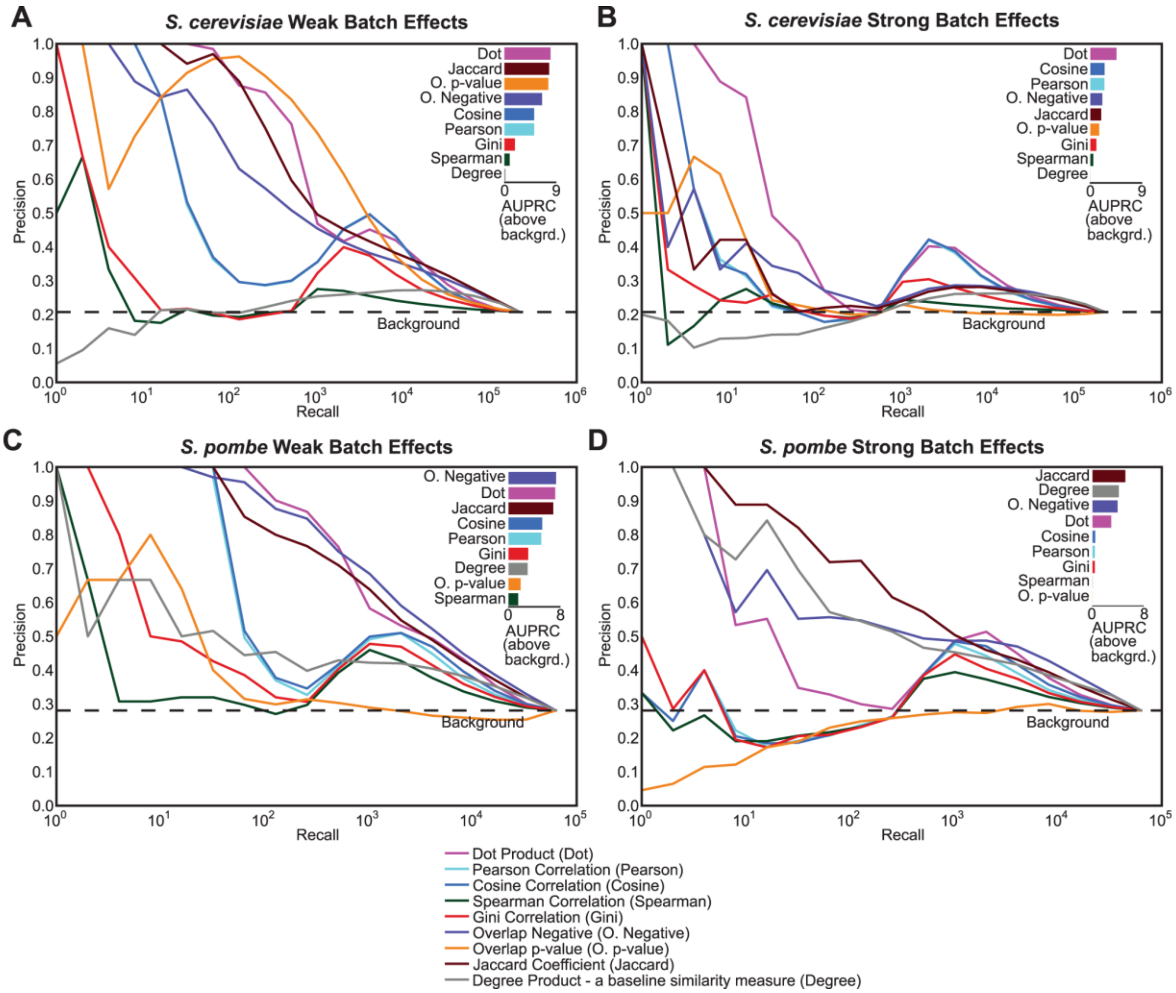


Figure 4.6 Role of simulated batch effects in genetic interaction data on similarity measure performance. (A) shows the performance of similarity measures on the query side of the *S. cerevisiae*genetic interaction network when simulated intermediate batch effects were added to the data. The batch effects were added by creating random batches of size 5 and for each batch, Gaussian noise ($\mu = 0$ and $\sigma = 0.02$) was added. Furthermore, Gaussian noise ($\mu = 0$ and $\sigma = 0.02$) was added to entire dataset. (B) A stronger batch effect signature and noise was added ($\mu = 0$, $\sigma = 0.04$ for both batch effect and noise) (C), (D) are similar plots for the query side of the *S.pombe* genetic interaction data ($\mu = 0$, $\sigma = 1$ for (C), and $\mu = 0$, $\sigma = 2$ for (D)). The bar plot on the upper right corner in each section shows the area under the precision-recall curve (AUPRC) above the background for each similarity measure. The area was calculated by summation of the areas of trapezoids at in increments of 2^n ($\log_2$ units). The bars are sorted by their respective areas above background.

We find that batch effect conditions are destructive for most similarity measures, especially those that include normalization (Figure 4.6). This reduction in performance is because normalization magnifies the batch signature for profiles with fewer interactions, and therefore, falsely predicts genes with weak profiles in the same batch to be functionally similar (Figure 4.6).

Other similarity measures that do not use normalization perform much better. For example, we observe that the dot product and Jaccard similarity measures are robust to batch effects. The robustness of Jaccard coefficient to batch effects is surprising given that it was one of the most sensitive similarity measures to different noise conditions. The overlap p-value is another surprising candidate: it performs poorly in the noise conditions (Figure 4.5) but is the best performer for query correlations of Costanzo *et al.* data under intermediate batch effect conditions (Figure 4.6A). However, the performance of overlap p-value is not consistent for stronger batch effects or for the *S. pombe* dataset (Figure 4.6B,C,D). When the batch effect was severely increased, all similarity measures break down in terms of their performance (Figure 4.6B,D).

There is a peak in the precision-recall performance for higher recall for Pearson, cosine and dot product which corresponds to the point at which all within-batch gene-pairs are exhausted (Number of within batch gene-pairs for 1800 query genes in *S. cerevisiae* 1800/5 * 5 choose 2 = 3600).

4.8 Conclusion

Profile comparison is a widely used approach for genetic interaction studies. Profile similarity measures are used to validate genetic interaction data [142], construct functional profile similarity maps [10], and predict gene functions or drug targets [10,34]. We routinely use Pearson or cosine correlations to cluster genetic interaction data for viewing in clustergram format, and we often look at several versions of the genetic interactions data some of which are thresholded. In this study, we have shown that some similarity measures may not reflect reliable gene-associations when applied to thresholded data, so we need to be aware of this vulnerability. Furthermore, similarity measures have varying sensitivities to different noise and batch effects conditions, so controlling for these conditions may require the use of specific similarity measures.

The dot product is one of the most consistent performers across all datasets and conditions evaluated here. Furthermore, it seems to be more robust to most noise conditions and batch-effects compared to other similarity measures. The dot product can be seen as a hybrid between cosine correlation and overlap measures: applying dot product on binarized data results in the overlap measure and applying dot product on normalized (L_2 norm) data produces the same result as cosine correlation. Indeed, the dot product seems to combine the best properties of these two measures. For instance, cosine correlation performs well on the full dataset but its performance degrades as the threshold on the data is increased. The overlap measure, on the other hand, improves with an increase in the stringency of the threshold and performs well at higher recalls. The dot product is able to retain the good performance of cosine correlation at low recall, but is not affected by thresholds applied to the data and also retains good performance at higher recall. Furthermore, the dot product is among the fastest profile operations computationally, so it is also attractive from that perspective. However, there are a few disadvantages of dot product, the most significant being that unlike several of the other correlation measures, its value is not directly interpretable as it is not bounded (e.g. between -1 and 1). Thus, choosing the right threshold for analysis is highly dataset specific and depends on the size and scale of the values in the interaction dataset. Another disadvantage is that dot product may not be readily available in commonly used clustering softwares. We have tried to address this disadvantage by implementing a dot product measure in cluster 3.0 [165], which is frequently used in combination with Java Treeview [166] for cluster visualization. Our implementation is freely available and can be downloaded from http://csbio.cs.umn.edu/people/RaameshDeshpande/profileSim.

For future work, we propose to develop approaches for combining top similarities from different methods. This proposal is based on the observation that different correlation measures detect different top similarities that all seem to provide relatively high performance in predicting functional associations. This suggests the potential for developing a combined measure that is able to combine the strengths of the different measures to provide superior performance.

4.9 Methods and Materials

4.9.1 Creation of GO standard

The *S. cerevisiae* Gene Ontology (GO) standard was created using the approach described in [162], which generates a co-annotation matrix based on the *S. cerevisiae* Gene Ontology [45] and annotations. Both of the datasets were downloaded on January 22, 2012. The final standard includes annotations for all pairs of 5513 genes with some denoted as positives (functionally related), some as negatives (not functionally related), and some as zero (neither). The GO standard for S. *pombe* was created in a very similar manner to *S. cerevisiae* and contains 4598 genes. The *S. pombe* Gene Ontology data was downloaded on July 4, 2012. The GO standards for both species are available for download on the Supplementary website (http://csbio.cs.umn.edu/people/RaameshDeshpande/profileSim).

4.9.2 Precision-Recall analysis

Pairwise similarities were calculated for all genes appearing in each genetic interaction dataset, except for pairs between multiple alleles of the same gene (multiple alleles for some genes were screened in Costanzo *et al.*), in which case the allele with highest degree was chosen. The calculated gene similarities were sorted and precision and recall were calculated at different similarity thresholds, above which similarities were considered to be positive predictions. These predictions were compared with gold standard positive and negative sets derived from Gene Ontology [162]. For programmatic analyses, an all genes by all genes GO standard matrix is created where values in the matrix are 1, 0 or -1 depending on whether the two genes are co-annotated, undetermined and definitely not co-annotated in Gene Ontology respectively. A prediction that matches with 1 in the GO standard is a True Positive, -1 is a False Positive, and 0s are undetermined and therefore ignored. Pairs with a 1 in the GO standard, but a similarity score below the threshold are considered False Negatives. Precision is given by TP/(TP+FP) and recall by TP/(TP+FN), where TP, FP, FN stand for numbers of True Positives, False Positives and False Negatives respectively. In this study, we have used Recall = TP because the denominator TP+FN is equal to number of 1s in the GO standard, which is a constant.

4.9.3 Genetic interaction datasets

We used two genetic interaction datasets, one from *S. cerevisiae* and another from *S. pombe*. Similarities between the replicates or different point mutation alleles of the genes with other genes could confound precision-recall performance analysis. So in both the datasets, in case of replicates or multiple alleles of the same gene, we have retained the profile with the highest negative interaction degree ($\varepsilon < -0.08$, pval < 0.05). The details of each dataset after processing are as follows:

1. *S. cerevisiae:* Costanzo *et al.* Synthetic Genetic Array (SGA) genetic interaction dataset (after removing replicates) dimensions are 1672 query (rows) by 3885 array (columns) genes.
2. *S. pombe*: Ryan *et al.* Epistatic MiniArray Profiles (EMAP) genetic interaction dataset (after removing replicates), dimensions are 879 query (rows) genes by 1955 array (columns) genes.

Chapter 5: Strategies for optimizing genetic interactions and chemical genomics experiments

5.1 Overview

Screening all pairwise or higher order genetic interactions in an organism can be infeasible because of the combinatorial explosion in the number of experiments that need to be conducted. Here we present strategies for optimizing the number of genetic interaction screens needed for different objectives, which are common to many studies in this area. While we can easily optimize some objectives such as discovering as many genetic interactions as possible with simple strategies, optimizing gene function discovery based on interaction profiles with a limited number of experiments can be more challenging. To optimize this objective, we developed the COMPRESS-GI algorithm, with which we can get interaction profile similarity information nearly equivalent to the complete space with a small fraction of the total pairs screened. This optimization is directly applicable to several screening scenarios including chemical genomics, condition-specific genetic interactions and higher order genetic interaction screens. Moreover, we have developed iterative version of the algorithm, LAF, which can be used for screening genetic interactions in new organisms where no prior interaction data is available (*de novo*).

The results described in this chapter are currently being prepared for submission to a journal. This project is part of a larger chemical genomics effort, which includes a large group of collaborators at the University of Toronto. Jeff Piotrowski, Sheena Li, Michael Costanzo, Kerry Andrusiak and Anastasia Baryshnikova were part of the experimental team who actively discussed the development and constructed the minipool from the genes I selected. Chad supervised the project.

5.2 Background

High-throughput genetic interaction experiments have become very popular because of their utility in revealing the functional relations between the genes [10,11]. A genetic interaction experiment between two genes measures how different the phenotype of the double gene knock-out is compared to the expected double knock-out phenotype predicted from the single mutants. Typically, in high-throughput genetic interaction screens, a mutant strain (query) is crossed into a library of mutants (array). Despite improvements in technology that have made several large-scale genetic interactions studies possible; the largest two studies are ~30% [10] and ~10% [11] complete in *S. cerevisiae* and *S. pombe*, respectively. The single overwhelming challenge for completing the screens is that the space of all possible gene pairs is enormous, and this space becomes even larger when other dimensions such as conditions or higher-order genetic interactions are considered.

The community has made some attempts to optimize the screening in the past. From a practical standpoint, in the Costanzo *et al.* study [10], genes with low single mutant fitness were prioritized for screening when it became clear that these genes were hubs in the genetic interaction network. Another heuristic popularly used in the community is to pick representative genes spanning across all known functional categories [11]. On the other hand, Casey *et al.* [167] proposed to solve this problem by prioritizing genes with least uncertainty, i.e., genes that can be clearly placed in any cluster. Casey *et al.* also suggested another heuristic, standard deviation of the genetic interaction profiles.

Though several ideas have been mentioned in various genetic interaction studies, no systematic attempt has been made to evaluate the different strategies with respect to their utility for how genetic interactions are generally used. In particular, the most popular use cases of genetic interactions are (i) discovering similar genes by finding gene pairs with similar genetic interaction profiles (profile similarity) [10], (ii) discovering important genes in the genome by finding hubs in genetic interaction networks (hub estimation) [10], (iii) discovering pathway level interactions by identifying local structure in the genetic interaction network (network structure) [9], and (iv) use direct genetic interactions for specialized cases such as cancer synthetic lethality (interaction coverage). We have devised screening strategies to optimize performance for each of these use cases, and in particular, we focus on the interaction profile similarity case, which is the most popular genetic interaction use case and is also the most challenging one to optimize.

5.3 Results

5.3.1 **Number of genome wide genetic interaction screens required for various genetic interaction use cases**

To investigate how the number of genes included in a genetic interaction screen affects the utility of the resulting data for various use cases, we randomly selected a subset of the array genes of varying sizes and evaluated the performance of each application on these partial profiles. To mimic selection from the whole non-essential genome, we have selected random subsets of genes from the array side of the *S. cerevisiae* genetic interactions [10] which is comprised of nearly the complete non-essential deletion collection.

When the genes are selected randomly, we find that the performance of the profile similarity and degree estimation use cases increases rapidly in the beginning with diminishing improvements for later screens as the performance saturates. For instance, the performance of the profile similarity use case with just random 10% genome screened is on average around 80% of the performance with complete genome screened (Figure 1C). To estimate the performance for the profile similarity use case, we used precision at a recall of 2048 (~2% of all annotated gene pairs) based on a standard of gene co-annotation of genes in the Gene Ontology. Similar to the profile similarity use case, for the degree estimation use case, we observe a correlation of 0.8 between degree estimates of genes with just 10% of the genome screened and the actual degree based on the complete genome screens (Figure 5.1B). Surprisingly, screening genes with low single mutant fitness is worse for the performance of the hub prediction use case compared to a random screening strategy. This observation is also true for the profile similarity use case, where prioritizing genes by the severity of their single mutant fitness defects is worse than a random screening strategy until ~50 screens; however, they improve upon random selection for a higher number of screens.

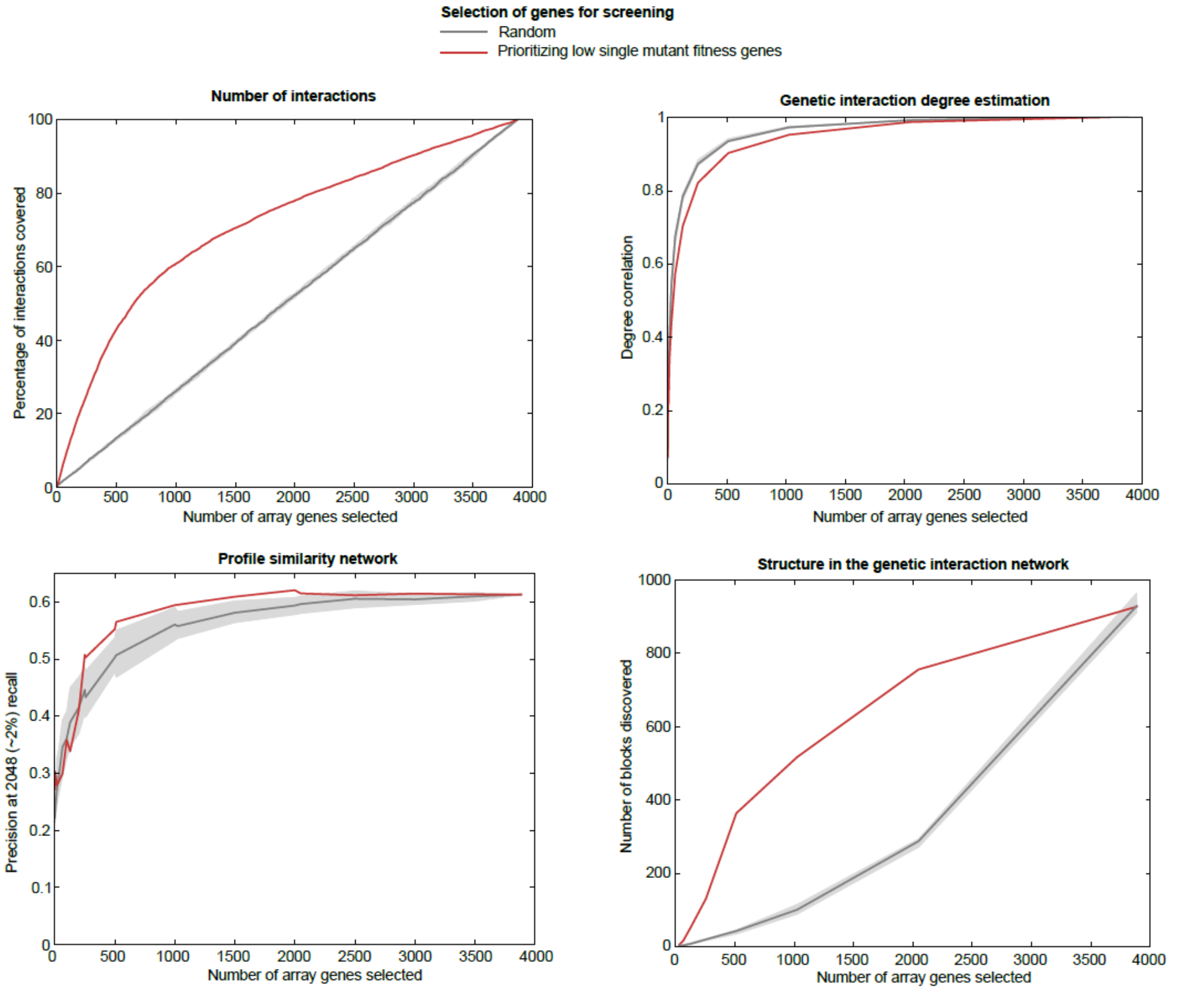


Figure 5.1 Information content for various genetic interaction usecases with selections of random and low single mutant fitness genes.

(A) The figure show how different selections of random genes (gray) and low single mutation fitness genes from the array side of the genetic interactions representing almost the complete non-essential deletion collection affect the percentage of the interactions covered. The random selection was conducted 21 times and the lighter gray region are the inter quartile range while the darker gray line is the median. Similar analyses are repeated for other genetic interaction use cases (B) Genetic interaction degree estimation: The genetic interaction degree is estimated with partial profiles and Pearson correlation is used to compare with degrees obtained from the complete dataset. (C) Profile similarity network: The information context in the profile similarity network was measured using precision-recall curve (see Online Methods). The precision-recall curves are summarized using precision at 2048 recall (approximately equal to 2%). (D) The number of bipartite or the block structure in the genetic interaction interaction network: The number of significant blocks was discovered by running XMOD [9] with different sets of array genes.

This observation suggests that screening low single mutant fitness genes may not be a good strategy relative to the screening of random genes for the degree estimation or for the profile similarity use case for small scale studies (for screening less than 50 genes).

For the genetic interaction coverage and block structure use-cases, the performance scales linearly with the number of random genes screened (Figure 5.1A, D). However, when genes with the lowest single mutant fitness are prioritized, a large fraction of the genetic interactions as well as the block structures are quickly recovered by screening only a small fraction of the genes. For example, around 60% of the interactions and around 70% of the block structure are covered by screening only 25% of the genome (Figure 5.1A, D). The reason low single mutant fitness genes perform well is because the hubs in the genetic interaction network are well-predicted by single mutant fitness, and therefore, prioritizing low single mutant fitness genes is a good proxy for prioritizing hubs.

5.3.2 **COMPRESS-GI: A method for discovering informative set of genes to optimize genetic interaction profile similarity use case**

Genetic interaction profile similarity is one of the popular use cases used especially for discovery of functionally similar genes, validation and visualization of genetic interaction networks. Since complete genetic interaction profiles may not be required for estimation of the profile similarity network, we have developed an algorithm, COMPRESS-GI (**COM**press **P**rofiles **R**elated to **E**pistasis by **S**electing **I**nformative **G**enes), that can select an informative set of genes for screening to optimize the performance of the profile similarity use case. The method requires genetic interactions and Gene Ontology datasets as input and outputs an informative set of genes that maximizes the precision-recall statistics of the profile similarity network. Precision-recall statistics are a metric for measuring the information content in the genetic interaction data by comparing the profile similarity network with known gene relations (Figure 5.2A; see Methods). We use co-annotations from Gene Ontology, which is a repository of all gene annotations, to obtain a gold standard set of known gene relations. The COMPRESS-GI algorithm was inspired by the Matching pursuit algorithm from the signal processing field [168], which is an algorithm designed for selecting representative signals from a dictionary of signals such that the selected signals span the entire dictionary. In our case, the objective is to select a set of genes such that the precision-recall curve of the similarity network generated from the partial profile based on those genes is maximized. Based on this objective, we use a step-wise exhaustive greedy approach, where we select the most informative gene, and for later iterations, pick the gene that is the most informative gene when added to the already selected set (Figure 5.2B; see Method for details). However, this process is heavily biased by the starting gene, so we repeated the process with different starting genes each taken from the top 50 informative based on the single gene's profile. In addition, we know from previous experience that precision-recall performance on just the global GO standard could be biased by genes belonging to one or few of the functional categories [162]. So we ran the COMPRESS-GI algorithm for each of the 13 major functional categories in the genetic interaction network by modifying the GO standard to retain positive co-annotations belonging to the category while removing other positive co-annotations (see Methods for details).

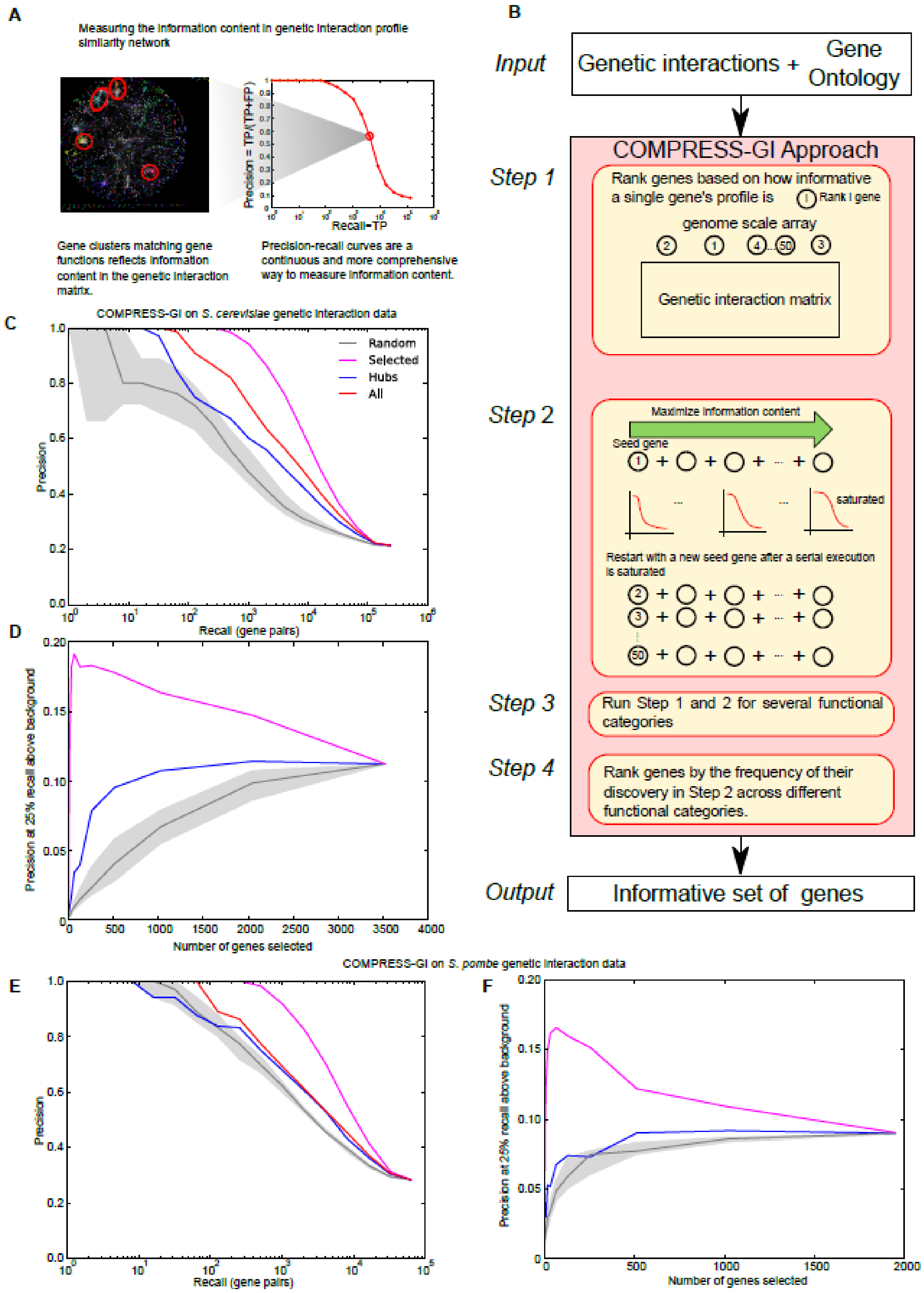

A
Measuring the information content in genetic interaction profile similarity network
Precision = TP/(TP+FP)
Recall=TP
Gene clusters matching gene functions reflects information content in the genetic interaction matrix.
Precision-recall curves are a continuous and more comprehensive way to measure information content.
B
Input
Genetic interactions + Gene Ontology
COMPRESS-GI Approach
Step 1
Rank genes based on how informative a single gene's profile is
Rank i gene
genome scale array
Genetic interaction matrix
Step 2
Maximize information content
Seed gene
saturated
Restart with a new seed gene after a serial execution is saturated
Step 3
Run Step 1 and 2 for several functional categories
Step 4
Rank genes by the frequency of their discovery in Step 2 across different functional categories.
Output
Informative set of genes
C
COMPRESS-GI on S. cerevisiae genetic interaction data
Random
Selected
Hubs
All
Precision
Recall (gene pairs)
D
Precision at 25% recall above background
Number of genes selected
COMPRESS-GI on S. pombe genetic interaction data
E
Precision
Recall (gene pairs)
F
Precision at 25% recall above background
Number of genes selected

Figure 5.2 COMPRESS-GI algorithm

(A) The information content in the genetic interaction profile similarity network is capture by a precision-recall curve. Given a profile similarity network at a particular similarity threshold, it can be compared with positive and negative co-annotation in Gene Ontology to calculate precision and recall. The precision and recall when plotted at several thresholds give precision-recall curve. The flowchart describing the COMPRESS-GI method. The inputs for the COMPRESS-GI algorithm are genetic interaction data, Gene Ontology standard and broad gene membership in functional categories, and it outputs informative set of genes. The COMPRESS-GI algorithm selects array genes (column genes) to optimize the precision-recall statistics of genetic interactions data with those array genes. (C) Precision-recall evaluation of the selected genes and comparison with equal number of random set of genes, equal number of hubs, and the entire genetic interaction dataset. (D) The evaluation in (C) was repeated for different functional categories by modifiying the Gene Ontology standard, and the precision at xxxx recall (~25%) was averaged across the different functions. The randomization for (C) and (D) was done 100 times and the middle line is the median across those runs and the ends of the shaded area represents the first and third quartile respectively.

We combined all these different lists of informative set of genes generated to obtain a list of around 200 genes.

5.3.3 Evaluation of the COMPRESS-GI algorithm

The set of genes derived from the COMPRESS-GI algorithm provide a major improvement over a random screening strategy, both when evaluated globally on the complete Gene Ontology (Figure 5.2C) and for the functional category-specific evaluations (Figure 5.2D). Our selected set of genes also perform better than an equal number of hubs, which one might guess could provide a reasonable strategy for maximizing the functional information derived from genetic interaction screens (Figure 5.2C,D). More strikingly, the selected genes perform better than even the complete dataset for the global evaluation (Figure 5.2C), and comparably to the complete dataset on the function-specific evaluation (Figure 5.2D). We note that the precision-recall performance metric is not a monotonically increasing function with the number of genes selected. The better performance of the selected genes over the complete dataset suggests that there are non-informative or noisy genes in the complete dataset that actually detract from functional information in the genetic interaction data and thus, bring the precision-recall performance down relative to a smaller set of informative genes.

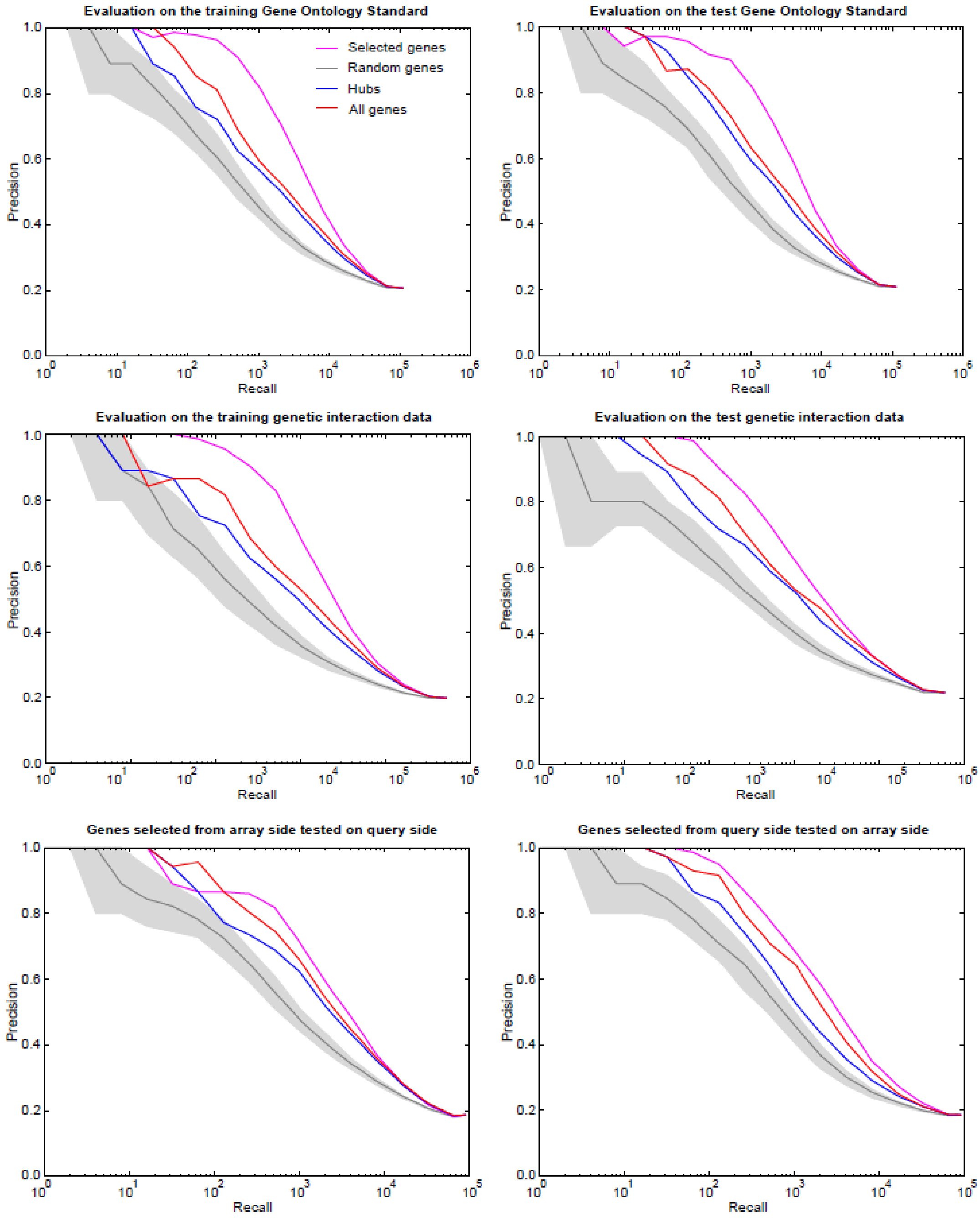


Figure 5.3 Evaluating the robustness of COMPRESS-GI algorithm to overfitting.
The Gene Ontology is split into training and test samples for cross-validation, and COMPRESS-GI is run on the training GO standard and informative set of genes is discovered. The informative set of genes is evaluated using precision-recall curves on training (A) and test GO standards (B). Likewise, the cross-validation is repeated for genetic interactions data, by split it into training and test halves on the query side. The informative set of genes discovered on the training genetic interaction data is evaluated on both training (C) and test halves (D). The

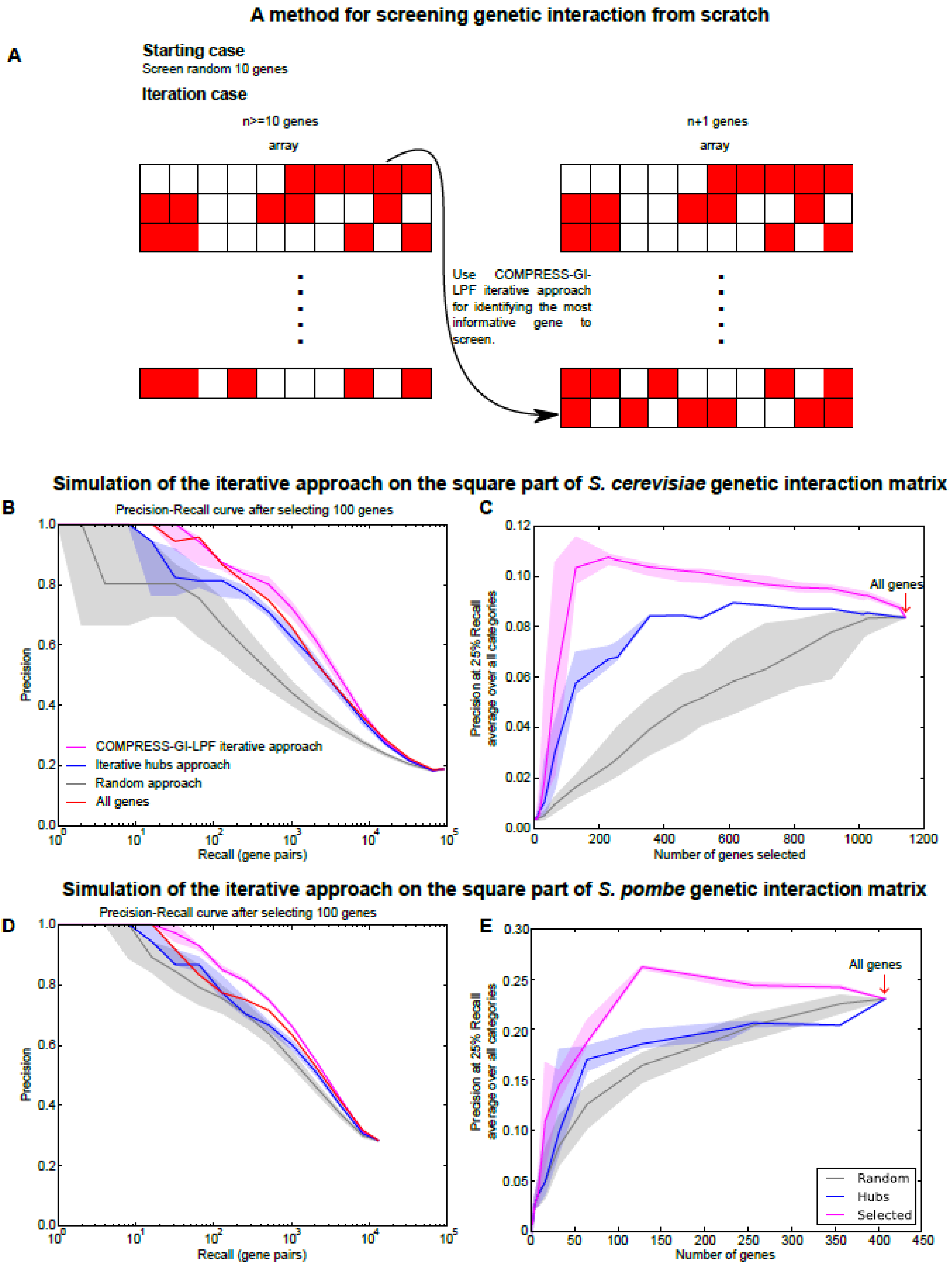
informativeness of the informative set of genes discovered on the array side is tested on the query side (E) and vice versa (F). Square genetic interaction matrix is used for tests E, F. The hubs in each of the plot are dataset specific, for example, in (D) hubs from the test genetic interactions are used. For random baseline, the randomization was run 100 times and median precision-recall curve was plotted. The lower and upper bounds of the gray area represent first and third quartile.

To make sure that the algorithm is not simply overfitting, we conducted two cross-validation experiments. In the first cross-validation experiment, we identified informative genes by applying COMPRESS-GI on only 50% of the randomly selected genetic interaction query screens and then tested the ability of these selected genes to provide profile information for the held out 50% of the genetic interactions screens. In the second experiment, we held out 50% of the GO annotations when selecting the informative gene set, and tested the selected set for functional information on the held-out pairwise GO co-annotations. In both cross-validation experiments, the informative genes discovered in the training data are equivalently informative even for held-out data (Figure 5.3 A,B), suggesting that the approach is not overfitting.

Since we are proposing that informative genes discovered from the array set be prioritized for genetic interaction screening as query mutants, we also checked whether informative array genes are also informative on the query side, and vice versa. To do this, we considered only the square part of the *S. cerevisiae* genetic interaction data (genes that appeared on both sets of the matrix) and discovered informative sets of genes by running COMPRESS-GI on the array side and then checked the information content of the same set genes on the query side, and vice versa. We find that the genes are informative in both cases, suggesting our method will indeed work for selecting new queries in practice (Figure 5.3E,F).

5.3.4 Iterative LAF: an iterative approach for screening genetic interactions *de novo*

The informative set of genes in COMPRESS-GI is well-suited for applications such as chemical genomics screens and condition-specific genetic interaction screening (see Application of COMPRESS-GI to optimize large-scale chemical genomic screens). However, the algorithm may not be suitable for genetic interaction screens where prior data is not present as the algorithm requires a sizeable genetic interaction matrix as an input. For this *de novo* genetic interaction screening scenario, we developed an iterative LAF (COMPRESS-GI **L**inear **A**lgebra **F**ormulation) approach to prioritize genes (Figure 5.4A). This method will be especially useful for genetic interaction screening in new organisms and also for new conditions in already established model organisms. The LAF approach is an approximation of the COMPRESS-GI approach but is orders of magnitude faster (see Methods). LAF optimizes the sum of products of similarities between genes and their known GO co-annotations between them (0, 1, or -1), which can be summarized as Hadamard's product or element-wise matrix multiplication. Using properties of the trace on Hadamard's product along with the cyclic property of the trace product of matrices, the problem can be reduced to a simple 0-1 knap-sack problem,


A method for screening genetic interaction from scratch

Simulation of the iterative approach on the square part of *S. cerevisiae* genetic interaction matrix

Simulation of the iterative approach on the square part of *S. pombe* genetic interaction matrix

Figure 5.4 Screening genetic interactions *de novo*.

(A) The schematic describes the iterative genetic interaction screening scenario, for which COMPRESS-GI-LAF was developed. (B,C) The method is evaluated by simulating iterative genetic interaction screening on the square genetic interaction Costanzo et al. data. (B) compares the precision-recall curves for 100 genes selected by this approach with iterative hub and random approaches. (C) compares of the precision at 25% recall performance of the different approaches averaged across different functional categories. In addition to these approach, the performance of the complete genetic interaction data has also been added in (B,C). Since there is a random component for selecting first 10 genes, the approaches were repeated 10 times with different initial set of genes. Each of the random case, where the rest of the 90 genes are random, were repeated 10 times.

giving each gene a score that is related to the gene's informativeness (see Methods for derivation). The genes can be ranked by their scores and top genes can be selected and screened.

Given its computational efficiency, the LAF approach is suited for the iterative genetic interaction screening scenario where screens are selected in an online fashion after each additional screen. For comparison, we have developed a baseline approach, iterative hubs, which is based on screening the highest-degree unscreened hub after each screen. Both iterative LAF and iterative hub method involves first screening of 10 random genes, after which the genes are iteratively prioritized and screened.

We simulated and evaluated the *de novo* genetic interaction screening scenario onthe Costanzo *et al.* genetic interaction data. We selected a submatrix of the genetic interaction matrix such that genes on the array side are also on the query side which will ensure that we have screens for the genes we will select from the array side (1141 query genes by 1141 array genes). As mentioned above, 10 genes were randomly screened first followed by 90 iteratively selected genes, for a total of 100 query gene screens. To measure the performance of each approach, a profile similarity network was constructed by measuring similarity between all pairs of array genes based on the 100 selected query genes, and evaluated with the Gene Ontology co-annotation standard using precision-recall analysis. Similar simulations were conducted to select 100 genes using the baseline iterative hubs approach. We observe that the iterative LAF method performs better than both the iterative hubs approach and random screen selection (Figure 5.4B). For a broader perspective of how the algorithm performs as more genes are selected for screening, the selection was continued beyond 100 genes to the completion of the square matrix. The genes were then evaluated across different functional contexts using precision-recall statistics. Again, the precision at 25% recall performance averaged over the 13 functional contexts is higher for iterative LAF approach compared to iterative hubs and random (Figure 5.4C).

To make sure that the method is generalizable to other species, we carried out a similar simulation approach on the published *S. pombe* genetic interaction data [11] (Figure 5.4D, E). Similar to results in *S. cerevisiae* simulation, we observed that the genes selected by the iterative LAF approach perform better than both random and iterative hub baseline approaches. The positive results in both species suggest that the algorithm will be useful in other organisms as well.

5.3.5 **Application of COMPRESS-GI to optimize large-scale chemical genomic screens**

The informative set of genes discovered by our COMPRESS-GI approach is directly applicable to the chemical genomics screening setting (Figure 5.5A). In chemical genomics, mutant strains are treated by a compound to find the strains' resistance or sensitivity to the compound. This sensitivity profile can then be compared against other compounds to discover mechanistic similarities between the compounds or the sensitivity profile can be compared to a database of genetic interaction profiles to predict the target (protein) of the compound. A chemical genomic profile across the complete, whole genome collection of mutant strains would be ideal, but if the major aim of the experiment is to conduct profile correlation of the chemical profiles with other chemical or genetic profiles, then chemical genomics screens against an optimally selected set of genes discovered by COMPRESS-GI should perform as well as a whole genome screen. Importantly, this optimization will save resources and make large-scale chemical genomics experiments feasible especially when amounts of compounds are a limiting constraint. The diagnostic set of genes we have selected roughly comprise 5% of the non-essential genome (200 genes), thus conducting chemical genomics experiments using this diagnostic set will reduce both the experimental cost and amount of compounds required by 20-fold.

To evaluate the set of genes selected by the COMPRESS-GI algorithm for the chemical genomics application, we compared the drug-target prediction capability of the compound profile restricted to selected genes with the compound profile restricted to an equal number of random genes. The drug-target prediction is conducted by finding a gene in genetic interaction data whose profile is most correlated with the compound's profile, and is based on the assumption that the compound's behavior will mimic the knock-out of the target gene [169]. Using Parsons *et al.* chemical genomics dataset [33], which is a yeast whole genome chemical genomics screening comprising of 82 compounds, we show that the correlation of the compound with its predicted target is

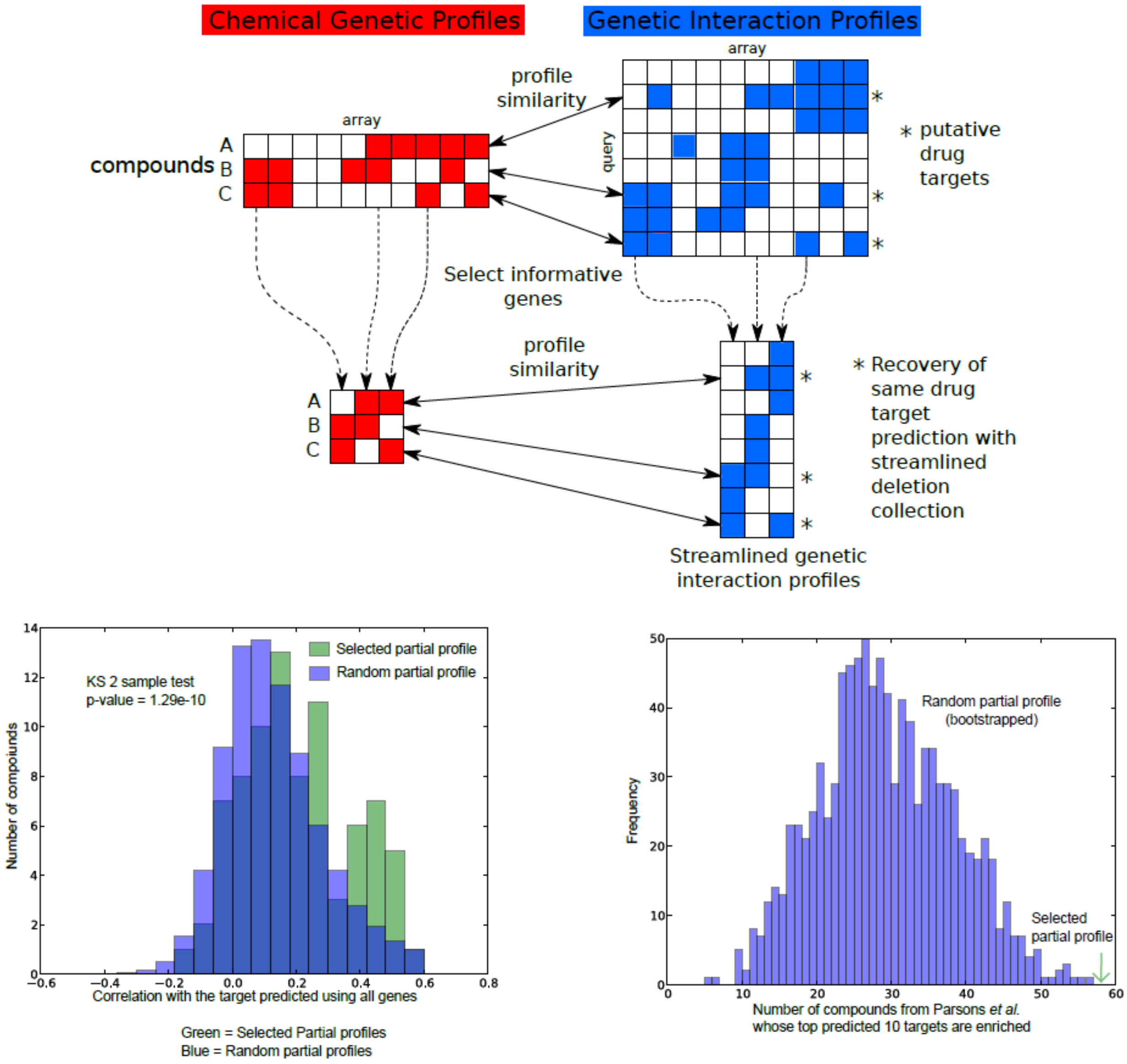

Chemical Genetic Profiles
Genetic Interaction Profiles
array
profile similarity
array
compounds
A
B
C
query
* putative drug targets
*
*
*
Select informative genes
profile similarity
A
B
C
* Recovery of same drug target prediction with streamlined deletion collection
*
*
*
Streamlined genetic interaction profiles
KS 2 sample test
p-value = 1.29e-10
Selected partial profile
Random partial profile
Number of compoiunds
0
2
4
6
8
10
12
14
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Correlation with the target predicted using all genes
Green = Selected Partial profiles
Blue = Random partial profiles
Frequency
0
10
20
30
40
50
0
10
20
30
40
50
60
Random partial profile (bootstrapped)
Selected partial profile
Number of compounds from Parsons *et al.* whose top predicted 10 targets are enriched

Figure 5.5 Application of informative set of genes discovered by COMPRESS-GI to chemical genomics

(A) The schematic shows that we would like to select genes such that even with partial profiles, we can make the same drug-target predictions. (B) Distributions of correlations of chemical compound's chemical genetic partial profile and its target's genetic interaction partial file based on the selected genes and random genes are compared. 82 compounds in the Parsons et al. chemical genomics data is used for this validation. The target for each of the compounds is the query gene (row) in the genetic interaction data that has maximum correlation with the chemical compound's complete chemical genetic interaction profile. (C) In parallel, the top 10 predicted targets for the compounds are checked for enrichment in Gene Ontology terms. The figure shows distribution of the number of compounds out of 82 compounds in Parsons et al. study whose targets based on the random partial profile targets are enriched. The randomization is conducted 100 times. The number of compounds whose targets are enriched with selected profile is shown with the arrow.

higher for the partial profile consisting of selected genes compared to the correlation of the random partial profile of equal size (Figure 5.5B; p-value < $1.3*10^{-10}$). This observation suggests that our diagnostic set of genes outperforms an equal number of random genes. Further, when we look at the enrichment of the top predicted targets, we find that the targets are more likely to be enriched for our selected partial profile compared to the random partial profile (Figure 5.5C; p-value < 0.01). Significant enrichment of the top predicted targets give an indication that the target prediction based on the partial profile is not noisy and is focused on a particular functional neighborhood, which is most likely the actual target neighborhood. Based on our observation, we discover that target prediction based on the diagnostic set of genes is less noisy compared to equal number of random genes. Moreover, the diagnostic set of genes performs (59 compounds show enrichment) better than even the entire profile (42 compounds show enrichment) for the enrichment of target prediction metric.

5.4 Discussion

In this study, we have developed methods that can select diagnostic sets of genes for chemical and genetic interaction screening. We show that screening against these diagnostic sets of genes works just as well as screening against a complete deletion collection for common genetic interaction use cases. In particular, we show that if the objective is to discover functionally similar genes using genetic interactions or predict drug-targets using chemical and genetic interaction data, our diagnostic set of genes is comparable, and sometimes better than even the complete deletion collection.

Complete screening of genetic interaction networks is important for model organisms as it will provide a reference for other chemical genomics and condition specific genetic interaction studies. The genetic interaction community is addressing this issue by conducting exhaustive screening combining all pairs of possible mutations in several model organisms including *S. cerevisiae* and *E. coli*. However, there are several contexts such as chemical genomics and condition-specific genetic interaction experiments, where complete screening may be prohibitive because the condition dimension or the chemical compounds space can be very large. Furthermore, as the community begins to establish screening technologies in higher organisms, they may use our algorithms to prioritize and dramatically increase the information in the dataset even with few screens.

Currently, there are two general screening strategies used for genetic interaction screening in species where such technologies have been developed: the rectangular screen design and square design. The rectangular approach refers to the scenario where query gene mutant strains are crossed against a complete (or near-complete) deletion collection [10]. When several query genes are screened against the complete deletion collection, this creates a rectangular genetic interaction matrix. The square approach involves screening a small set of genes against the same set of genes on the array side [143], an approach which has been adopted in several cases to rapidly cover small sets of functionally related genes. We believe that the community should adopt the rectangular approach because it will provide an unbiased view of the complete genome. We have shown here that only a small number of query genes are required to generate a useful profile similarity network. This number can be further reduced by intelligent selection of genes using the approaches we describe here. Thus, the rectangular approach can give an unbiased picture on a genome scale and, at the same time, be cost effective. Further, the rectangular approach is better suited for distributed screening efforts in which screen data from multiple labs is pooled together. In contrast, scaling up using the square approach by increasing both query and array dimensions can be difficult as it will involve stitching together separate square sections, which may be challenging due to inherent technical biases in the genetic interaction data [142].

5.5 Methods

5.5.1 COMPRESS-GI

Given genetic interaction data (m query genes crossed against n array genes) and Gene Ontology standard for the query genes (size m by m), the COMPRESS-GI method discovers an informative subset of array genes. The optimization objective for selecting the informative subset of genes is to maximize the match between the gene profile similarities based on the selected partial profiles and gene co-annotations in Gene Ontology. The matching is quantified using precision-recall statistics by treating gene profile similarities as predictions and co-annotations from Gene Ontology as the gold standard positive and unrelated genes in the Gene Ontology as gold standard negative. The informative set of genes is discovered by exhaustively searching for genes that when added to the selected set of genes will best improve the precision-recall statistics. For example, for discovering the first gene, we conduct an exhaustive search of all the array genes and the gene that gives the best precision-recall statistics is selected. For the second gene, we search for all the array genes except for the first selected gene, and select the gene that gives best precision-recall statistics along with the first gene. This process is continued until the precision-recall statistic saturates and the increase by adding any gene does not increase the precision-recall statistic significantly. In practice, a maximum of around 20 genes can typically be selected before the precision-recall statistic saturates.

The genes selected by COMPRESS-GI are influenced by genes already selected by the algorithm. For example, different starting gene may give different sets of informative genes. So to make sure that the genes selected are robust to the selections of the starting gene, we ran the COMPRESS-GI algorithm with different starting genes. For example, instead of starting with the best gene as the first gene, we started with the second best gene and allowed the first gene to occur in the COMPRESS-GI selections. We repeated this process with each of the 50 best genes ranked high in the precision-recall statistics based on the single gene profile as starting gene.

Further, to make sure that all the major functional categories are represented by the selected set of genes, we repeated the COMPRESS-GI algorithm for several different functional contexts. The functional context was created by limiting the Gene Ontology standard to only genes that are related to the function.

The different sets of genes obtained by running with different start genes and in different functional contexts are combined and the genes are sorted by their frequency of occurrence in these sets. The optimal number of the genes to be selected is decided based on where precision at 25% recall averaged across different functional categories peaks (see Figure 5.2 E).

5.5.2 **Precision-recall statistic**

The precision-recall statistic is a way to assess both precision as well as recall of predictions against a gold standard truth. In a typical machine learning setting, there is a positive and a negative class which are being predicted. If a positive prediction is found correct according to the positive gold standard, then the prediction is called True Positive otherwise it is called False Positive. Likewise, if a negative prediction is correct according to the negative gold standard it is True Negative otherwise it is False Negative. Precision is TP/(TP+FP) and Recall is TP/(TP+FN), where TP, FP, FN are number of True Positives, False positives, and False negatives predictions respectively.

In our case, precision-recall statistics are used to assess the match between the gene similarities based on partial profiles with co-annotations in Gene Ontology. To evaluate the predictions and compute precision we also need gold standard positive co-annotations and gold standard negative co-annotations. These gold standard co-annotation are generated from Gene Ontology using GRIFN[162] (see Creation of GO standard). The similarities are thresholded at different points (recall equal to integral powers of 2 and the last recall) where precision and recall statistics are calculated and the precision-recall curve is plotted. Since the denominator for recall is constant for all similarity thresholds (TP + FN = number of 1s in the GO standard matrix), we have ignored the denominator and used Recall = TP.

5.5.3 **Comparing Precision-recall curves**

For the COMPRESS-GI approach, precision recall curves are compared to find the best gene to select at each iteration. The precisions are compared at recall equal to powers of 2. The precisions at earlier powers of 2 are compared first. If one of the PR curves has higher precision at that recall, that one is considered to be a better PR curve. In case of tie, precisions at higher recalls are considered. One problem with this approach is that after the PR curve has saturated, even weak profiles can become slightly better by chance. To safeguard against this situation, in addition to checking that the PR curve improves we also check that the increase is greater than the sum of standard error in the two precisions. Given precision p = TP/(TP+FP), where TP, FP are number of true positive and false positives, respectively, the standard error on p is calculated as $\sqrt{\frac{p(1-p)}{TP+FP}}$.

5.5.4 Category specific precision-recall statistics

The COMPRESS-GI is run at several different functional contexts, that is, we want to select informative set of genes for the different functional category. To compute category specific precision-recall statistics and optimize on that objective, we modify the Gene Ontology standard to be specific to the functional category. The GO standard, M, is changed as follows:

(1) $M_{i,j}$ is unchanged if genes i, j both belong to the functional category,

(2) $M_{i,j} = 0$ if originally $M_{i,j} == 1$ and only one of the genes i, j belong to the functional category.

(3) $M_{i,j}$ is unchanged if $M_{i,j} = -1$ originally

The GO standard for the genes within the functional category remain unchanged (1), but co-annotations of gene pairs outside the functional category are set to 0. Even though the focus of the optimization is to select genes informative for a particular functional context, the -1s in GO standard are never changed so that predicting unrelated genes as related is always penalized.

5.5.4.1 Complexity

The complexity of this algorithm depends on the complexity of calculating a precision recall curve, which is repeated the size of the array (n) multiplied by number of genes picked in each run (average a = 10). Also, each of the runs are repeated for different start genes (fraction of array size) and for different functional contexts (f). The main contributor for the complexity in precision-recall statistics is the sorting of the query gene similarities. The computation of the gene similarity takes O(am^2) time and sorting of m C 2 similarities take O(m^{2}logm) time. In total, the algorithm has complexity of O(am^2 + m^{2}logm)(a)(n)(f)(n) = O(m^{2}n^2(a+logm)af). The sorting of the gene similarities is the most expensive part of the algorithm so instead of sorting the entire list of query gene similarities, only the top 100,000 gene similarities were considered. Finding the 100,000th largest number is an O(m^2) operation, which was further implemented in C to increase the speed. Similarities greater than the 100,000th largest similarity were considered and sorted. Here, 100,000 is a small fraction of the m^2 so it does not change the complexity of the algorithm but the run time is greatly improved.

5.5.5 **LAF**

Like COMPRESS-GI, the LAF method is also based on a similar objective of optimizing the match between the similarities of the genes with Gene Ontology standard (G). The similarities of the genes based on the partial profiles can be written as $XW(XW)^T = XWW^TX^T = XWX^T$ where W is the diagonal matrix with $W_{ii} = 1$ if array gene i is selected. However, unlike COMPRESS-GI where precision-recall statistics are used to assess the match between XWX^T and G, we optimize on the sum of element wise multiplication of XWX^T and G. This objective can be written as

$\max_W sum\ of\ matrix\ elements(\ XWX^T \odot G)$ (where $\odot$ is the element wise multiplication and more formally known as Hadamard product)

$= \max_W e^T(\ XWX^T \odot G)e$ (sum of element in matrix $M = e^TMe$)

$= \max_W e^T(\ A \odot G)e$ (Let $A = XWX^T$)

$= \max_W tr(D_e^* A D_e G^T)$ (property of Hadamard product)

$$= \max_{W} tr(AG^T)$$

$$= \max_{W} tr(XWX^TG^T) = \max_{W} tr(G^TXWX^T) = \max_{W} tr(X^TG^TXW)$$

$$= \max_{W} (X^TG^TX)_{ii} * W_{ii}$$

This reduces the problem to the simplest type of 0-1 knapsack problem which can be solved by a greedy algorithm. To solve this problem, we rank the genes by$(X^TG^TX)_{ii}$, and pick the top n genes.

5.5.5.1 Complexity

The complexity of this algorithm mainly lies in the matrix multiplication X^TG^TX. So if X is the genetic interaction matrix composed of m queries and n arrays, the complexity for X^TG^TX matrix multiplication is O(nmm)+(nmn) = O($(mn)(m+n)$). The complexity of the knapsack problem is O(n), so the overall complexity of the algorithm is O($(mn)(m+n)$). This complexity makes the algorithm perfectly reasonable to run on genetic interaction datasets that are several folds larger than the current largest genetic interaction datasets [Costanzo]. Further, the algorithm can be used even for organisms with a much larger number of genes (m,n = 60,000). This complexity allows the algorithm to be run very quickly for iterative approaches, which has been specifically demonstrated in the results.

5.5.6 **Interaction profile similarity measure**

We have used the dot product similarity measure because we showed earlier that it is among the best performers on genetic interaction data[140]. More importantly, similarity measures that use normalization such as Pearson correlation, Cosine correlation and Spearman correlation are unstable on smaller profiles, so they could not be used for this study where we are building the informative set of genes. Further, dot product has been shown to be more robust to noise and batch effects, which are typical in genetic interaction data [140].

5.5.7 Block discovery

To discover block structure in the genetic interaction data, we have used a block discovery method published earlier by our lab, XMOD [9]. We streamlined the XMOD implementation and provide a python interface to XMOD so that it can be run from the command line. For all block discovery applications (Figure 5.1), we use a minimum support threshold of 6 and item-set size of 3, which means all the discovered blocks are greater than or equal to size of 6 (on the query side) by 3 (on the array side). The blocks discovered are compared with blocks obtained by running XMOD on degree distribution preserved randomized genetic interaction network (see Bellay *et al.[9]*), and we use a p-value of 0.001 to filter out insignificant blocks. The discovered set of blocks may be overlapping, so we remove blocks which share an overlap of 10% or more with a larger block.

5.5.8 Gene Ontology

Please refer to Methods and Materials in Chapter 4 for details.

5.5.9 Genetic Interactions

Please refer to Methods and Materials in Chapter 4 for details.

Chapter 6: Conclusion and Future work

6.1 Conclusion

My main contributions in my Ph.D. have been to develop computational methods and approaches to solve biological problems including interpreting cross-species expression data in a network context (neXus; Chapter 2), discovering cancer targets in human using genomic data from model organisms (Chapter 3) , and optimizing screening strategies for chemical genomics experiments (COMPRESS-GI; Chapter 5). Further, I have conducted systematic comparison of different profile similarity measures on genetic interaction networks, frequently used by our lab and the broader community, for their utility under various noise and batch effect conditions (Chapter 4).

Portions of my research contribute to a larger personalized medicine strategy, which is key to treating several diseases that cannot be treated using conventional approaches. Cancer is a disease that particularly falls into this domain and this is why treatments for it have been elusive for decades. Instead, these diseases need to be solved on an individualized basis. The cancer synthetic lethality idea to discover cancer targets (Chapter 3) is a step towards a personalized approach for treating cancer. Based on this idea, the personalized medicine strategy would be to sequence the tumor(s), identify the spectrum of mutations associated with the tumor, predict the genes that would specifically sensitize tumor cells, and treat the cancer using an appropriate drug that inhibits these target proteins. With the dramatic reduction in whole genome sequencing costs, this strategy is becoming increasingly feasible, but there are several key missing pieces: firstly, we need more experimentally mapped genetic interactions in human cells, which will enable a more reliable prediction of target genes, and secondly, we need a comprehensive library of drugs for which the precise molecular targets are characterized. In this dissertation, due to the lack of genetic interaction data in the human, I have attempted to translate genetic interaction data from yeast to human; however, screening genetic interactions in human will cover substantially more human specific genes, and will likely be more effective as technology for genetic perturbations matures. Screening genetic interactions in human cell lines is not a trivial undertaking, which is one of the motivations for my work on optimized strategies for genetic interaction screens. I have also contributed to a large scale natural compound screening against the diagnostic set of genes I discovered in Chapter 5, which will lead to drug-target characterization for many compounds.

One of the key lessons I have learned over the course of my Ph.D. from several of my projects is that setting up appropriate random controls is key to successful methods development in computational biology. Random approaches are often a surprisingly good baseline strategy for many computational biology applications. For example, we showed in Chapter 5 that a random subset of 500 selected genes performs nearly as well as a complete dataset at recovering a profile similarity network from genetic interactions. Also, in the active subnetwork discovery project (Chapter 2), we showed that subnetworks can be discovered even from sets of genes with randomized scores. Random baselines are especially critical for algorithm development in the network setting because genes in the networks are often closely interconnected simply due the small-world nature of most biological networks. Thus, random baseline should be incorporated from the first steps of analysis or algorithm development for such projects.

During the development of the COMPRESS-GI algorithm, I found that I get results that are no better than random when I used the already established genetic interaction profile similarity measure, Pearson correlation. This observation led me to systematically compare many profile similarity measures for their utility on genetic interactions described in Chapter 4. I discovered that dot product, one of the simplest profile similarity measures performs better than some of the most sophisticated measures, including similarity measures developed specifically for genetic interaction data. Sometimes revisiting our basic assumptions is required before constructing more sophisticated algorithms.

6.2 Future work

6.2.1 Network-based discovery of coherent subnetworks in time-series genomics data

neXus was developed to address the growing need to compare and interpret large amounts of genomics data in multiple species using networks as the scaffold for comparison. A fruitful next step would be to extend the algorithm to compare time series genomics data using networks to discover functional modules that coherently vary with time. Such a method could be applied on time-series cancer expression data and will be useful to discover functional modules whose expression is tightly controlled. Also the method could be modified to discover modules that are highly variable which would be interesting to discover functional modules that have been disrupted in the condition. These different functional modules will provide insight for understanding the biological processes and will also be helpful for generating hypotheses to conduct experiments.

6.2.2 Computational methods for discovering mechanism of action of chemical compounds

Trial and error has been the primary way compounds with interesting properties have been discovered, and, in most cases, only after compounds are recognized to be of therapeutic value is the mechanism of action investigated. In many cases, including for widely used drugs, the mechanism of action is still not fully characterized. However, the drug discovery process has slowed significantly and become increasingly expensive, which has increased the need new approaches to drug discovery that can provide functional information for the compounds' activity without extensive and expensive tests. To address this need, we are currently screening thousands of natural compounds against a diagnostic library of yeast deletion mutant strains identified using the COMRPESS-GI method described in Chapter 5. We have screened around 10,000 naturally occurring and synthetic compounds using Bar-seq method [170], and we are currently working on predicting mechanism of actions for these compounds. We have already developed a pipeline to process the Bar-seq data; however, there are still few systematic effects, one of which is related to the multiplex tags primers used in PCR step to code the conditions, which need to be corrected. After normalization is done, we will predict the targets for the compounds and validate the most confident predictions in yeast and human cell lines. Though the basic principle for predicting mechanism of action from chemical genomic screens was described in [34], the method needs to reevaluated and most likely, a new method will need to be developed for a large-scale chemical genomics dataset generated by the Bar-seq approach. We are also planning to conduct chemical genomic screening in other organisms in the immediate future and one of the longer term goals is to conduct chemical genomics experiments in human cell lines.

6.2.3 Computational analyses for discovering rules governing conservation of genetic interactions

In chapter 3, we discussed the prediction of genetic interactions in human from data in model organisms to discover new cancer drug targets. We prioritized strong synthetic lethal genetic interactions in yeast between human orthologs which did not have paralogs in human. However, we only tested a very small set of interactions in human, which did not provide a true estimation of the level of conservation of genetic interactions between yeast and human. Testing of a larger set of interactions will enable the discovery of rules for genetic interaction in addition to providing statistics for rates of yeast-human genetic interaction conservation. Further, the interaction testing could be expanded to human genes with paralogs and conservation rules related to them could also be understood. Discovery of conservation rules will enable a better translation of large amounts of genomics data in yeast and several other model organisms for applications in human and will also generate a better understanding of how genetic interactions evolve.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
2. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, et al. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2: 2006 0008.
3. Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387-391.
4. Kim DU, Hayles J, Kim D, Wood V, Park HO, et al. (2010) Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nat Biotechnol 28: 617-623.
5. Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, et al. (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science 320: 362-365.
6. Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, et al. (2005) High-dimensional and large-scale phenotyping of yeast mutants. Proc Natl Acad Sci U S A 102: 19015-19020.
7. Mani R, St Onge RP, Hartman JLt, Giaever G, Roth FP (2008) Defining genetic interaction. Proc Natl Acad Sci U S A 105: 3461-3466.
8. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol 23: 561-566.
9. Bellay J, Atluri G, Sing TL, Toufighi K, Costanzo M, et al. (2011) Putting genetic interactions in context through a global modular decomposition. Genome Res 21: 1375-1387.
10. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, et al. (2010) The genetic landscape of a cell. Science 327: 425-431.
11. Ryan CJ, Roguev A, Patrick K, Xu J, Jahari H, et al. (2012) Hierarchical Modularity and the Evolution of Genetic Interactomes across Species. Mol Cell 46: 691-704.
12. Butland G, Babu M, Diaz-Mejia JJ, Bohdana F, Phanse S, et al. (2008) eSGA: E. coli synthetic genetic array analysis. Nat Methods 5: 789-795.
13. Roguev A, Talbot D, Negri GL, Shales M, Cagney G, et al. (2013) Quantitative genetic-interaction mapping in mammalian cells. Nat Methods 10: 432-437.
14. Laufer C, Fischer B, Billmann M, Huber W, Boutros M (2013) Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. Nat Methods 10: 427-431.
15. Marcotte R, Brown KR, Suarez F, Sayad A, Karamboulas K, et al. (2012) Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov 2: 172-189.
16. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141-147.

17. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637-643.

18. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, et al. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322: 104-110.

19. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403: 623-627.

20. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98: 4569-4574.

21. Kaiser P, Meierhofer D, Wang X, Huang L (2008) Tandem affinity purification combined with mass spectrometry to identify components of protein complexes. Methods Mol Biol 439: 309-326.

22. Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, et al. (2008) An in vivo map of the yeast protein interactome. Science 320: 1465-1470.

23. Middendorf M, Ziv E, Wiggins CH (2005) Inferring network mechanisms: the Drosophila melanogaster protein interaction network. Proc Natl Acad Sci U S A 102: 3192-3197.

24. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3: 140.

25. Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, et al. (2006) Large-scale identification of protein-protein interaction of Escherichia coli K-12. Genome Res 16: 686-691.

26. Wu G, Feng X, Stein L (2010) A human functional protein interaction network and its application to cancer data analysis. Genome Biol 11: R53.

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

28. Koch EN, Costanzo M, Bellay J, Deshpande R, Chatfield-Reed K, et al. (2012) Conserved rules govern genetic interaction degree across species. Genome Biol 13: R57.

29. Yellaboina S, Goyal K, Mande SC (2007) Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: comparison with high-throughput experimental data. Genome Res 17: 527-535.

30. Costello JC, Dalkilic MM, Beason SM, Gehlhausen JR, Patwardhan R, et al. (2009) Gene networks in Drosophila melanogaster: integrating experimental data to predict gene function. Genome Biol 10: R97.

31. Guan Y, Myers CL, Lu R, Lemischka IR, Bult CJ, et al. (2008) A genomewide functional network for the laboratory mouse. PLoS Comput Biol 4: e1000165.

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

33. Parsons A, Lopez A, Givoni I, Williams D, Gray C, et al. (2006) Exploring the Mode-of-Action of Bioactive Compounds by Chemical-Genetic Profiling in Yeast. Cell 126: 611-625.

34. Parsons AB, Brost RL, Ding H, Li Z, Zhang C, et al. (2004) Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nat Biotechnol 22: 62-69.

35. Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. Nature 411: 1046-1049.

36. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314: 1041-1052.

37. Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, et al. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38: D196-203.

38. Remm M, Sonnhammer E (2000) Classification of transmembrane protein families in the Caenorhabditis elegans genome and identification of human orthologs. Genome Res 10: 1679-1689.

39. O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33: D476-480.

40. Schulze A, Downward J (2001) Navigating gene expression using microarrays [mdash] a technology review. Nat Cell Biol 3: E190-E195.

41. Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18 Suppl 1: S233-240.

42. Ulitsky I, Shamir R (2007) Identification of functional modules using network topology and high-throughput data. BMC Syst Biol 1: 8.

43. Ulitsky I, Shamir R (2009) Identifying functional modules using expression profiles and confidence-scored protein interactions. Bioinformatics 25: 1158-1164.

44. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 24: i223-231.

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

46. Deshpande R, Sharma S, Verfaillie CM, Hu WS, Myers CL (2010) A scalable approach for discovering conserved active subnetworks across species. PLoS Comput Biol 6: e1001028.

47. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98: 5116-5121.

48. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530-536.

49. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881-885.

50. Rajagopalan D, Agarwal P (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. Bioinformatics 21: 788-793.

51. Guo Z, Wang L, Li Y, Gong X, Yao C, et al. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. Bioinformatics 23: 2121-2128.

52. Cabusora L, Sutton E, Fulmer A, Forst CV (2005) Differential network expression during drug and stress response. Bioinformatics 21: 2898-2905.

53. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, et al. (2004) PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res 32: W83-88.

54. Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S (2006) Graemlin: general and robust alignment of multiple large interaction networks. Genome Res 16: 1169-1181.

55. Peregrin-Alvarez JM, Xiong X, Su C, Parkinson J (2009) The Modular Organization of Protein Interactions in Escherichia coli. PLoS Comput Biol 5: e1000523.

56. Simonis N, Rual JF, Carvunis AR, Tasan M, Lemmens I, et al. (2009) Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. Nat Methods 6: 47-54.

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

58. Linghu B, Snitkin ES, Hu Z, Xia Y, Delisi C (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol 10: R91.

59. Ihmels J, Bergmann S, Berman J, Barkai N (2005) Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program. PLoS Genet 1: e39.

60. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957-968.

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

62. Adewumi O, Aflatoonian B, Ahrlund-Richter L, Amit M, Andrews PW, et al. (2007) Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. Nat Biotechnol 25: 803-816.

63. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788-791.

64. Harrington L, McPhail T, Mar V, Zhou W, Oulton R, et al. (1997) A mammalian telomerase-associated protein. Science 275: 973-977.

65. Li H, Zhao LL, Yang ZY, Funder JW, Liu JP (1998) Telomerase is controlled by protein kinase C alpha in human breast cancer cells. Journal of Biological Chemistry 273: 33436-33442.

66. Weber GF, Ashkar S, Glimcher MJ, Cantor H (1996) Receptor-ligand interaction between CD44 and osteopontin (Eta-1). Science 271: 509-512.

67. Botquin V, Hess H, Fuhrmann G, Anastassiadis C, Gross MK, et al. (1998) New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. Genes Dev 12: 2073-2090.
68. Chen S, Choo A, Chin A, Oh SK (2006) TGF-beta2 allows pluripotent human embryonic stem cell proliferation on E6/E7 immortalized mouse embryonic fibroblasts. J Biotechnol 122: 341-361.
69. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, et al. (2004) Protein interaction networks from yeast to human. Curr Opin Struct Biol 14: 292-299.
70. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947-956.
71. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, et al. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet 38: 431-440.
72. Rappolee DA, Basilico C, Patel Y, Werb Z (1994) Expression and function of FGF-4 in peri-implantation development in mouse embryos. Development 120: 2259-2269.
73. Yuan H, Corbi N, Basilico C, Dailey L (1995) Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. Genes Dev 9: 2635-2645.
74. Feldman B, Poueymirou W, Papaioannou VE, DeChiara TM, Goldfarb M (1995) Requirement of FGF-4 for postimplantation mouse development. Science 267: 246-249.
75. Kurimoto K, Yabuta Y, Ohinata Y, Ono Y, Uno KD, et al. (2006) An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. Nucleic Acids Res 34: e42.
76. Guo G, Huss M, Tong GQ, Wang C, Li Sun L, et al. (2010) Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell 18: 675-685.
77. Brons IG, Smithers LE, Trotter MW, Rugg-Gunn P, Sun B, et al. (2007) Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature 448: 191-195.
78. Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, et al. (2007) New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature 448: 196-199.
79. Kunath T, Saba-El-Leil MK, Almousailleakh M, Wray J, Meloche S, et al. (2007) FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. Development 134: 2895-2902.
80. Xu RH, Sampsell-Barron TL, Gu F, Root S, Peck RM, et al. (2008) NANOG is a direct target of TGFbeta/activin-mediated SMAD signaling in human ESCs. Cell Stem Cell 3: 196-206.
81. Xu RH, Peck RM, Li DS, Feng X, Ludwig T, et al. (2005) Basic FGF and suppression of BMP signaling sustain undifferentiated proliferation of human ES cells. Nat Methods 2: 185-190.

82. Greber B, Wu G, Bernemann C, Joo JY, Han DW, et al. (2010) Conserved and divergent roles of FGF signaling in mouse epiblast stem cells and human embryonic stem cells. Cell Stem Cell 6: 215-226.

83. Faast R, White J, Cartwright P, Crocker L, Sarcevic B, et al. (2004) Cdk6-cyclin D3 activity in murine ES cells is resistant to inhibition by p16(INK4a). Oncogene 23: 491-502.

84. Neganova I, Zhang X, Atkinson S, Lako M (2009) Expression and functional analysis of G1 to S regulatory components reveals an important role for CDK2 in cell cycle regulation in human embryonic stem cells. Oncogene 28: 20-30.

85. Fujii-Yamamoto H, Kim JM, Arai K, Masai H (2005) Cell cycle and developmental regulations of replication factors in mouse embryonic stem cells. J Biol Chem 280: 12976-12987.

86. Tarasov KV, Tarasova YS, Tam WL, Riordon DR, Elliott ST, et al. (2008) B-MYB is essential for normal cell cycle progression and chromosomal stability of embryonic stem cells. PLoS One 3: e2478.

87. Tanaka Y, Patestos NP, Maekawa T, Ishii S (1999) B-myb is required for inner cell mass formation at an early stage of development. J Biol Chem 274: 28067-28070.

88. Kong M, Barnes EA, Ollendorff V, Donoghue DJ (2000) Cyclin F regulates the nuclear localization of cyclin B1 through a cyclin-cyclin interaction. EMBO J 19: 1378-1388.

89. Lim LS, Loh YH, Zhang W, Li Y, Chen X, et al. (2007) Zic3 is required for maintenance of pluripotency in embryonic stem cells. Mol Biol Cell 18: 1348-1358.

90. Heo JS, Lee MY, Han HJ (2007) Sonic hedgehog stimulates mouse embryonic stem cell proliferation by cooperation of Ca2+/protein kinase C and epidermal growth factor receptor as well as Gli1 activation. Stem Cells 25: 3069-3080.

91. Mizugishi K, Aruga J, Nakata K, Mikoshiba K (2001) Molecular properties of Zic proteins as transcriptional regulators and their relationship to GLI proteins. J Biol Chem 276: 2180-2188.

92. Yang X, Meng X, Su X, Mauchley DC, Ao L, et al. (2009) Bone morphogenic protein 2 induces Runx2 and osteopontin expression in human aortic valve interstitial cells: role of Smad1 and extracellular signal-regulated kinase 1/2. J Thorac Cardiovasc Surg 138: 1008-1015.

93. Wang T, Tamakoshi T, Uezato T, Shu F, Kanzaki-Kato N, et al. (2003) Forkhead transcription factor Foxf2 (LUN)-deficient mice exhibit abnormal development of secondary palate. Dev Biol 259: 83-94.

94. Chatzi C, van den Brink CE, van der Saag PT, McCaig CD, Shen S (2010) Expression of a mutant retinoic acid receptor beta alters lineage differentiation in mouse embryonic stem cells. Stem Cells Dev 19: 951-960.

95. Hannenhalli S, Kaestner KH (2009) The evolution of Fox genes and their role in development and disease. Nat Rev Genet 10: 233-240.

96. Miyashita T, Hanashita T, Toriyama M, Takagi R, Akashika T, et al. (2008) Gene cloning and biochemical characterization of the BMP-2 of Pinctada fucata. Biosci Biotechnol Biochem 72: 37-47.

97. Mullor JL, Dahmane N, Sun T, Ruiz i Altaba A (2001) Wnt signals are targets and mediators of Gli function. Curr Biol 11: 769-773.
98. Parsons SJ, Parsons JT (2004) Src family kinases, key regulators of signal transduction. Oncogene 23: 7906-7909.
99. Anneren C, Cowan CA, Melton DA (2004) The Src family of tyrosine kinases is important for embryonic stem cell self-renewal. J Biol Chem 279: 31590-31598.
100. Sun Y, Li H, Liu Y, Mattson MP, Rao MS, et al. (2008) Evolutionarily conserved transcriptional co-expression guiding embryonic stem cell differentiation. PLoS One 3: e3406.
101. Ying QL, Nichols J, Chambers I, Smith A (2003) BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. Cell 115: 281-292.
102. Perlingeiro RC (2007) Endoglin is required for hemangioblast and early hematopoietic development. Development 134: 3041-3048.
103. Wallin R, Cain D, Hutson SM, Sane DC, Loeser R (2000) Modulation of the binding of matrix Gla protein (MGP) to bone morphogenetic protein-2 (BMP-2). Thromb Haemost 84: 1039-1044.
104. Gansner ER, Koren Y, North S (2004) Graph drawing by stress majorization. Graph Drawing 3383: 239-250.
105. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.
106. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA (2008) The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res 36: D724-728.
107. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (1997) GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13: 163.
108. Affymetrix (2002) Statistical Algorithm Description Document. Tech Rep Available: http://wwwaffymetrixcom/support/technical/whitepapersaffx.
109. Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 101: 4164-4169.
110. Lopez-Fernandez L, Robles G, Gonzalez-Barahona J (2004) Applying Social Network Analysis to the Information in CVS Repositories. In Proceedings of the International Workshop on Mining Software Repositories: 101-105.
111. Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res 36: D263-266.
112. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710-3715.
113. Hartwell LH, Szankasi P, Roberts CJ, Murray AW, Friend SH (1997) Integrating genetic approaches into the discovery of anticancer drugs. Science 278: 1064-1068.

114. Tutt A, Robson M, Garber JE, Domchek SM, Audeh MW, et al. (2010) Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial. Lancet 376: 235-244.

115. Scholl C, Frohling S, Dunn IF, Schinzel AC, Barbie DA, et al. (2009) Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. Cell 137: 821-834.

116. Luo J, Emanuele MJ, Li D, Creighton CJ, Schlabach MR, et al. (2009) A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. Cell 137: 835-848.

117. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, et al. (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462: 108-112.

118. Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, et al. (2005) Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. Nature 434: 913-917.

119. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, et al. (2009) Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. N Engl J Med 361: 123-134.

120. Cheung HW, Cowley GS, Weir BA, Boehm JS, Rusin S, et al. (2011) Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc Natl Acad Sci U S A 108: 12372-12377.

121. Iorns E, Lord CJ, Turner N, Ashworth A (2007) Utilizing RNA interference to enhance cancer drug discovery. Nat Rev Drug Discov 6: 556-568.

122. McManus KJ, Barrett IJ, Nouhi Y, Hieter P (2009) Specific synthetic lethal killing of RAD54B-deficient human colorectal cancer cells by FEN1 silencing. Proc Natl Acad Sci U S A 106: 3276-3281.

123. Welch BL (1947) The generalisation of student's problems when several different population variances are involved. Biometrika 34: 28-35.

124. Satterthwaite FE (1946) An approximate distribution of estimates of variance components. Biometrics 2: 110-114.

125. Baryshnikova A, Costanzo M, Dixon S, Vizeacoumar FJ, Myers CL, et al. (2010) Synthetic genetic array (SGA) analysis in Saccharomyces cerevisiae and Schizosaccharomyces pombe. Methods Enzymol 470: 145-179.

126. Amberg DC, Burke DJ, Strathern JN (2006) Tetrad dissection. CSH Protoc 2006.

127. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. (2004) A census of human cancer genes. Nat Rev Cancer 4: 177-183.

128. VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, et al. (2010) Genetic interactions reveal the evolutionary trajectories of duplicate genes. Mol Syst Biol 6: 429.

129. DeCristofaro MF, Betz BL, Wang W, Weissman BE (1999) Alteration of hSNF5/INI1/BAF47 detected in rhabdoid cell lines and primary rhabdomyosarcomas but not Wilms' tumors. Oncogene 18: 7559-7565.

130. Liu Y, Liu P, Wen W, James MA, Wang Y, et al. (2009) Haplotype and cell proliferation analyses of candidate lung cancer susceptibility genes on chromosome 15q24-25.1. Cancer Res 69: 7844-7850.

131. Modena P, Lualdi E, Facchinetti F, Galli L, Teixeira MR, et al. (2005) SMARCB1/INI1 tumor suppressor gene is frequently inactivated in epithelioid sarcomas. Cancer Res 65: 4012-4019.

132. Schneppenheim R, Fruhwald MC, Gesk S, Hasselblatt M, Jeibmann A, et al. (2010) Germline nonsense mutation and somatic inactivation of SMARCA4/BRG1 in a family with rhabdoid tumor predisposition syndrome. Am J Hum Genet 86: 279-284.

133. Lee RS, Stewart C, Carter SL, Ambrogio L, Cibulskis K, et al. (2012) A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. J Clin Invest 122: 2983-2988.

134. Weissman B, Knudsen KE (2009) Hijacking the chromatin remodeling machinery: impact of SWI/SNF perturbations in cancer. Cancer Res 69: 8223-8230.

135. Reisman D, Glaros S, Thompson EA (2009) The SWI/SNF complex and cancer. Oncogene 28: 1653-1668.

136. Folpe AL, Deyrup AT (2006) Alveolar soft-part sarcoma: a review and update. J Clin Pathol 59: 1127-1132.

137. Ladanyi M, Lui MY, Antonescu CR, Krause-Boehm A, Meindl A, et al. (2001) The der(17)t(X;17)(p11;q25) of human alveolar soft part sarcoma fuses the TFE3 transcription factor gene to ASPL, a novel gene at 17q25. Oncogene 20: 48-57.

138. Hideshima T, Richardson P, Chauhan D, Palombella VJ, Elliott PJ, et al. (2001) The proteasome inhibitor PS-341 inhibits growth, induces apoptosis, and overcomes drug resistance in human multiple myeloma cells. Cancer Res 61: 3071-3076.

139. Nijhawan D, Zack TI, Ren Y, Strickland MR, Lamothe R, et al. (2012) Cancer vulnerabilities unveiled by genomic loss. Cell 150: 842-854.

140. Deshpande R, VanderSluis B., Myers, C.L. (2013) Comparison of Profile Similarity Measures for Genetic Interaction Networks. PLoS ONE 8(7): e68664.

141. Avery L, Wasserman S (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. Trends Genet 8: 312-316.

142. Baryshnikova A, Costanzo M, Kim Y, Ding H, Koh J, et al. (2010) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nat Methods 7: 1017-1024.

143. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, et al. (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature 446: 806-810.

144. Roguev A, Bandyopadhyay S, Zofall M, Zhang K, Fischer T, et al. (2008) Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. Science 322: 405-410.

145. Dixon SJ, Fedyshyn Y, Koh JL, Prasad TS, Chahwan C, et al. (2008) Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. Proc Natl Acad Sci U S A 105: 16653-16658.

146. Typas A, Nichols RJ, Siegele DA, Shales M, Collins SR, et al. (2008) High-throughput, quantitative analyses of genetic interactions in E. coli. Nat Methods 5: 781-787.
147. Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, et al. (2004) Genome-wide RNAi analysis of growth and viability in Drosophila cells. Science 303: 832-835.
148. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nat Genet 38: 896-903.
149. Falkowski BJ (1998) On certain generalizations of inner product similarity measures. Journal of the American Society for Information Science 49: 854-858.
150. Wang ZW, Wong SKM, Yao YY (1992) An analysis of vector space models based on computational geometry. Proceeding SIGIR '92 Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval.
151. Leydesdorff L (2008) On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. Journal of the American Society for Information Science & Technology 59: 77-85.
152. Egghe L, Leydesdorff L (2009) The relation between Pearson's correlation coefficient r and Salton's cosine measure. Journal of the American Society for Information Science & Technology 60: 1027-1036.
153. Adler J, Parmryd I (2010) Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient. Cytometry A 77: 733-742.
154. Haranczyk M, Holliday J (2008) Comparison of similarity coefficients for clustering and compound selection. J Chem Inf Model 48: 498-508.
155. Yona G, Dirks W, Rahman S, Lin DM (2006) Effective similarity measures for expression profiles. Bioinformatics 22: 1616-1622.
156. Dalirsefat SB, da Silva Meyer A, Mirhoseini SZ (2009) Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, Bombyx mori. J Insect Sci 9: 1-8.
157. Obayashi T, Kinoshita K (2009) Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression. DNA Research 16: 249-260.
158. Leiserson MD, Tatar D, Cowen LJ, Hescott BJ (2011) Inferring mechanisms of compensation from E-MAP and SGA data using local search algorithms for max cut. J Comput Biol 18: 1399-1409.
159. Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2006) A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biol 7: R63.
160. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, et al. (2011) Detecting novel associations in large data sets. Science 334: 1518-1524.
161. Gillis J, Pavlidis P (2011) The impact of multifunctional genes on "guilt by association" analysis. PLoS One 6: e17258.

162. Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG (2006) Finding function: evaluation methods for functional genomic data. BMC Genomics 7: 187.

163. Pearson K (1920) Notes on the History of Correlation. Biometrika 13: 25-45.

164. Schechtman E, Yitzhaki S (1999) On the proper bounds of the Gini correlation. Economics Letters 63: 133-138.

165. de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20: 1453-1454.

166. Page RDM (1996) Tree View: An application to display phylogenetic trees on personal computers. Computer applications in the biosciences : CABIOS 12: 357-358.

167. Casey FP, Cagney G, Krogan NJ, Shields DC (2008) Optimal stepwise experimental design for pairwise functional interaction studies. Bioinformatics 24: 2733-2739.

168. Mallat SG, Zhifeng Z (1993) Matching pursuits with time-frequency dictionaries. Signal Processing, IEEE Transactions on 41: 3397-3415.

169. Parsons AB, Brost RL, Ding H, Li Z, Zhang C, et al. (2004) Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nat Biotech 22: 62-69.

170. Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, et al. (2009) Quantitative phenotyping via deep barcode sequencing. Genome Res 19: 1836-1842.

Appendix A: Appendix for chapter 2

Note 1: Implications of using functional linkage vs. physical interaction networks for active subnetwork discovery

While previous work has focused on subnetwork discovery in the context of physical (protein-protein) interaction networks, we propose here to use functional linkage networks, which are now available in a range of organisms from yeast to human. There are a number of implications of this decision, which we briefly discuss here. In general, protein-protein interactions reflect direct physical interactions between proteins, which means that they have the ability to capture stably bound protein complexes or potentially interactions that mediate signaling events. On the other hand, functional linkage networks are constructed by incorporating a much broader range of relationships: for example, physical interactions, co-expression, shared regulatory motifs, common protein localization patterns, or shared phylogenetic relationships. All of these are indicative of a function in a common biological process, and integration methods weight the input datasets according to their ability to predict known relationships, but the resulting linkages are less direct, and often less functionally specific. The broader relationships captured by functional linkage networks provide both unique advantages as well as additional challenges in the context of active subnetwork discovery.

In terms of advantages, they can potentially reveal broader functional modules or pathways which are not directly connected by physical protein-protein interactions, but are nonetheless functionally coherent and show similar patterns of differential expression. Protein-protein interaction networks will miss a large number of relationships captured by a functional linkage network, and thus, may not allow detection of the same coherently expressed modules. For example, it is not likely that transcription factors will show up in the same module as downstream targets in networks that truly reflect physical interactions between proteins, while this is common in subnetworks identified using functional linkage networks.

Functional linkage networks also introduce new challenges to the subnetwork discovery problem. Due to the generality of relationships captured by these networks, interpretation of the resulting active modules is often more difficult. They can include relationships based on a variety of underlying physical or genetic evidence, so even when coherent functions are represented among a set of genes, the underlying mechanisms supporting the putative relationships are not always clear. Each module must be investigated individually to assess the underlying evidence for the relationships before clear hypotheses can be formed. The higher density of functional linkage networks, which can be orders of magnitude more dense, may also present more technical problems in the discovery of active subnetworks. For example, the probability of finding dense subnetworks among even randomly chosen genes increases with the increased density of linkages. The higher density also adds complexity to the subnetwork discovery problem, which may require new algorithms as we discuss here.

In this work, we have chosen to use functional linkage networks as the basis of our approach because we feel that the added sensitivity to a broader range of functional relationships and availability of comprehensive functional linkage networks for several organisms, including human and mouse, are major advantages. Particularly when relatively independent expression data and networks are available in related species, the more comprehensive coverage offered by functional linkage networks can be a major advantage. Furthermore, given that our current understanding of physical protein-protein interactions is still limited in higher eukaryotes, functional relationships inferred from patterns in genomic data can offer a powerful approach to discovering new biology.

Note 2: neXus applied to single dataset differential expression study

The main contribution of this work is the cross-species subnetwork search algorithm, which is completely independent of the method for generating gene lists and activity scores. To illustrate the application of our algorithm, we compiled a large compendium of stem cell expression data for both mouse and human and derived a set of differentially expressed genes as described in Materials and Methods. However, to demonstrate that the search algorithm is independent of the differential expression analysis method, we also ran our cross-species search algorithm on gene lists derived from a simple application of SAM (Significance Array of Microarrays [1]) to a single mouse expression dataset (GSE 3653) [2] and a single human expression dataset (GSE 9940) [3]. We also randomized the resulting expression values and searched for subnetworks on the randomized data for comparison (Figure 2.2). The conclusion from this analysis is similar to that from the analysis of our original differential expression list: our approach is able to find many significant subnetworks from the real differential expression list but very few based on the randomized differential expression data. For example, at an activity score cutoff of 0.2, our approach discovers 48 subnetworks on the real differential expression values but an average of 2 on the randomized data (Figure 2.2).

Note 3: Independence of the datasets

An important issue that affects the significance of active subnetworks discovered is the independence of the various input datasets, including the relationship between the differential expression data and the functional networks within each species as well as the relationship between the functional networks across species. Because expression data is one of the major sources of input data for constructing both the mouse and human functional linkage networks, we checked whether the datasets we compiled for our stem cell differential expression analysis overlapped with those used in constructing the mouse and human functional linkage networks. None of the 20 mouse datasets (Supplemental Table S3) or 13 human datasets (Supplemental Table S4) used for our differential expression analysis were used to construct the mouse or human functional linkage networks. Thus, these data are independent. With regard to the independence of the mouse and human functional linkage network, the mouse network was constructed first (2008), and was not used as input in constructing the human functional linkage network. The human network incorporates physical and genetic interactions, sequence information (shared protein domains, transcription factor binding sites), and gene expression profiles [4]. The mouse network incorporates physical interaction data, shared phenotype data, phylogenetic profile information, the yeast functional network where orthologs exist, and gene expression information [5]. Both resulted from naïve Bayes classifiers that were trained using Gene Ontology annotations, which are of course not independent between mouse and human, but we would argue that this is also true of any interaction network available for mouse or human, so it is not an easy issue to avoid. Another source of dependence between the mouse and human network is that the mouse functional network incorporates putative protein interactions from the Online Predicted Human Interaction Database (OPHID), which were mapped from human orthologs (commonly referred to as interologs). OPHID itself was not directly used in constructing the human functional linkage network, but is based on several of the protein-protein interactions that were. Thus, the physical interaction data incorporated in the mouse network are not independent of the physical interactions incorporated in the human network. For our analysis in this paper, we have chosen to keep the mouse network as originally constructed because of the limited availability of mouse-specific interactions. We should note this dependence between the two functional networks is accounted for in the randomization procedure, which we used to statistically validate our results (see Figure 2.3). For all randomization experiments, the functional linkage networks were held fixed in both species (only the differential expression values are randomized), so whatever dependence exists should also help increase the number of networks discovered during random instances. Our approach recovers ~20-fold more subnetworks than the average random run (see Figure 2.3), suggesting that the algorithm is accomplishing something useful even if the human and mouse functional networks are not completely independent.

Note 4: Comparison of the overlap of mouse and human subnetworks discovered through MATISSE and neXus

To check whether single-species approaches could be used to discover conserved active subnetworks, we applied MATISSE [6], the existing method for single-species network discovery with the best performance (see Results, Figure 2.9). Both approaches were applied independently to the mouse and human differential expression data and functional linkage networks. As discussed in the Results, we were not able to apply

MATISSE to the complete mouse or human functional linkage network, so we reduced the size of the networks to the size where we could load and run the algorithm on the networks, which was to 50,000 edges (7909 genes) and 25,000 edges (9281 genes) for mouse and human, respectively. For both, the edges were restricted to the highest weight edges. We then ran MATISSE independently on these reduced networks and the corresponding expression profiles. Mouse subnetworks and human subnetworks were then compared for overlap to assess how it compared to our cross-species algorithm. In terms of the number of genes, the mouse and human subnetworks covered around three thousand genes (Table S1), roughly half of which were orthologs (Table S1). However, the MATISSE subnetworks showed relatively low agreement in terms of the sets of genes present. Only a single pair was found to have a Jaccard index ($A \cap B/A \cup B$) greater than 0.2, suggesting that even where orthologs overlap, the sets of modules discovered in human and mouse are quite distinct (Table S2). In contrast, our algorithm is designed to find completely overlapping subnetworks. Thus, we find nearly 100% overlap in terms of the genes covered in our approach, and all subnetworks have a Jaccard index near 1 (Table S2). We should note that this is not surprising given the fact that our search algorithm identifies networks on orthologous genes in parallel, but this result does demonstrate the utility of this type of search in the sense that it enables a more direct comparison of the human and mouse expression patterns

Note 5: Other Randomizations

In addition to the randomization scheme described in the Results section, which involves shuffling the differential expression values in both species, we evaluated three other schemes as well: randomizing differential expression values in only mouse, randomizing differential expression values in only human, and randomizing the orthology links between mouse and human. At the same parameters at which we discover 255 real subnetworks (mouse and human clustering coefficient = 0.1 and 0.2 respectively and network score cutoff = 0.15), we found an average of ~11 with our original randomization approach, an average of ~30 with the mouse-only randomization, an average of ~24 with the human-only randomization, and an average of ~3 with the orthology randomization (2.14). Even by the most conservative randomization scheme, our approach finds ~10-fold more real networks than random