Using interspecies biological networks to guide drug therapy

Alexandra Jacunski

Submitted in partial fulfillment of the requirements for

the degree of Doctor of Philosophy under the Executive

Committee of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Using interspecies biological networks to guide drug therapy

Alexandra Jacunski

The use of drug combinations (DCs) in cancer therapy can prevent the development of drug resistance and decrease the severity and number of side effects. Synthetic lethality (SL), a genetic interaction wherein two nonessential genes cause cell death when knocked out simultaneously, has been suggested as a method of identifying novel DCs. A combination of two drugs that mimic genetic knockout may cause cellular death through a synthetic lethal pathway. Because SL can be context-specific, it may be possible to find DCs that target SL pairs in tumours while leaving healthy cells unscathed.

However, elucidating all synthetic lethal pairs in humans would take more than 200 million experiments in a single biological context – an unmanageably large search space. It is thus necessary to develop computational methods to predict human SL.

In this thesis, we develop connectivity homology, a novel measure of network similarity that allows for the comparison of interspecies protein-protein interaction networks. We then use this principle to develop Species-INdependent TRAnslation (SINaTRA), an algorithm that allows us to predict SL between species using protein-protein interaction networks. We validate it by predicting SL in *S. pombe* from *S. cerevisiae*, then generate over 100 million SINaTRA scores for putative human SL pairs. We use these data to predict new areas of cancer combination therapy, and then test fifteen of these predictions across several cell lines. Finally, in order to better understand synergy, we develop DAVISS (Data-driven Assessment of Variability In Synergy Scores), a novel way to statistically evaluate the significance of a drug interaction.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Page

DEDICATION

For my mom, who taught me that being prepared is half the victory.
For my dad, who taught me to cross the bridge when I get to it.
With thanks to Walt Whitman, who taught me to live with contradictions.
(I am large, I contain multitudes.)

CHAPTER 1 – INTRODUCTION

To begin, we will introduce three concepts that are necessary to understand this body of research: systems biology, synthetic lethality, and drug synergy. We will then describe how we integrate these sections by outlining the results presented in subsequent chapters of this thesis.

SYSTEMS BIOLOGY

The need for systems biology

In a 2002 article [1], Yuri Lazebnik asked a simple question – "Can a biologist fix a radio?" – in order to illustrate the methodology of 'traditional,' experimental biology. When presented with a broken radio, a biologist would acquire a number of functioning radios and try to replicate the problem by breaking or removing different components within it. This can be called a reductionist, or bottom-up, approach: to understand
a complex problem, one must take it apart and understand how the parts interact.

There are many benefits to reductionism, including a detailed understanding of all moving parts within the system. Furthermore, reductionist methodologies have been essential to a number of important biological discoveries, especially pertaining to Mendelian diseases.

However, in certain cases, reductionism is insufficient. For example, the removal of several different components may cause the same problem; the issue may not be replicable without altering a number of parts simultaneously; and tunable components may be involved, making the search space infinitely complex. Furthermore, there may be factors that unnecessarily confound the analysis, such as whether the colour of a part alters its function. In short, simply cataloguing the function and importance of each part will be a time-consuming process, and may only help fix a subset of all possible problems. Intuitively, we know that a high-level understanding of how

a radio functions — for example, understanding the schematic — will make the process of fixing it easier.

Removing components in a radio has a parallel in biology: genetic knockouts. Applying a reductionist approach to human biology can result in similar drawbacks to what we observe in radios. First, multiple proteins may be essential to a process, and knocking out any one will cause a disease; for example, mutations in a number of different genes can cause Seckel syndrome [2]. Furthermore, complex diseases, such as cancer, are caused by the malfunction of multiple genes simultaneously [3]; thus, finding the correct combination requires an extremely large number of experiments – something near impossible without guidance or sheer luck. Some diseases show differences in penetrance or severity depending on a variety of factors aside from genetics, such as the presence of short-term gestational hypoxia as a potential trigger for the genesis of scoliosis [4]. As with the radio, these cases are too complex to approach with a traditional knock-out approach.

A schematic of human systems would be useful in such cases, and systems biology attempts to provide just such a map. It can be used to compliment traditional biological experiments, providing a top-down perspective to guide and inform our understanding of human health.

An introduction to systems biology

Systems biology is an interdisciplinary field of study that focuses on the "bird's-eye view" of biology. Here, a system can be defined as a set of relationships between biological concepts. They can be as large-scale as the relationships between diseases, or as specific as the interconnected metabolic pathways of a cell. At its core, systems biology is about providing relational structure to observations, such that no single datum lives in a vacuum.

Importantly, systems biology integrates work from areas such as biology, computer science, statistics, physics, and bioinformatics. Unifying these fields allows researchers to use existing

data to develop predictive biological models, which are then used to generate hypotheses that can be tested in experimental settings. Arguably the best work in systems biology establishes a consistent positive feedback loop between the development of computational models, the creation of hypotheses based on these models, and the experimental testing thereof. This allows for the continuous refinement of our understanding of a particular problem.

Another important aspect of systems biology is the integration of new technology. Although systems biology isn't synonymous with Big Data, much of the work in the field has used new technological developments in areas such as sequencing and large-scale assays to develop multidimensional models of human systems.

Finally, systems biology represents another step towards a new paradigm of looking at human health: from *reactive* to *predictive* [5]. Historically, the treatment of human disease has shifted from purely curative, relying on the identification and treatment of symptoms, to preventative, such as the use of vaccines. With the advent of genomic sequencing, it has added a predictive element, such as in prenatal and carrier testing. The integrative nature of systems biology means that patient-specific data can be incorporated into models, allowing for individualized predictions for their health [6].

Work in systems biology has influenced the study of cancer, viruses, and neuroscience, among many others. For example, one paper differentiated gene mutations causative to, rather than simply associated with, cancer by identifying the frequency of genetic interaction within biological subsystems [7]. Another investigated the aetiology of viral disease by integrating virus-host and host-host interactions to understand how viruses manipulate host cell machinery [8]. A third looked at the topology of an individual's brain during MRI scans to predict subjects at high risk for schizophrenia [9]. Although these publications are vastly different, they hold one

particular tenet of systems biology in common: the use of networks to represent the relationships between elements of a system.

Network theory

Networks, also known as graphs, are highly versatile, visual representations of connections among data. Nodes denote objects, and the edges that join them symbolize associations. They can be used to depict almost any type of relationship across almost any area of study: social interactions, as in the famous 'small world' experiment that led to the idea of six degrees of separation [10]; the Internet, and how Google returns search results [11]; and the inner metabolic workings of a cell [12].

In brief, a node with *n* connections is of degree *n*; more edges signify higher degree. Causal relationships are represented with the use of directed edges, while relationships that vary in magnitude can be indicated using weighted edges.
Multigraphs may be used to depict relationships where two nodes are connected by more than one edge, or one node has an edge leading to itself.

One important measure of networks is distance. Two nodes connected by an edge are a distance of one step from each other; they are neighbours. If Node A connects only to B, and B connects only to C, then A and C are a distance of two steps from each other. Often, there are many paths between two nodes; in this case, the shortest path between them can be used to measure their distance. If there are no paths between two nodes, the distance between them is infinite.

The topology of the network can be used to infer information about its components. A list of common terms is available in **Table 1.1**. Two important topological characteristics are hubs and modules. Hubs are nodes that are considered central due to their high degree. Modules are highly connected subcomponents of a network;

a specific subtype of modules is a clique, in which every node is connected to every other node in that subset. Cliques offer a large benefit in analysis because of their closed nature; they are much less computationally intensive to find than modules, which are open subnetworks [13].

Term	Definition
Network	Also called a graph, it consists of nodes (e.g., genes) that are connected by defined relationships (e.g., coexpression) with edges.
Neighbour	Any node connected by an edge to the node of interest.
Bipartite network	A network in which nodes of one group (e.g., genes) are connected to nodes of another (e.g., diseases), but no within-group edges exist (i.e., no gene is connected to another gene).
Hub	A central, highly connected node within a network; often represents essential genes when applied to biological networks.
Clique	A subset of a module, in which all nodes are connected to all other nodes in the clique. A maximal clique is the largest clique that can be found within a given module.
Motif	A recurrent, statistically significant subgraph or pattern. In biology, these can include negative autoregulation, feed-forward loops, and so on. They can be particularly important in metabolic networks.
Scale-free network	The degrees of nodes in a network tend to be distributed according to a power law, such that a new edge being assigned to the graph tends to be given to a node of high degree. Biological networks tend to have this property.
Small-world property	Most nodes are not directly connected, but the majority of nodes can be reached from all others by crossing a relatively small number of edges. The strict definition states that the average path length is of the order of log(N), where N is the size of the network. Biological networks also tend to have this property.

Table 1.1: The vocabulary of networks

Network topology can also be defined mathematically, making networks highly quantifiable. Several examples of this are depicted in **Figure 1.1**. More complex parameters include closeness centrality, which measures how close a node, *A*, is to the rest of the network by calculating the shortest distance of *A* to all other nodes. Betweenness centrality describes the number of times Node *A* appears in the shortest path of all pairs of nodes in the network. Methods of calculating some of these are listed in **Table 1.2**. These parameters facilitate the comparison of features not just within one network, but also between many.

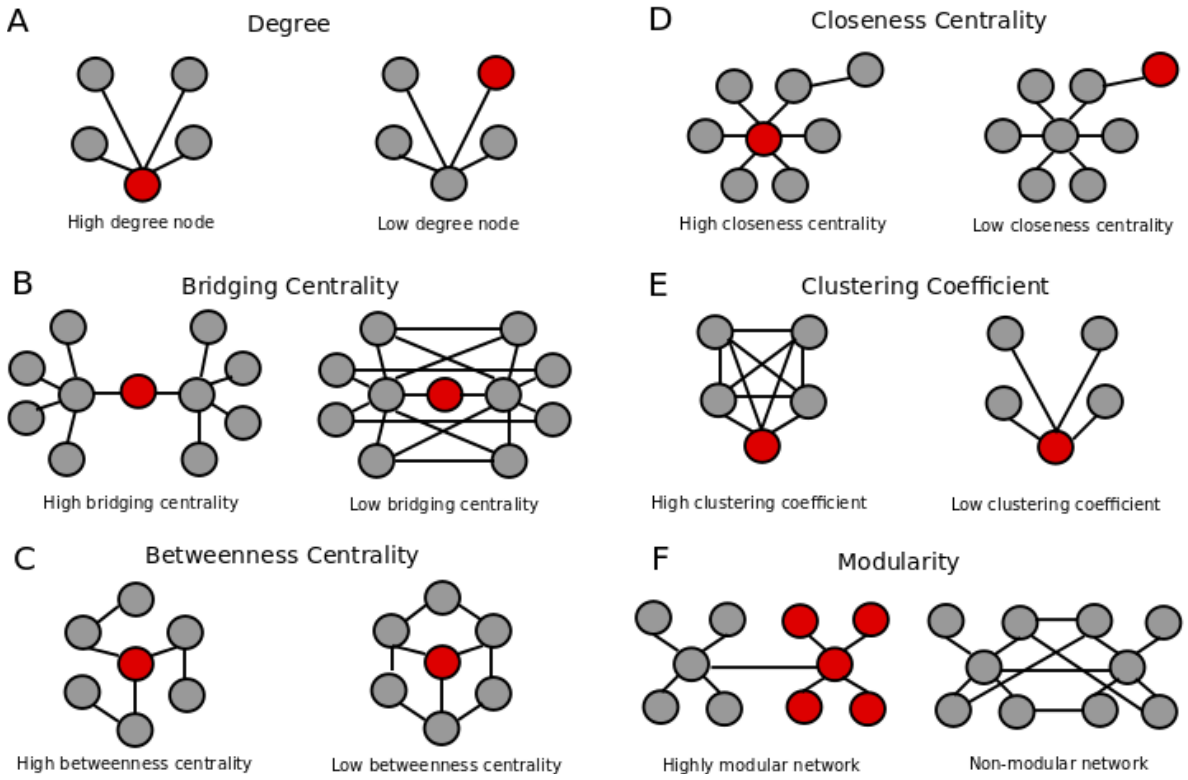


Figure 1.1: Network measures

Red nodes in A.) through E.) indicate the node of interest. A.) Degree is the number of edges connected to a node. B.) Bridging centrality represents the extent to which nodes link highly connected subcomponents (modules). C.) Betweenness centrality is a representation of the "traffic" that a node experiences, and it measures how many times a given node appears in the shortest path between all other node pairs in the network. D.) Closeness centrality measures how close a node is to all other nodes in the network. E.) The clustering coefficient shows how connected the direct neighbours of a node are. F.) Modularity describes the degree of network separation into modules.

	Property	Equation	Description
A	Degree	$k_i = \sum_{j=1}^{N} A_{ij}$ where A_{ij} =1 if there exists an edge between nodes *i* and *j*, and 0 otherwise	The degree is the sum of all edges connected to a node in an undirected network; for a directed network, it may be calculated as the sum of the incoming, outgoing, or total edges.
B	Bridging centrality	$\mathrm{BC}_i = \frac{{k_i}^{-1}}{\sum_{A_{ij}=1} {d_j}^{-1}}$	The bridging centrality is the inverse degree of the node divided by the sum of inverse degrees of its neighbours.
C	Betweenness centrality	$k_i = \sum_{j=1}^{N} \frac{\sigma_{hj}(i)}{\sigma_{hj}}$	The betweenness centrality of *i* is the number of shortest paths (σ) between nodes h and j that pass through node *i* divided by all shortest paths between *h* and *j*.
D	Closeness centrality	$c_i = \frac{1}{\sum_{j=1}^{N} \sigma_{ij}}$	Closeness centrality is the inverse sum of the shortest paths between *i* and all other nodes in the network.
E	Clustering coefficient	$C_i = \frac{\sum_{A_{ij}=1} k_j}{k_i(k_i - 1)}$	The clustering coefficient describes how connected the neighbors of *i* are by calculating the number of connections between all the neighbors of *i* and dividing it by the maximum possible number of connections among them.
F	Modularity (for two groups)	$M = \frac{1}{\sum_i^N k_i} \sum_{i \neq j} g(A_{ij} - \frac{k_i k_j}{\sum_i^N k_i})$ where *g*=1 if *i* and *j* are in the same group, and 0 otherwise	First, all edges are cut in half into "stubs" such that there are $\sum k_N$ total stubs in the network. These are randomly reconnected into edges. *M* is calculated as the actual number of edges in a group minus the expected number. Positive modularity indicates the likelihood of community structure

Table 1.2: Calculating network properties
In this table, we list the network properties described in Figure 1.1 and how to calculate them.

In all, networks are a viable and robust approach to representing biological systems without oversimplifying the complex nature of the cell. In addition, the tools previously described allow networks to be efficiently analyzed, paving the way for drawing meaningful conclusions about biological function, disease, and potential treatments.

Network biology

Networks can be used to define a number of biological systems. These data can be of experimental origin, such as genetic co-expression or protein–protein interactions (PPIs), or taken from clinical findings, such as correlations between genetic mutations and disease phenotype.

Previously, two overarching types of biological networks have been described: molecular and phenotypic [14]. Molecular networks include those depicting PPIs, metabolic reactions, regulatory relationships such as those between transcription factors and genes, and RNA

networks such as microRNA-associated gene expression. Phenotypic networks include those depicting gene co-expression or gene–phenotype relationships, such as gene–disease associations.

Biological networks have a number of similar properties. For example, in molecular networks, hub nodes have been associated with essential genes [15]. Modules often represent subnetworks that are associated with a unique biological function. Disease modules are an interpretive extension of functional modules; a disruption in various parts of a functional module can lead to identical or related diseases.

Hypothesis generation with networks

Networks can be used to generate hypotheses. For example, Ciriello *et al.* integrated the human PPI network with the hypothesis of mutual exclusivity in cancer, based on the observation that cancer patients tend to harbour only one mutation per pathway, though each pathway may be altered in various locations [7]. They first identified genes most likely to participate in tumour progression, and then found modules in the PPI network significantly enriched for those genes. Maximal cliques were extracted from network modules and assessed for mutual exclusivity. This method successfully confirmed previously recognized altered pathways and also found unexpected mutual exclusivity between the gene *RBBP8* and the BRCA/Rb pathway that may indicate a novel role for the gene in different parts of the cell cycle.

Network-based predictive models

Biological networks can be used to create predictive models using machine learning (ML). ML is a branch of computer science, wherein computers can "learn" without being programmed. There are a number of methodologies in ML; here, we will focus on classification, where the property being learned is a status, such as diseased vs. healthy. Typically, ML requires two types of inputs: labels, such as disease status, and feature vectors, such as gene expression. In the case

of supervised learning, labels and features are provided to create a model. A new feature vector can then be fed into the model to predict its label.

Using networks in machine learning is possible because of their mathematical properties. In this case, the features are a series of values that describe the node, node pair, module, or network, such as shortest path, betweenness centrality, degree, *etc.* For example, Lorberbaum *et al*. developed the Modular Assembly of Drug Safety Subnetworks (MADSS), a network analysis-based algorithm that identifies adverse event neighbourhoods within the human interactome [16]. Drugs targeting proteins within this neighbourhood are predicted to be more likely to cause the ADR than drugs targeting proteins outside the neighbourhood. Beginning with a small "seed" set of highly interconnected proteins with a direct genetic link to an ADR of interest, the authors then scored every protein in the human PPI network on how well-connected it was to the seed set using multiple network connectivity functions, including shortest path and shared neighbours. They trained a random forest classifier using each of the connectivity metrics as features to generate drug safety subnetwork models, then evaluated drug safety using both known and predicted drug targets.

SYNTHETIC LETHALITY

Biological understanding and theory

Synthetic lethality (SL) occurs when changes in two otherwise nonessential genes results in an unviable cell or organism (Figure 1.2A). Although these perturbations can be of various types, most commonly, they refer to the removal or full inhibition of a particular gene. Synergistic interaction between genes to the point of lethality was first described in *Drosophila melanogaster* in 1922 [17], then confirmed in *Drosophila pseudoobscura* in 1964 [18]. To date, synthetic lethality has been described in numerous organisms, including humans. For example, in *Saccharomyces cerevisiae*, cytidine 5′-triphosphate synthetase catalyzes the conversion of uridine 5′-triphosphate to cytidine 5′-triphosphate. *URA7* and *URA8* are nonessential genes that redundantly code for the same enzyme; however, knocking out both *URA7* and *URA8* results in synthetic lethality [19].

SL has a number of possible mechanisms through which it can occur [20,21]; several of these are illustrated in Figure 1.2B. For example, in the case of parallel pathways that both lead to an essential product, knocking out one gene in one of the arms won't affect the viability of the pathway, as the other arm serves as a backup. Similarly, knocking out two genes within the same arm will still leave the other one as a backup. However, knocking out two genes in different arms (A/X, A/Y, B/X, B/Y) will lead to lethality, as both arms of the pathway will collapse, and the essential downstream protein will no longer be produced. Other possible mechanisms include homodimerization and functional redundancy. In short, synthetic lethality can be thought of as a function of genetic buffering [22].

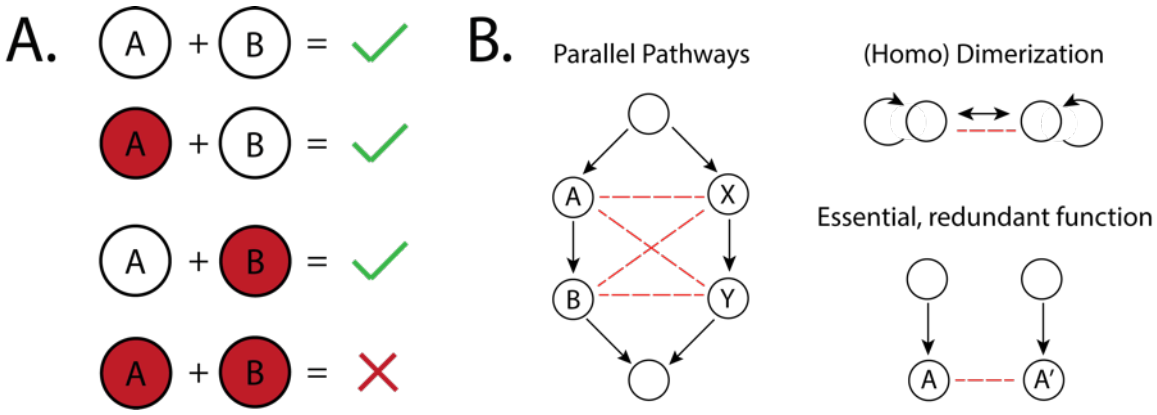


Figure 1.2: Mechanisms of synthetic lethality
A.) Synthetic lethality occurs when two genes that can be knocked out individually with no harm to the viability of the cell, but knocking out both together causes cell death. B.) There are a number of possible mechanisms by which synthetic lethality can occur. In a parallel pathway, knocking out A and B or X and Y has no effect on the viability of the cell, but knocking out two genes on either arm of the pathway will cause cell death. In the case of potential homodimerization, knocking either gene out may not affect viability, but both will. Finally, in the case of an essential product with a redundant copy, knocking out only one will still leave a backup, but losing both proves fatal.

Experimental approaches

Typically, synthetic lethality has been studied in model organisms such as *S. cerevisiae* [23], *S. pombe* [24], and *C. elegans* [25] due to the ease with which they can be manipulated. To identify synthetic lethality, organisms are observed for growth and fecundity; a complete lack thereof indicates synthetic lethality. Significantly decreased growth may be dubbed "synthetic sickness." However, depending on the organism, different methods must be used to create double-mutant organisms or colonies, and the type of organism (single-cell vs. multicellular) may affect assessment of growth.

In organisms like *C. elegans*, RNA interference (RNAi) may be used to knock out genes. RNAi uses double-stranded RNA to trigger a gene-silencing pathway, and is common in a number of eukaryotic species. In one study [26], researchers used *C. elegans* with wild-type *bmk-1* or *bmk-1(ok391)*, which harbours a genetic deletion allele. They induced genetic knockdown by RNAi with a genome-wide library, and assessed synthetic lethality using fluorescence microscopy.

RNAi has also been used to test synthetic lethality in humans. However, testing for lethality on humans is considered unethical; therefore, these screens are conducted in human cells. These can be both established cell lines, or patient-derived ones [27].

In *S. cerevisiae*, RNAi is ineffective [28], so a different approach must be used. There, two parental strains that each have a single gene deletion are used to create double-mutant progeny [23]. The growth of these offspring may be compared to the single-mutant colonies in order to identify pairs that have fitness defects that are greater than expected [24].

Computational models of synthetic lethality

Previous work has predicted synthetic lethality in yeast using the protein-protein interaction (PPI) network [29]. The authors hypothesized that the topology of node pairs in the network will change depending on the SL status of their associated genes. This hypothesis relies in part on the existence of functional modules in biological networks, as often, the topological similarity of two nodes indicates shared function, and many SL pairs have also been shown to share functional annotations [30].

In this paper, the authors used machine learning to create these models. Each gene pair was associated with a label and a feature vector. For the label, the researchers used experimental data to describe the pair as SL or non-SL. For each feature vector, they used node or node-pair properties such as degree, shortest path, and number of shared neighbours. These were fed into a support vector machine to create a classifier that predicts a gene pair's label from its associated feature vector. The classifier was successful, achieving an area under the receiver operating curve (AUC) of at least 0.89 for all cross-validation runs. This suggests that PPI networks are highly informative for predicting SL in yeast.

DRUG SYNERGY

Introduction to drug interactions

In certain situations, drug effects may change due to the presence of other drugs or chemicals. For example, grapefruit juice interacts with a number of drugs, including simvastatin [31], because it decreases the activity of the enzyme that metabolizes them. This causes the drug to remain in the system for longer, and may lead to overdose. Drug interactions may also decrease the effect of a drug. For example, nalaxone acts as a competitive antagonist of the mu-opioid receptor. Therefore, nalaxone can be used to reverse the effects of a morphine overdose [32].

Each drug interaction occurs through one of two mechanisms: pharmacokinetic or pharmacodynamic. Pharmacokinetics can be considered the study of the body's effect on the drug — how a drug is absorbed, distributed, and metabolized. Therefore, in a pharmacokinetic interaction, one drug affects how another is processed. For example, ciprofloxacin is an inhibitor of CYP3A4, the prime metabolizer of the antidiabetic glyburide [33]. If they are given together, in some patients, ciprofloxacin may increase the effects of glyburide and lead to hypoglycemia [34].

Pharmacodynamics, on the other hand, is the drug's effect on the body. A pharmacodynamic interaction therefore occurs when two drugs exhibit similar mechanistic spheres of influence. Combining antipsychotics (dopamine antagonists) with levodopa (a Parkinson's drug that raises dopamine levels) can result in an interaction where one drug negates the effect of another. Taking the drugs simultaneously could therefore cause a relapse of psychosis, or a worsening of motor function [35].

Although the potential for adverse drug interactions is possible, multidrug therapy can also have significant benefits. For example, in HIV therapy, using multiple drugs at once prevents the rise of resistance in the virus [36]. Furthermore, the use of multidrug therapy in cancer can both prevent drug resistance, and also reduce the dose of each drug required for an effect, reducing drug side effects [37] .

The effect of multiple drugs simultaneously falls into one of three bins: "synergy," "additivity," and "antagonism" (Figure 1.3). In the case of additivity, the effect of the two drugs is unchanged from taking them separately - that is, taking A and B together results in an effect of A + B (Figure 1.3B). Additivity may occur because the drugs are completely unrelated both pharmacodynamically and pharmacokinetically, or because they are very similar in both respects — for example, acetaminophen and aspirin.

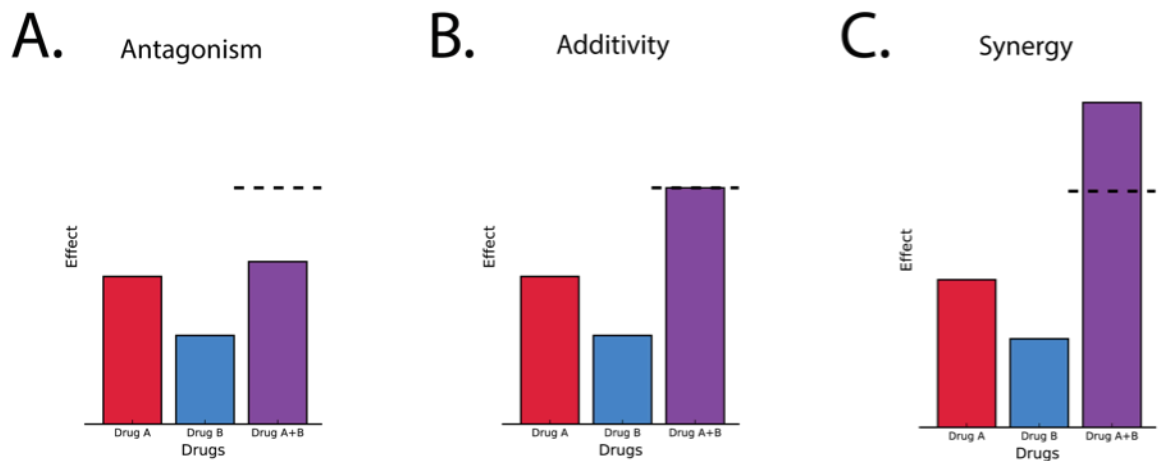


Figure 1.3: Illustrations of drug interactions
A. In an antagonistic interaction the effect of giving drugs A and B simultaneously lessens the effect of one or both drugs given independently of each other. B. Additivity means that the effect of giving two drugs at the same time is the sum of giving them independently. C. In synergy, the effect of one or both drugs is magnified when they are given in combination.

Drug synergy occurs when the effect of using Drug A and Drug B together is greater than the expected additive effect (Figure 1.3C). This may be because one drug slows the metabolism

of the other, or because they affect different pathways that produce the same final result. Simvastatin and grapefruit juice produce a synergistic drug interaction, as does ciprofloxacin with glyburide. Finally, drug antagonism occurs when drugs taken together have a smaller effect than expected (Figure 1.3A). Morphine taken with nalaxone is an example of this.

Although the effects listed above have been described *in vivo*, it is also possible to test for drug interactions in the laboratory. The first of these tests is usually *in vitro*, in a cell line; here, cell growth or death may be measured as a proxy for effect. These drugs can later be tested *in vivo* in model organisms, such as rats or mice, although these experiments have their own limitations in translation to humans [38].

Models of measuring drug synergy

Drug synergy can be measured using a number of different methods [39,40]. They may be chosen on the basis of drug mechanism and effect. In this section, we will outline several of the most common ones.

Loewe additivity was developed in 1953 [41]. Here, a desired response level is chosen for each drug. In our case, let us say that X μM of Drug A is required for this effect, and YμM of Drug B. Next, the concentration of Drug A is plotted on the x-axis, and that of Drug B on the Y. A straight line is drawn between X μM and Y μM, and it is used to represent the isobole. If the drugs have an additive effect, varying the concentration of A or B should result in an effect existing on that line. This can be described by the equation:

$$\frac{x}{X} + \frac{y}{Y} = 1$$

where X is the concentration of Drug A alone for the desired effect; Y is the concentration of Drug B alone for the desired effect; and x and y are the concentrations of Drugs A and B, respectively, when the two are taken together. An effect occurring above the line ($\frac{x}{X} + \frac{y}{Y} > 1$)

indicates synergy, while an effect below the line ($\frac{x}{X} + \frac{y}{Y} < 1$) indicates antagonism. This model has several drawbacks. In some cases, the isobole may not be a straight line, and the equation can be adjusted based on the expected response. This may occur in cases where the maximum effect of Drug A and Drug B are significantly different. In addition, the model is based on the idea that the two inhibitors act through similar mechanisms, and is thus inappropriate for very different drugs.

A number of other methods exist that relate the combination to the effects of individual components — that is, the expected effect of Drug A and B taken together (E_{AB}) is a function of the effects of Drug A (E_A) and Drug B (E_B) alone [39]. These are an improvement over Loewe additivity in that they are not limited to combinations of similar drugs.

Combination subthresholding [42] relies on identifying doses at which Drug A and B are ineffective on their own, but become significantly effective when given together. The significance of these effects is determined by comparing them to a control group given neither drug; however, the reliance on p-values means that statistical blips may falsely indicate synergy. For example, at a cut-off of p=0.05, single-drug effects at p=0.05001 compared to a combined effect of p=0.04999 would be defined as significantly synergistic.

Highest single agent [43,44] compares the combined effect (E_{AB}) to the highest effect of each individual drug (*max*(E_A,E_B)) in order to determine synergy and assess significance. The combination index (CI) can be described as $\mathrm{CI} = \frac{\max(E_A, E_B)}{E_{AB}}$, where CI < 1 indicates synergy. This method shows improvement over a single drug, but doesn't necessarily indicate synergy. However, it may be useful in cases where the second drug shows little to no effect on its own.

Response additivity [45] follows a similar principle, but CI is calculated using the formula $\mathrm{CI} = \frac{E_A+E_B}{E_{AB}}$. Although this formula assesses synergy as a combination of both drugs' effects, it assumes linear dose-effect curves, which is inaccurate for many drugs.

Bliss independence [46] relies on a probabilistic model of drug action. The expected effect of a drug pair is defined as $E_A + E_B - E_A \times E_B$, where E_A and E_B are values between 0 and 1. In this case, therefore, raw cell counts or specific concentrations cannot be used; the effect must be described in relation to the control (e.g. as growth inhibition). Excess over Bliss (EOB) can be used to describe synergy:

$$\mathrm{EOB} = \mathrm{Observed} - \mathrm{Expected} = E_{AB} - (E_A + E_B - E_A \times E_B)$$

An EOB > 0 indicates synergy, as the observed effect is greater than the expected; EOB < 0 indicates antagonism. Bliss Independence is one of the most popular methods of predicting drug synergy due to its versatility and simplicity, and it is the method that we will focus on for the remainder of this work.

Shortcomings of Bliss independence

Although Bliss Independence is an extremely popular method of calculating drug synergy, it has certain limitations. These are of two kinds: theoretical and practical.

Theoretical concerns

The first of these is that drugs themselves are often messy, with multiple targets and many known (and, often, unknown) mechanisms of action. Therefore, the use of a model that assumes independence may be inappropriate for a number of drug combinations, and this may not be clear at the time of experiments.

Next, because EOB is a probabilistic model, the numbers used in its calculation must be [0,1]. Therefore, if the control sample has 14,000 cells, and the drug-treated one has 7,000, the effect of the drug would be calculated as (14,000-7,000)/(14,000) = 0.5. However, if the drug

causes accelerated growth, and cell count after drug is 21,000, the effect would be calculated as (14,000–21,000)/14,000 = -0.5, which is not feasible for use in calculating EOB.

Finally, the assumption that drugs have exponential dose-effect curves [44] may lead to an incorrect calculation that a drug is synergistic with itself – which, by definition, is an impossibility [39].

Practical concerns

There are two primary practical concerns in the use of EOB to measure drug synergy. The first is the use of replicates in experimental biology. These are necessary to ensure that statistical fluctuation isn't the reason for a designation of drug synergy. However, this does affect the calculation of percent inhibition in EOB.

Let us assume that we perform an experiment with three replicates per dose level. Therefore, to measure effect of a given drug or drug pair at a specific concentration, we must compare the three dosed replicates to three control samples. Typically, the effect is calculated for each dosed replicate against the median of control samples. Then, the median EOB is reported to describe synergy. However, in so doing, researchers do not account for variance in either the control samples or dosed ones; at best, the standard error of EOB can be reported.

In addition, experimental fluctuation may also lead to 'impossible' effect scores (Figure 1.4). For example, in the case where the lowest dose of a drug is generally ineffective, comparing the replicates to the median of control scores may still lead to net growth — *e.g.* a score of -0.05. Because the maximum effect score in the calculation of EOB is 1, the replicate must be either invalidated, or artificially set to a score of 0.00 (no effect). This will artificially inflate the expected score of a combination.

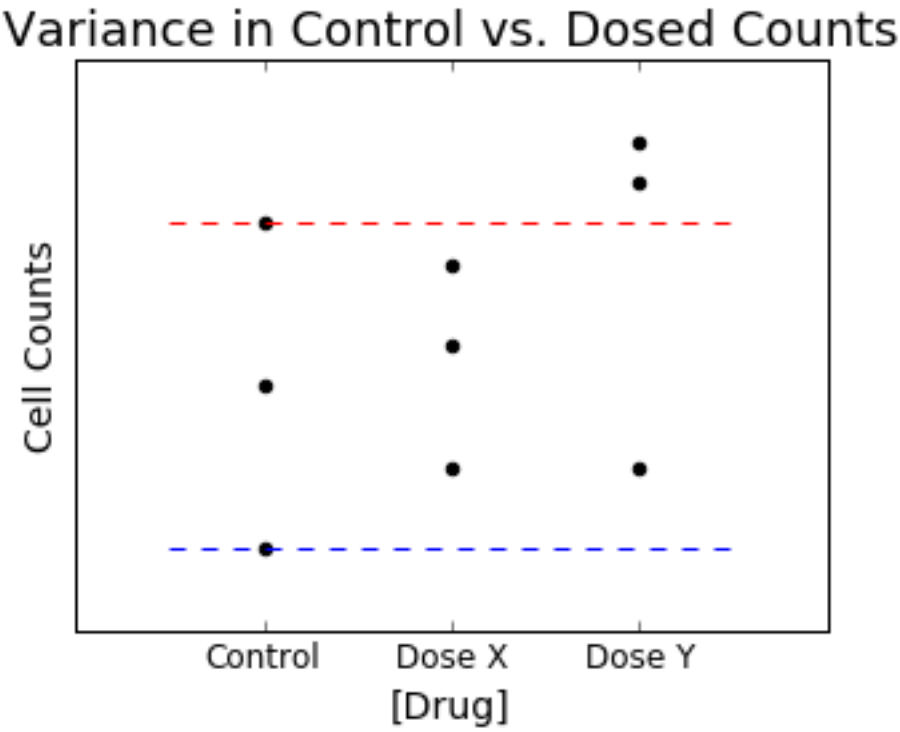


Figure 1.4: Bliss independence and experimental replicates
We present the simulated distribution of cell counts in three replicates each of control, dose X, and dose Y experiments. In typical use of EOB, the median score of the control samples would be used to determine the effect (e.g. percent inhibition) of each replicate of a dosed sample. This would mean that two replicates at Dose X and one replicate at Dose Y would be seen as having negative effects; however, in the case of Dose X, all effects actually fall within the range of control wells, meaning that the effect being reported is not accurate.

Furthermore, it doesn't account for experimental variance properly; at a dose of X, the drug shows negative inhibition compared to the control median, even though the distribution of responses is within the distribution of control wells. In contrast, the results at dose Y are mostly out of the control range, but they would be considered equivalent to dose X if adjustment occurred. This is clearly inaccurate.

Second, the establishment of statistical significance is a problem in the practical application of Bliss Independence. In particular, there is no statistic of significance to describe EOB. An EOB of 0.001 is considered synergistic — as is 0.0001, 0.05, or 0.99. Thus, even if the standard error indicates that the EOB is net positive, it is hard to trust that a score of 0.0001 is true synergy and not merely a statistical blip.

Based on these shortcomings, a version of Bliss independence that takes into account the variation in control and experimental replicates would be a useful and necessary step forward in assessing drug synergy.

PREDICTING AND EVALUATING DRUG SYNERGY USING SYNTHETIC LETHALITY

In this work, we aim to synthesize the fields described in this introduction to address two questions: can we develop a computational model of human synthetic lethality, and can these predictions be used to inform combination cancer drug therapy?

Computational models of human synthetic lethality

Given the previous success of developing computational models of synthetic lethality (SL) [29], we hypothesize that we can leverage the similar structures of protein-protein interaction (PPI) networks and experimental yeast data to predict SL in humans. We do so by first developing the notion of connectivity homology, a method through which we can compare interspecies protein-protein interaction networks (Chapter 2). We then use connectivity homology to create an algorithm, Species-INdependent TRAnslation (SINaTRA): an interspecies, machine-learning model of synthetic lethality based on *S. cerevisiae* experimental data and validate it in *S. pombe* (Chapter 3). Finally, we apply the model to human PPI data (Chapter 4).

Synthetic lethality and drug synergy

In Chapter 5, we test ten putative human SL pairs and five predicted non-SL gene pairs for synergy using drug combinations in a number of cell lines. To do so, we develop DAVISS (Data-driven Assessment of Variability In Synergy Scores), a novel method of testing the statistical significance of drug synergy based on Bliss independence that takes into account the variance of control wells.

ACKNOWLEDGEMENTS

Sections of this introduction are adapted from articles in Clinical Pharmacology and Therapeutics by Jacunski *et al*. [47] and in WIREs Systems Biology and Medicine by Boland *et al*. [48]. I thank Benjamin Sally for his careful reading of the former manuscript. Furthermore, I thank my coauthors on the latter paper for their help in writing and assembling the review.

CHAPTER 2 – CONNECTIVITY HOMOLOGY

INTRODUCTION

Biological networks have a number of similar properties. For example, protein-protein (PPI) networks tend to be similar in terms of connectivity patterns, regardless of species. Research has suggested that they are connected according to a scale-free, power law distribution, where a new node being added to a network is more likely to connect to an existing node of high degree [49]. Furthermore, biological networks can also be described as "small world" [50], where each node is connected to every other one with a relatively small number of steps.

In spite of these structural similarities, PPI networks are typically constructed using genes as nodes; thus, species with more genes will necessarily have larger networks. Furthermore, different networks may have different levels of completeness [51];
an organism that is well studied will have more nodes and edges than a less-studied one, even if the two organisms have similar genome sizes. Therefore, upon calculating the parameters of the two networks, such as shortest path and degree, the distributions will be different between the two species. This, in turn, will mean that they the two networks not be immediately comparable; what is considered a high value in one network may be low in another.

Here, we introduce the concept of *connectivity homology*, a measure of relatedness between genes based on protein-protein interaction networks. Connectivity homology is independent of structure, function, or genetic homology. We first illustrate this concept with two toy networks. Next, we perform a brief experiment illustrating the principle of connectivity homology in networks evolved *in silico*. We show that a node in a network evolved via preferential attachment retains similar properties throughout network growth, and thus exhibits higher connectivity homology, compared to one in an evolving random network.

Finally, we explore connectivity homology in *S. cerevisaie*, *S. pombe*, and human PPI networks using well-known graph properties, such as degree centrality and shortest path [14,47,52,53]. We find that both orthologous and non-orthologous genes of the same function have similar connectivity patterns between species.

These results suggest that connectivity homology is an inherent property of biological networks based on their evolutionary patterns; thus, it is useful for the understanding of biological phenomena on an interspecies level.

RESULTS

Defining connectivity homology

We define two proteins as being *connectively homologous* if they share similar connectivity profiles in their respective networks. A connectively homologous relationship may exist between two proteins in the same species, or between proteins of different species. This concept can be generalized for pairs of proteins, or even groups of proteins (*i.e*. modules). For example, two pairs of proteins may be connectively homologous because both pairs are connected to each other in a similar way.

We illustrate this concept in Figure 2.1, where we present two networks of different sizes and topologies. We used two network parameters to describe the network: degree and betweenness centrality. These network parameters are not immediately comparable; for example, the range of degrees in Network 1 is [1,3], while it is [2,5] in Network 2. However, we can compute very simple connectivity profiles for each parameter of each node, where it is classified as either low (blue), medium (white), or high (red). When comparing the connectivity profiles of various nodes, it becomes apparent that certain sets of nodes are connectively homologous with each other (*e.g.* Node B/Node 2/Node 3). In contrast, nodes with the same raw parameters (*e.g.* Node A/Node 1) may not necessarily be connectively homologous.

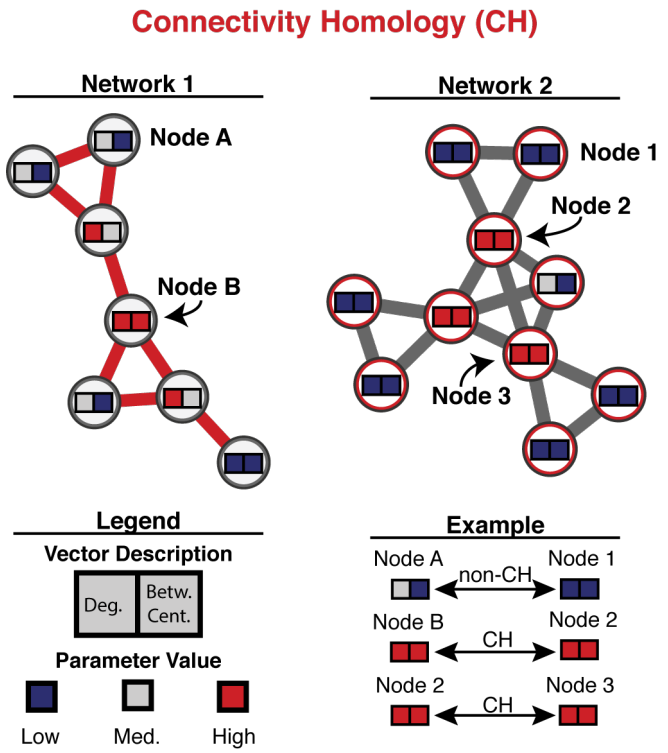


Figure 2.1: An illustration of connectivity homology
Each node is described by two parameters (degree [deg.] and betweenness centrality [bet.cent.]) at three levels: low, medium, and high. Certain nodes have the same vectors (Node B/Node 2/Node 3); these nodes can be said to be connectively homologous (CH). Other nodes do not (Node A/Node 1); these are non-connectively homologous (non-CH).

In silico evolution of networks indicates biological bases for connectivity homology

Biological networks have been suggested to follow a power law distribution, where nodes of high degree are more likely to receive new connections when a new node joins the network [49]. Given this model of preferential attachment, it is intuitive that overarching connectivity patterns will remain similar in biological networks as they grow. This means that orthologous genes are likely to maintain similar connectivity, regardless of time.

To confirm this, we generated and evolved two types of networks: one observing growth by preferential attachment (PA network), and another grown randomly (RD network) (Figure 2.2).

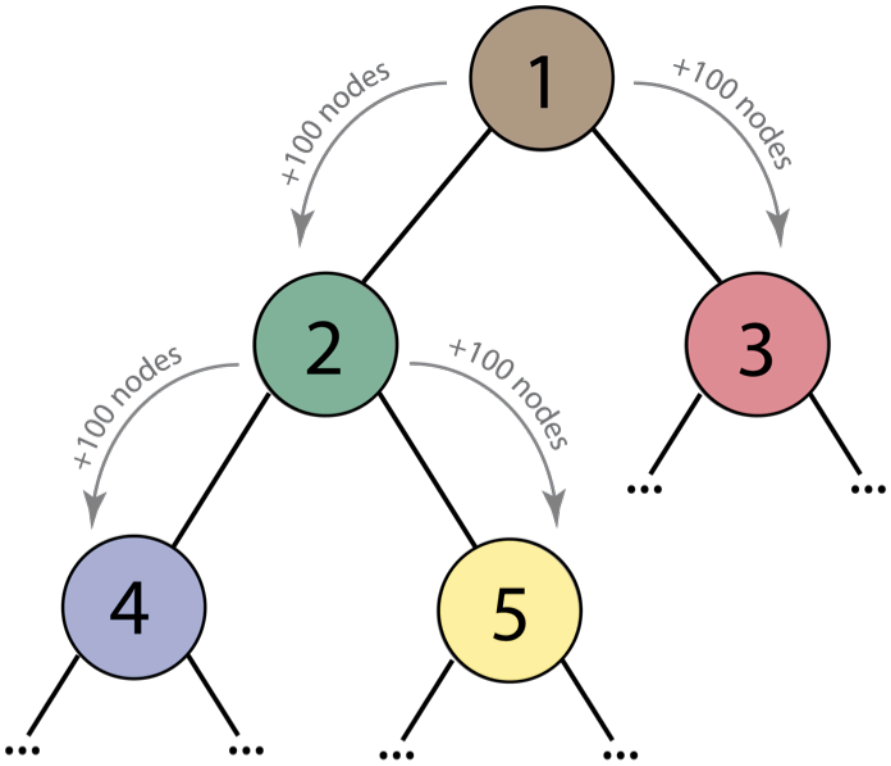


Figure 2.2: Network evolution
To evolve both random and preferential attachment networks, we first started with a parent node constructed according to each kind (Node 1). We then created two 'evolved' children (Nodes 2 and 3) by adding 100 nodes to the parent and connecting them to the network according to the method of attachment being used. Each child of the original parent would then be similarly evolved, until we had a perfect binary tree of 16 levels.

When we compared the connectivity patterns of the original 1,000 nodes of each network over the network's evolution, we found that the median differences in degree and betweenness centrality were significantly more similar in PA networks than in RD ones ($p<2.2e-16$, Mann-Whitney U test; Figure 2.3).

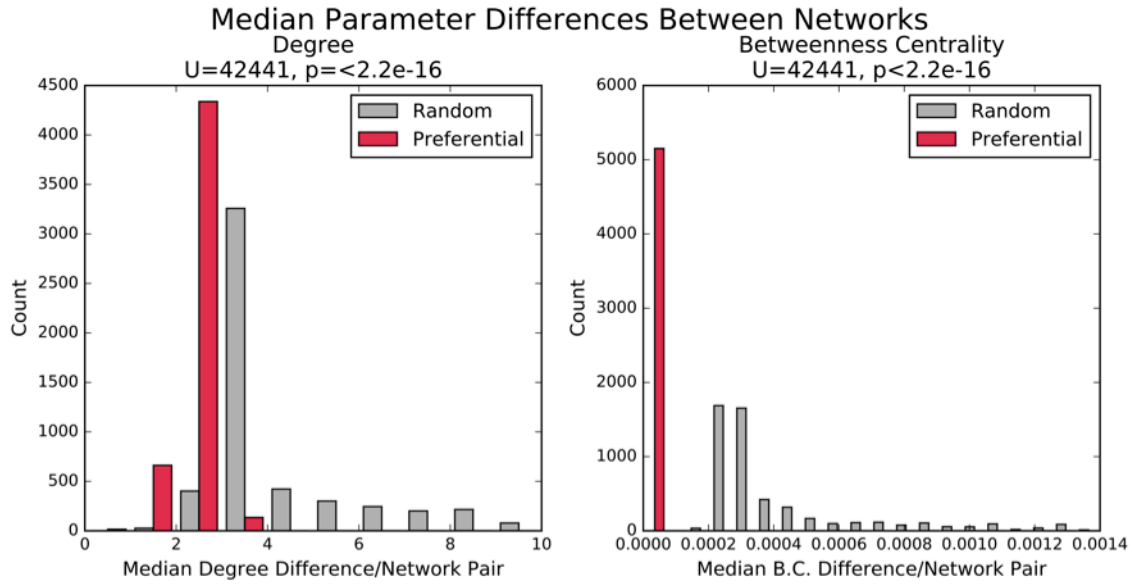


Figure 2.3: Median parameter differences between co-evolved networks
We selected 100 networks from two different evolved "phylogenies" (random and preferential attachment), then compare the median parameters of the 1,000 original nodes of the parent network: degree (left) and betweenness centrality (right).

Furthermore, when we compared Spearman correlation between a parameter of two networks (Figure 2.4), we found that PA networks had significantly higher Spearman correlation than RD networks for both degree and betweenness centrality ($p<2.2e-16$ for both, Mann-Whitney U test). This is partially because PA networks are significantly closer in 'evolutionary distance' than RD ones (Figure 2.A.1). Thus, a single "step" in a PA network's evolution brings about less change than one in an RD network.

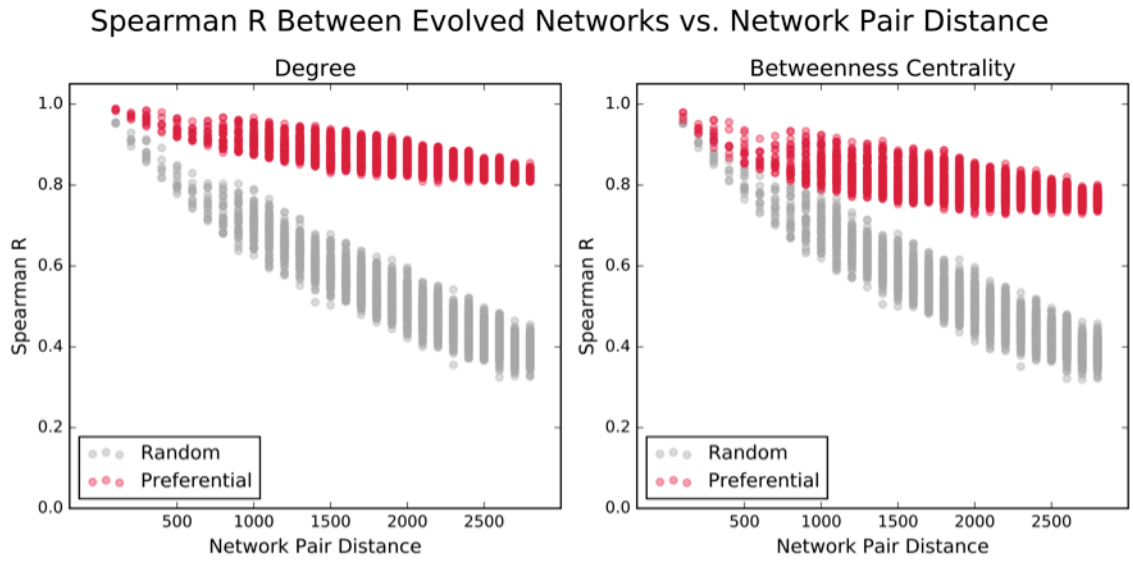


Figure 2.4: Spearman correlation between evolved networks
We selected 100 networks from each evolved phylogeny (random and preferential attachment), and then computed the Spearman correlation of two parameters (degree and betweenness centrality) for the 1,000 nodes from the original network. We find that networks with preferential attachment (red) have significantly higher Spearman R values than those with random attachment (gray), even when the networks have large distances between them.

Connectivity homology can be evaluated with network parameters

As in the previous section, we show connectivity profiles using vectors of network parameters. A vector of eight parameters represents each gene (Tables 2.1 and 2.A.1). Each gene pair is represented by four node-pair parameters as well as the individual profiles for each gene in the pair, leaving each pair with a connectivity profile defined by a vector of 20 network parameters. For the purposes of this investigation, we chose to use protein-protein interaction (PPI) networks because of the wide availability of data across many species. PPI data was downloaded from BioGRID [54] to construct graphs, which are pruned to contain one connected component (*Materials and Methods*). We computed the connectivity profiles for 5,810 proteins in *S. cerevisiae*, 1,919 in *S. pombe*, 4,233 in *M. musculus*, and 14,820 proteins in humans as well as for 16.8 million, 1.8 million, 8.9 million, and 109.8 million pairs of proteins for *S. cerevisiae*, *S. pombe*, *M. musculus*, and humans, respectively.

Parameter	Context	Description
2nd degree shared neighbours	Single	The sum of all nodes two edges away from the node

	node	of interest
Betweenness centrality	Single node	The sum of the fraction of shortest paths between two other nodes passing through the node of interest
Closeness centrality	Single node	The inverse sum of all shortest paths that originate at the node of interest
Communicability	Node pair	The sum of all closed walks between a pair of nodes
Current-flow betweenness centrality	Single node	Analogous to betweenness centrality, but with all paths instead of shortest paths. Also known as random walk betweenness centrality.
Degree centrality	Single node	The fraction of edges a node has of all possible edges
Eccentricity	Single node	The maximum distance from the node of interest to any other node in the network
Eigenvector centrality	Single node	The eigenvector for the largest eigenvalue of the matrix adjacency network
Inverse shortest path	Node pair	The inverse of the smallest number of edges connecting two nodes of interest
PageRank	Single node	The rank of a graph's nodes based on the number of incoming links
Shared neighbours	Node pair	The intersection of two nodes' sets of immediate neighbours.
Shared non-neighbours	Node pair	The number of nodes that are not immediate neighbours of the two nodes of interest

Table 2.1: Parameter descriptions
Here, we describe the network parameters used to explore connectivity homology in the S. cerevisiae, S. pombe, mouse, and human networks.

We found that the distributions and ranges of network parameter values differed significantly between species (Fig 2.A.2; Table 2.2). To correct for this (Figure 2.5), we chose to use rank normalization to rescale the values of each parameter between 0 and 1; this allows them to be comparable between species. We refer to normalized data as being translated.

		2nd Degree Shared Neighbors	Between-ness Centrality	Closeness Centrality	Communi-cability	Current-flow Betweenness Centrality	Degree Centrality	Eccen-tricity	Eigenvector Centrality	Inverse Shortest Path	PageRank	Shared Neighbors	Shared non-neighbors
Cerevisae-Pombe	MWU	29693.5	460889	2192	0	398166	413170	2	437106	101498.5	12745	256887.5	0
	p-value	9.84E-291	0.0011	<2.2E-16	<2.2E-16	1.31E-15	7.86E-12	<2.2E-16	5.56E-07	2.50E-224	<2.2E-16	3.10E-118	<2.2E-16
Cerevisae-Mouse	MWU	31873	442123	504	0	445945.5	194189	0	344992	80512	132671	233678	0
	p-value	4.56E-288	2.83E-06	<2.2E-16	<2.2E-16	1.25E-05	1.10E-125	<2.2E-16	1.69E-33	1.18E-247	2.70E-178	4.46E-145	<2.2E-16
Cerevisae-Human	MWU	311210.5	327722	369844	69	271969.5	196167	44895	350074	438865.5	31788	406604	0
	p-value	1.04E-48	3.70E-41	3.39E-24	<2.2E-16	2.96E-70	6.21E-123	<2.2E-16	1.81E-31	5.85E-08	3.40E-288	1.47E-15	<2.2E-16
Pombe-Mouse	MWU	495858	435956.5	445933	335172	387754.5	194674	20939	412643	480893	34263	488812.5	0
	p-value	0.3742	9.48E-08	1.41E-05	1.30E-37	1.62E-19	1.08E-125	<2.2E-16	6.66E-12	0.0602	3.67E-285	0.0205	<2.2E-16
Pombe-Human	MWU	98830	406552	19307	0	332705	197158	8371	436114	136340	3664	320172	0
	p-value	3.08E-212	6.56E-14	1.22E-303	<2.2E-16	2.27E-39	1.02E-122	<2.2E-16	3.76E-07	4.31E-186	<2.2E-16	2.45E-76	<2.2E-16
Mouse-Human	MWU	85225	447096	14237	0	451730	397249	0	477952	108443	10286	298150	0
	p-value	1.00E-226	1.22E-05	5.15E-310	<2.2E-16	6.97E-05	5.53E-16	<2.2E-16	0.0439	5.44E-216	<2.2E-16	2.76E-95	<2.2E-16

Table 2.2: Comparison of network parameter distributions
Distribution of all untranslated network parameters between species, described using the Mann-Whitney U test ("MWU") and associated p-values.

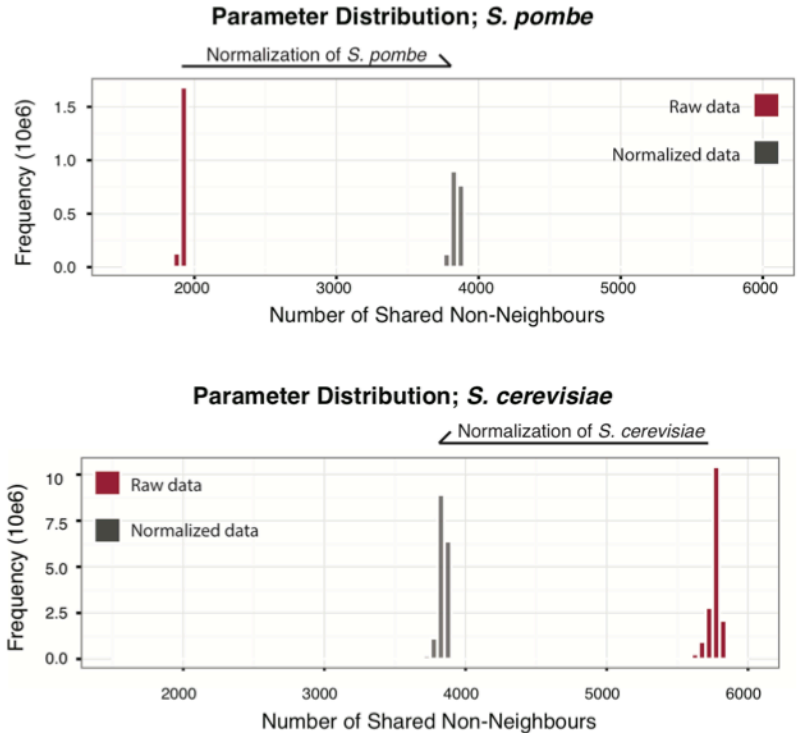


Figure 2.5: Parameter correction from *S. cerevisiae* to *S. pombe*
The use of normalization makes network parameters that are not comparable before translation (red) easily compared after (gray).

Similarity between connectivity vectors is indicative of shared function

We found that proteins with similar connectivity profiles (*i.e.* those that are connectively homologous) were more likely to share functional annotations. We used the Euclidean distance between connectivity profiles as a measure of connectivity homology (*Materials and Methods*).

We compared this distance between genes that share genetic homology (orthologs) and specific functional annotations (Gene Ontology [GO]) [55,56] between *S. cerevisiae* and *S. pombe* (Sc/Sp) (Figure 2.6A) and between
S. cerevisiae and humans (Sc/H) (Figure 2.6B). We found that proteins annotated with the same function had significantly lower distances (Sc/Sp median = 1.04, Sc/H median = 0.92) than those annotated with different functions (Sc/Sp median = 1.08, $p<2.2e-16$; Sc/H median = 1.04, $p<2.2e-16$).

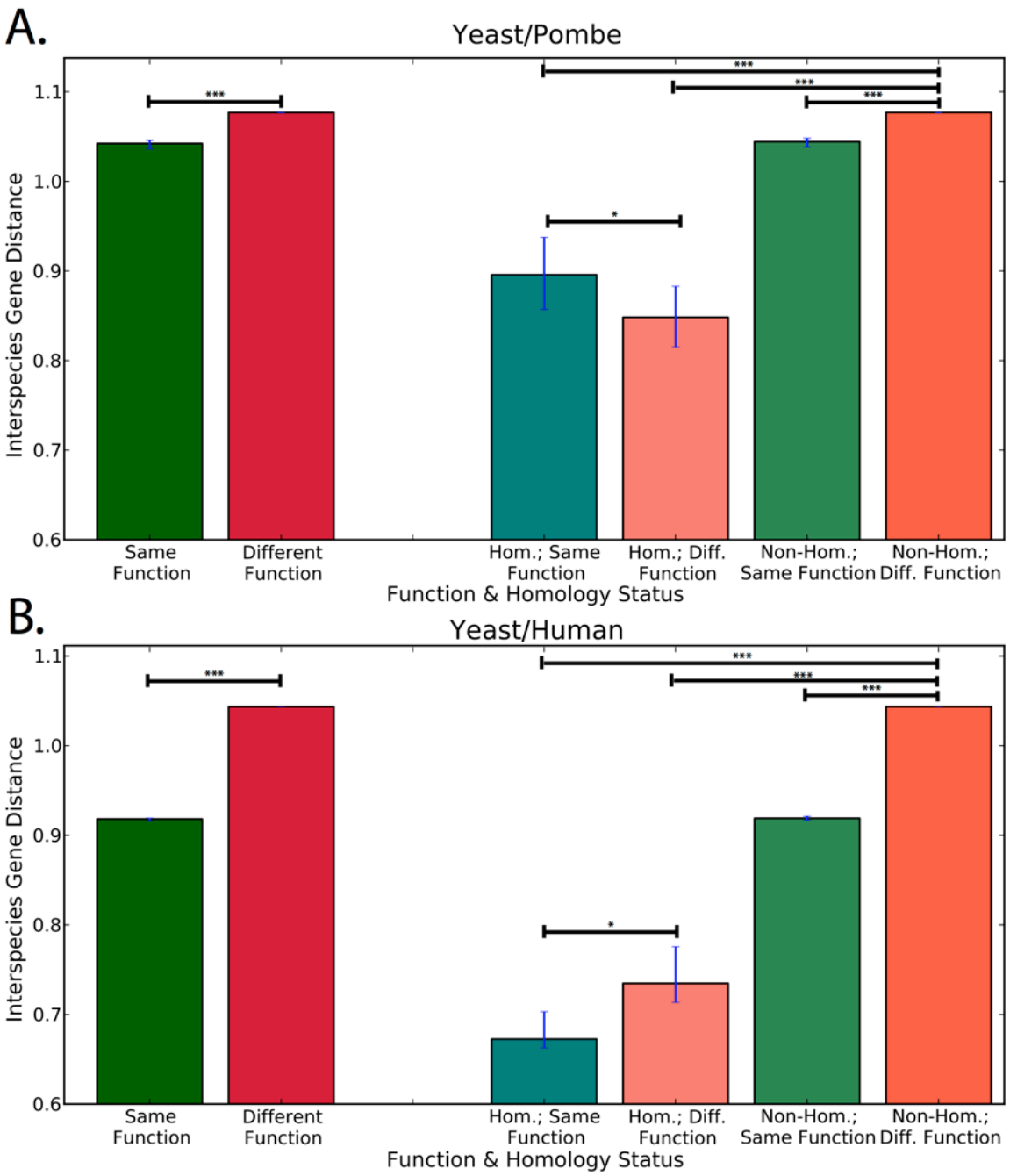


Figure 2.6: Interspecies gene-pair connectivity homology
We measure this using the Euclidean distance between vectors of single-node parameters for both genes (lower distance implies higher similarity). We find that gene pairs with the same specific function (≤100 genes annotated with that GO term) are significantly more similar to each other than gene pairs with different functions; this effect is consistent even when accounting for homology (: p<0.05; **: p<0.01; ***: p<2.2e-16. Mann-Whitney U test).*

This result holds even when orthologs are not considered. Non-orthologous genes annotated with the same function had significantly lower distances than non-orthologous genes annotated

with different functions (Figure 2.6, p<2.2e-16). We also found that orthologs had significantly lower distances than non-orthologous pairs (Figure 2.6, p<2.2e-16). These differences were consistent across all levels of functional specificity (Figure 2.7). These results suggest that network substructure, and therefore network signals, are conserved between species based on both homology and function.

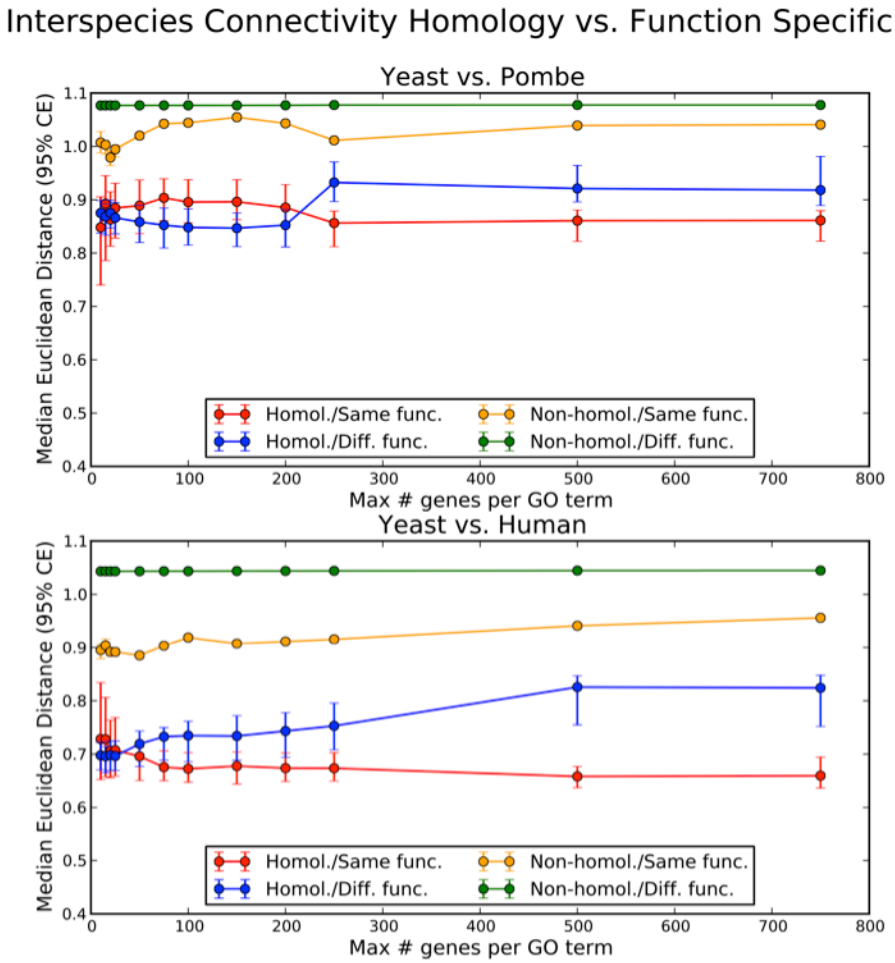


Figure 2.7: Interspecies connectivity homology vs. functional specificity
Interspecies gene-pair connectivity homology is measured using the Euclidean distance between vectors of single-node parameters for both genes (lower distance implies higher similarity). The maximum number of genes annotated by each GO term was changed to determine how specific each function is (x-axis). For each cutoff, the median distance between non-homologous gene pairs with different functions is higher than for all homologous gene pairs, and for non-homologous gene pairs with the same function.

DISCUSSION

In this chapter, we introduce the idea of *connectivity homology*, which exists when two genes share similar connectivity patterns quantified by network and graph theoretic parameters. We explored connectivity homology first by defining it in two toy networks, and then showed that orthologous nodes of networks evolved *in silico* have comparable parameters. Finally, we considered connectivity homology and its relation to genetic homology and function in *S. cerevisiae*, *S. pombe*, and *H. sapiens*. We found that homologous genes exhibit higher connectivity homology; in turn, interspecies gene pairs that share the same specific function have higher connectivity homology than interspecies gene pairs of different functions, regardless of orthology.

There are certain limitations to this exploration. In particular, our model of network evolution is oversimplified. Current network evolution theory believes that protein-protein interaction networks grow via a duplication-divergence model [57-60]. In this case, a node is duplicated, and each of its associated edges copied over with some probability. In addition, some nodes develop new edges connecting them to other nodes, which mimics spontaneous mutation generation. In a case where a node ends up with no edges, it may be considered a non-coding gene.

Although the preferential attachment model is much simpler than the duplication-divergence one, they are related. In the case of duplication-divergence, although each node has the same probability of being duplicated, “preference” is still shown because of existing edge distribution. That is to say, most nodes have a small degree, so
a duplication will not significantly affect the rank of existing nodes.

Based on these data, we hypothesize that there are connectivity patterns between pairs of genes that are indicative of a synthetic lethal relationship, and that, by using supervised machine learning, we will discover these patterns are discovered in a source species where synthetic lethality has been well-characterized and then identify them in a target species to predict synthetic lethal pairs of genes. We explore this hypothesis in the following chapter.

METHODS

Defining connectivity homology

We manually constructed the networks in Figure 2.1, then calculated their node parameters using Cytoscape [61]. We defined the connectivity profiles by binning the parameters of each species' network independently into one of three levels: low, medium, or high.

Network evolution

We evolve two types of networks: random and preferential attachment. We use modified versions of two NetworkX graph generators. The random is based on the Erdős-Rényi model [62], and the preferential on the Holme and Kim algorithm [63].

In both cases, we started with a seed network of 1000 nodes (node 1). A child is generated from a parent node by adding 100 nodes to the parent according to the appropriate evolutionary method. We generate 32,767 nodes for each method.

We seed the random network with 1000 nodes, with a probability of edge creation P=0.007. We ensure the presence of only one component by randomly wiring unattached nodes to other components.

We seed the preferential attachment network with 1000 nodes, with m=4 edges added for each node and an initial probability of $p_0 = 0.2$ for creating a triangle after adding a random edge. These parameters allowed for a similar starting density for parent nodes of both PA and RD networks.

We modified the Holme-Kim algorithm two ways. First, we updated the probability of adding a triangle with each level of the 'phylogeny,' such that $p = p_0+0.03*level$. Second, instead of attaching *m* edges to the newly generated node, we added one edge between the new node and an existing one according to preferential attachment, and then three other edges between random

pairs nodes in the network, again via preferential attachment. This allowed for a distribution of edges more similar to a protein-protein interaction network.

Although we "evolved" each parent network by 100 nodes at each step, that does not necessarily mean that the "distance" between two child nodes would be 200, as nodes may attach in similar patterns. Therefore, we calculated the actual distance between child nodes in both random and preferential attachment networks with the following equation:

$$\Delta_{c_1,c_2} = \text{sum}(\text{abs}([p_{c_1} - p_p] - [p_{c_2} - p_p]))$$

where p is a parameter (in this case, degree or betweenness centrality), p_p is the parameters of all nodes in the parent network, p_{c_1} is the parameter of the nodes in the parent network as they appear in the network of the first child, and p_{c_2} is the parameter of the nodes in the parent network as they appear in the network of the second child. We compared the children of all parent networks in this way, and plot histograms of their Δ_{c_1,c_2} distributions in Figure 2.A.1.

Comparison of network evolution

We next compared network parameters between networks of different 'evolutionary' distance. We hypothesized that, the farther the networks were in their 'phylogeny,' the less similar their parameters would be, but that networks evolved via preferential attachment would be more similar than those evolved randomly.

To do so, we calculated the degree and betweenness centrality of the original 1,000 nodes for each network in the phylogeny. Then, we sampled 100 networks from the phylogeny, choosing at least two from each level and excluding Node 1, the original network. We calculated the difference between the parameters using:

$$\Delta_{nw_x,nw_y} = \text{abs}\left(p_{nw_x} - p_{nw_y}\right)$$

where p_{nw_x} and p_{nw_y} are the degree or betweenness centrality for the original 1,000 nodes between Network X and Network Y, respectively, and Δ_{nw_x,nw_y} is a vector of length 1,000. We then compared the distribution of $\mathrm{med}(\Delta_{nw_x,nw_y})$ for each pair of networks for RD and PA networks. We tested the differences between evolution distributions using Mann-Whitney's U test [64].

Next, we calculated the Spearman correlation between p_{nw_x} and p_{nw_y} for both degree and betweenness centrality, for all pairwise combinations of X and Y in our sampled networks, for both RD and PA networks. We illustrated these results as scatterplots in Figure 2.4. We used Mann-Whitney's U to compare their distributions between RD and PA networks.

Calculation of translated network parameters

To rank-normalize data for a given species, we calculated all individual single- and paired-node parameters. Then, for each parameter, we ranked all calculated values from smallest to largest, resolving ties at random. We then divided all values by the total number of genes in the network (for single-node parameters) or the total number of gene pairs (for node-pair parameters). This resulted in all genes or gene pairs having all parameter values be a value between 0 and 1.

Similarity between connectivity vectors is indicative of shared function

We defined a vector of single-node network parameters (see Table 2.1) for each gene in the *S. cerevisiae*, *S. pombe*, and human networks. We calculated the connectivity homology of each interspecies node pair using Euclidean distance. A lower distance implies greater connectivity homology (similarity).

We first divided all gene pairs into same specific function or different specific function. We then further divided these groups into homologous/non-homologous. Specific functions were

defined as all GO terms related to process or function (excluding *molecular_function* or *biological_process*) where the number of genes annotated with that GO term in each species was less than or equal to a given cutoff. This cutoff was set to 100 at first, and then expanded to 10, 15, 20, 25, 50, 75, 100, 150, 200, 250, 500, and 750 genes per GO term.

APPENDIX

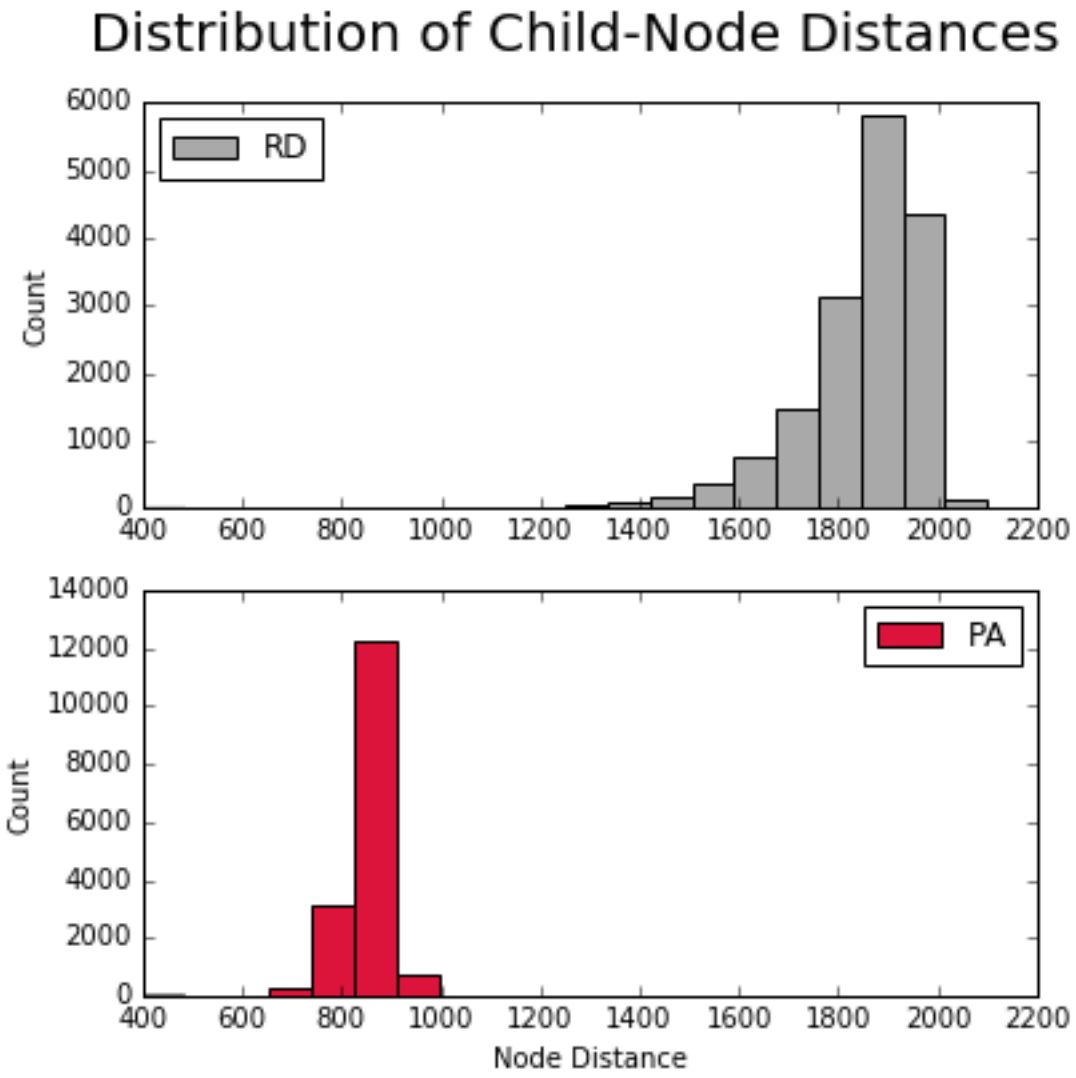


Figure 2.A.1: Distribution of distances between child networks of parents in evolved random and preferential attachment networks

We calculated an approximation of 'evolutionary distance' between all child nodes in each network model (Materials and Methods) and found that preferential attachment (PA) has significantly lower average distances than random attachment (RD) (Mann-Whitney U=268,387,251, p<2.2e-16). This is intuitive, as preferential attachment will necessarily mean that certain nodes are more likely to get picked than others.

Parameter	Context	Description	Equation
2nd degree shared neighbours	Single node	The sum of all nodes two edges away from the node of interest.	
Betweenness centrality	Single node	The sum of the fraction of paths passing through the node of interest of all shortest paths between the two other nodes.	$\sum_{s\neq v\neq t} \frac{\partial(s,t\|v)}{\partial(s,t)}$
Closeness centrality	Single node	The inverse sum of all shortest paths that originate at the node of interest.	$\frac{n-1}{\sum_{s\neq t}^{n}(\partial(s,t))}$
Communicability	Node pair	The sum of all closed walks between a pair of nodes.	$\sum_{s\neq t}^{n} \rho(s,t)$
Current-flow betweenness cenrality	Node pair	Analogous to betweenness centrality, but with all paths instead of shortest paths. Also known as random walk betweenness centrality.	$\frac{\sum_{s\neq v\neq t}\rho(s,t\|v)}{\rho(s,t)}$
Degree centrality	Single node	The fraction of edges a node has of all possible edges.	$\frac{\sum_{s\neq t}^{n}\epsilon(s,t)}{n-1}$
Eccentricity	Single node	The maximum distance from the node to any other node in the network.	$\max(\rho(s,t))$
Eigenvector centrality	Single node	The eigenvector for the larges eigenvalue of a matrix adjacency network.	$\frac{1}{\lambda}\sum_{s\neq t}^{n}\epsilon(s,t)x_t$
Inverse shortest path	Node pair	The inverse of the smallest number of edges connecting two nodes of interest.	$\frac{1}{\partial(s,t)}$
PageRank	Single node	The rank of a graph's nodes based on the incoming links.	See [65]
Shared neighbours	Node pair	The intersection of two nodes' sets of immediate neighbours (*i.e.* $\epsilon(s,v)=\epsilon(t,v)=1$)	$\sum_{s\neq t\neq v}\epsilon(s,v)\times\epsilon(t,v)$
Shared non-neighbours	Node pair	The number of nodes that are not immediate neighbours of either node of interest.	$\sum_{s\neq t\neq v}(1-\epsilon(s,v))(1-\epsilon(t,v))$

Table 2.A.1: Network parameter descriptions

When $\epsilon(s,t)=1$, there is an edge between nodes s and t. In addition, ∂ represents shortest path; ρ represents a path of any length.

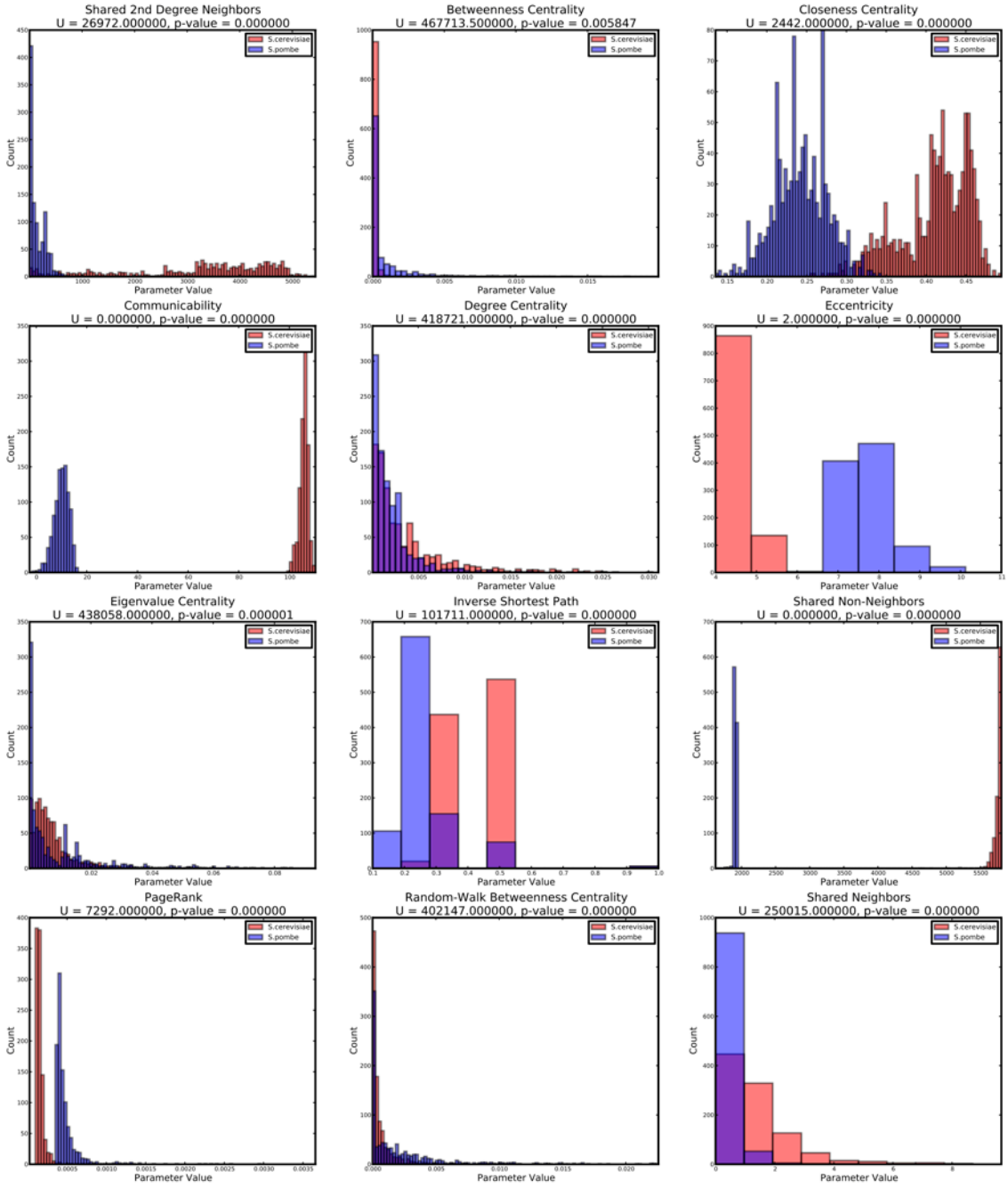


Figure 2.A.2: Distribution of network parameters for the *S. cerevisiae* (red) and *S. pombe* (blue) networks

Mann-Whitney U test indicates that the parameters are significantly differently distributed between species.

ACKNOWLEDGEMENTS

Sections of this chapter are adapted from a publication by Jacunski *et al*. in PLOS Computational Biology [66]. I thank Hossein Khiabanian for his role as a sounding board for the second section of this chapter. The expanded information in this chapter is currently in submission for publication.

CHAPTER 3 – INTERSPECIES MODELS OF SYNTHETIC LETHALITY IN MODEL ORGANISMS

INTRODUCTION

Synthetic lethality (SL) occurs when two nonessential genes cause cellular inviability after being knocked out simultaneously [22]. Although SL has mainly been studied in model organisms such as *D. melanogaster* [17] and *S. cerevisiae* [23], it can be a powerful tool for studying drug action in humans; for example, SL may guide the development of cancer combination therapy [67,68] and inform drug-drug interactions. SL interactions may differ between cellular contexts [69]; a gene pair that is SL in one cell type may not be SL in another. This can provide a tremendous therapeutic boon when two drugs targeting two gene products mimic an SL interaction in cancer cells and leave healthy cells unaffected. However, drug-induced SL interactions may also cause adverse events via unexpected cell death. Thus, mapping SL in humans is necessary to understanding mono- and polypharmacological effects.

Most gene pairs have not been interrogated for SL in humans, and several factors impede a species-wide evaluation of this interaction. These include the ethical implications of studying SL directly, the inability to discern state-specific SL interactions from global ones in experimental cell lines (*e.g.* cancer [69,70]), and – most significantly – the heavy experimental burden. Over 200 million assays would be required to determine the SL status of all human gene pairs in just a single cellular context. *In silico* methods are therefore necessary to guide the identification of SL in human systems and disease.

Previous work on leveraging model organisms to predict human SL has focused in particular on genetic homology, under the hypothesis that SL status will be maintained between

orthologous gene pairs [52]. This approach has two major limitations. First, there are only approximately 2,000 homologous genes between *S. cerevisiae* and humans (NCBI Homologene [71]). This accounts for a mere 1% of all possible human pairs, leaving the majority with no predictive data regarding SL status.

Second, genetic redundancies that developed independently in each species since deviation from a common ancestor may affect synthetic lethal status. For example, 228 gene duplication events have been suggested between *S. cerevisiae* and *S. pombe* [59] in the ~400 million years of evolution between the two species [72]; this number is certainly even higher between *S. cerevisiae* and humans. Each of these events may introduce a functional redundancy that alters SL relationships in the organism by causing a gain or loss of SL. Focusing solely on genetic homology does not account for these complexities.

In this chapter, we first evaluate the performance of genetic homology in predicting SL. We also consider structural similarity using protein structure families, domain similarity using protein domains, and functional similarity with gene ontology annotations. We additionally consider information centrality, a univariate network-based model. We show that homology, structural similarity, and information centrality are limited in their ability to predict SL.

We observe that relationships between genes and proteins, including redundancies, may be illustrated through the use of biological networks, and we hypothesize that the network connectivity profiles between two genes will better characterize their potential for an SL relationship. Therefore, we leverage the concept of connectivity homology to develop an algorithm, Species-INdependent TRAnslation (SINaTRA), that predicts interspecies SL using well-known graph properties, such as degree centrality and shortest path [14,47,52,53], and machine learning. We first develop the model in *S. cerevisiae*, and then validate it in *S. pombe*

and *M. musculus*. We show that SINaTRA significantly outperforms previously published models of predicting SL in translation, and that the method is robust to network incompleteness.

RESULTS

Previous methods of modeling synthetic lethality: genetic homology, structural similarity, and functional similarity

We began our study by considering two published methods of predicting SL, protein homology [73] and bi-nodal information centrality [59,74], and implemented the algorithms as described by the authors. In addition, we hypothesized that structural homology, domain homology, and functional homology may be able to predict SL and designed models based on these parameters for comparative analysis.

In Wu *et al*. [73], the authors constructed a model to predict SL in *S. cerevisiae*, then hypothesized that human gene pairs homologous to SL pairs in *S. cerevisiae* would also be SL in humans. We implemented the latter part of the approach and evaluated it by predicting SL in *S. pombe*. By restricting our analysis to only genes that are homologous between *S. cerevisiae* and *S. pombe*, we find a significant predictive effect (OR = 145, 95% CI: 93–219, $p < 2.2e\text{-}16$, Fisher's exact test), corresponding to an area under the receiver operating characteristic curve (AUC) of 0.60. Model performance decreased to OR = 45.9 ($p<2.2e\text{-}16$) and an AUC = 0.52 when expanding the model to include all gene pairs (*Materials and Methods*).

We next hypothesized that structural, domain, and functional similarity may be predictors of SL. We trained these models in *S. cerevisiae* and applied them to *S. pombe*. We used SCOP protein classifications to describe the former, and assigned each gene pair a value between 0 (no similarity) and 4 (same class) based on their products' structural similarity. The model was trained and tested only on pairs with SCOP data associated with both genes. Only 399 SL pairs and 109,357 non-SL pairs had SCOP data for *S. cerevisiae* (16,765,399 pairs skipped) and 2 SL/298 non-SL pairs had SCOP data in *S. pombe* (1,840,021 pairs skipped). The SCOP-based model had an AUC of 0.62. We additionally created a domain-based model from PFam [75,76]

to predict SL. Domain data exists for a larger number of proteins (9,424 SL/10,280,492 non-SL in *S. cerevisiae;* 514/1,431,764 for *S. pombe*), allowing us to score more pairs than the SCOP-based model (*Materials and Methods*). The AUC in the domain-based model was 0.56. We described functional homology using annotations from Gene Ontology (GO) (*Materials and Methods*). Functional similarity attained an AUC of 0.81.

Finally, we calculated the pairwise information centrality [74] in *S. pombe* and found no significant predictive performance identifying SL pairs (AUC = 0.46, Logistic Regression). Bi-nodal information centrality did not require interspecies translation.

We hypothesized that multivariate, network-based models of synthetic lethality can able to capture SL interactions both within and between species more accurately.

Networks successfully predict within-species synthetic lethality

We used machine learning algorithms to build two models of synthetic lethality (SL) using the connectivity profiles we derived for pairs of proteins – one for *S. cerevisiae* and one for *S. pombe*. We illustrate this in Figure 3.A.1. We trained these models using experimentally established SL gene pairs from BioGRID (N = 13,196 for *S. cerevisiae* and N = 628 for *S. pombe*) as our positive training examples. We randomly selected pairs not listed as SL in the database as non-synthetic lethal (non-SL) pairs and used these as negative examples. Our assumption that any pair without experimental evidence for synthetic lethality is non-SL will be incorrect for a small number of pairs that are SL but have not yet been investigated (*i.e.*, false negatives); however, this will introduce only negligible error due to the rarity of SL interactions (estimated 0.1% in diploid organisms [77]).

We evaluated these models using cross-validation and area under the receiver operating characteristic curve (AUC). Random forest (RF) significantly outperformed logistic regression (LR) for both *S. cerevisiae* (AUC_{RF} = 0.92, AUC_{LR} = 0.77; p<2.2e-16, De Long's Test) and *S.*

pombe (AUC_{RF} = 0.93, AUC_{LR} = 0.86; p<2.2e-16, De Long's Test) (Figure 3.1A). We found that within-species model performance is consistent regardless of normalization method (*Materials and Methods*; Figure 3.1B, C).

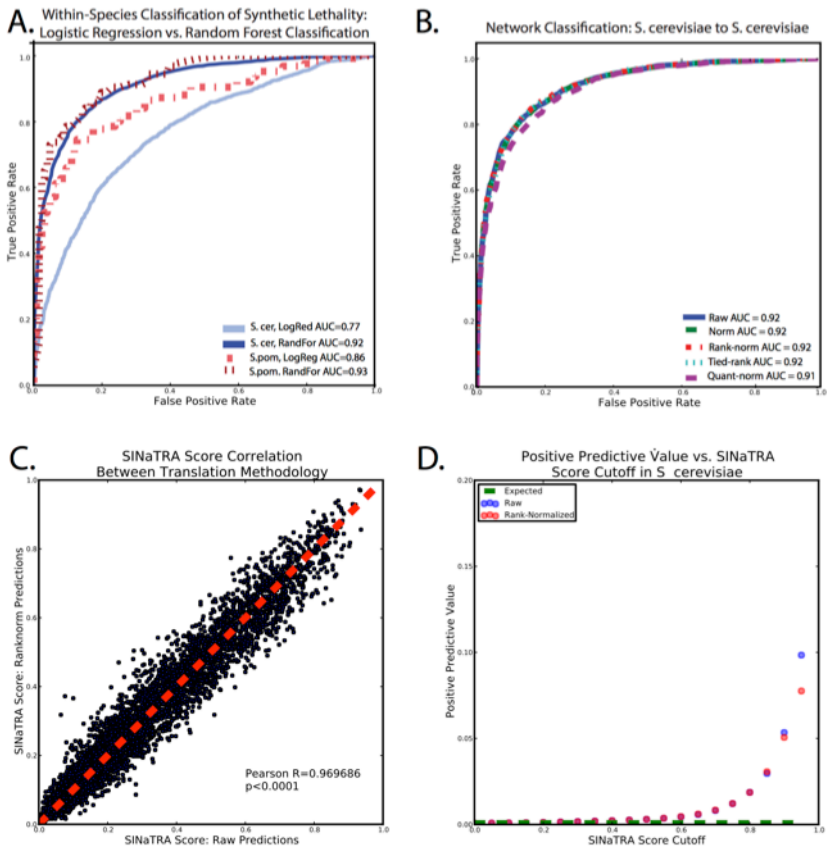


Figure 3.1: Within- and between-species classification of synthetic lethality
A.) *We performed classification of SL within two species:* S. cerevisiae *and* S. pombe*. We considered logistic regression (LogReg) vs. random forest (RanFor) to pick the more robust method. We found that random forest significantly outperformed logistic regression in both species (p<0.0001, De Long's Method).* B.) *Receiver operating characteristic for within-species classification of SL in* S. cerevisiae *using raw (red) and rank-normalized (yellow) data; both achieved an AUC of 0.91. In addition, SL labels were permuted (blue), achieving an AUC no better than chance.* C.) *Correlation between 5,000 gene pairs' SINaTRA scores using raw and rank-normalized data. Pearson R correlation is 0.97 (p<0.0001).* D.) *SINaTRA score cutoff vs. positive predictive value. We computed PPV at each SINaTRA score cutoff (all gene pairs with SINaTRA score greater than the cutoff were considered to be SL), and found that it increased to approximately 0.1 at a SINaTRA score cutoff of 0.95.*

Translation of synthetic lethality between *S. cerevisiae* and *S. pombe*

In order to create network models of synthetic lethality in translation, we developed the SINaTRA algorithm (Species INdependent TRAnslation). The schematic is illustrated in Figure 3.2.

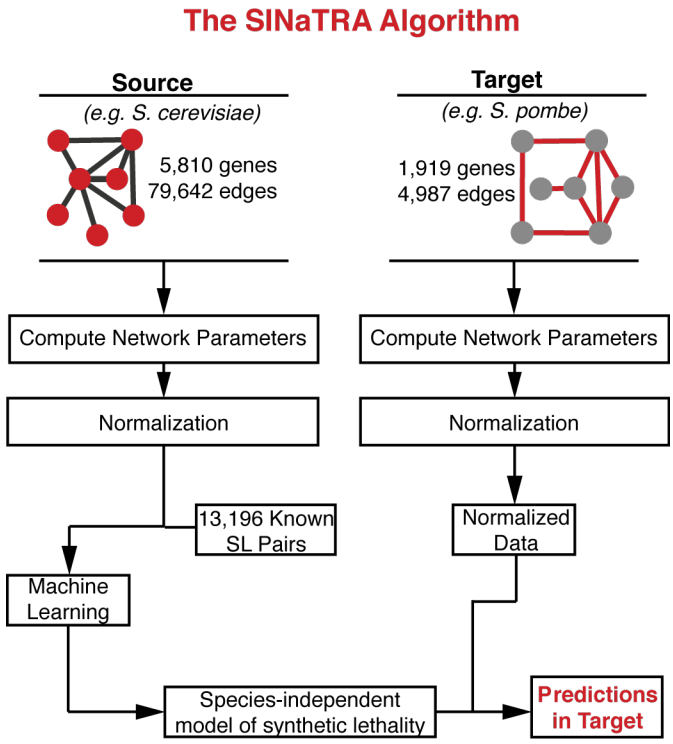


Figure 3.2: Schematic of the SINaTRA algorithm
We begin with the PPI networks of both our source and target species, calculate the network parameters (independently), and normalize the values of all parameters. Next, we use machine-learning methods on the normalized network parameters of our source species as well as experimentally derived labels of synthetic lethality to construct a species-independent model of SL. Finally, we apply this model to the normalized network data of our target species in order to attain SL predictions in our target.

To apply SINaTRA to *S. cerevisiae* and *S. pombe*, we created two translational, network-based models that use data from a source species to infer the SL status of gene pairs in a target species. The first was trained on *S. cerevisiae* to predict SL in *S. pombe*; the second was trained on *S. pombe* to predict in *S. cerevisiae*. For each model, we randomly selected an equal number of non-SL pairs as SL pairs (13,196 for *S. cerevisiae*; 628 for *S. pombe*). We built random forest models with 100 trees for each species. We evaluated these two models for their ability to predict

SL gene pairs in the target species. Each model generates a SINaTRA score for each pair between 0 (predicted non-SL) and 1 (predicted SL).

Using *S. cerevisiae* as the source and *S. pombe* as the target, we found that untranslated parameters resulted in poor inter-species SL prediction (AUC = 0.67). We tested all methods of normalization in translation (Figure 3.A.2) and found that the model significantly improves with any translational method with rank normalization performing best (AUC = 0.86; p<2.2e-16, De Long's method) (Figure 3.3A). We also found that parameter normalization improved the precision from 50% to 98% at a recall rate of 30% (Figure 3.3B) in our testing data. The translated model also significantly outperforms the untranslated one when using *S. pombe* as the source species and *S. cerevisiae* as the target ($AUC_{translated}$ = 0.74, AUC_{raw} = 0.67, p < 2.2.e-16, DeLong's method, Figure 3.A.3).

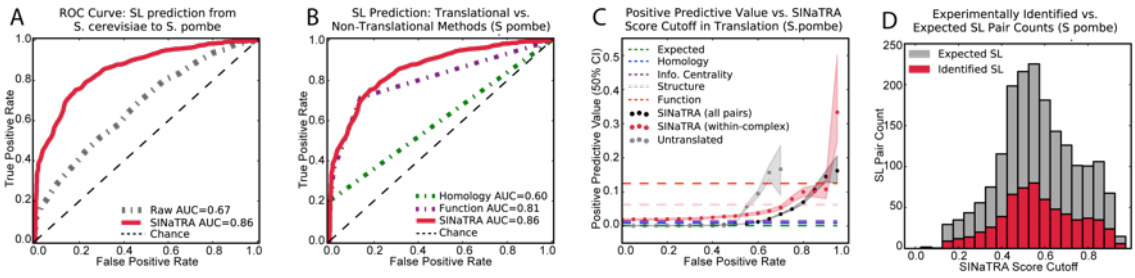


Figure 3.3: SINaTRA predictions, *S. cerevisiae* to *S. pombe*
A.) *Receiver operating characteristic (ROC) curves for classification of SL/non-SL gene in* S. pombe *using* S. cerevisiae *as source. Comparison of untranslated ("raw") parameters (gray, AUC = 0.67) and the translated parameters used in SINaTRA (red, AUC = 0.86).* B.) *ROC curve of SL predictions using SINaTRA (AUC = 0.86) compared functional homology of gene pair products (AUC = 0.81) and gene homology (AUC = 0.60). The model based on gene homology was created using only gene pairs with homology data.* C.) *Positive predictive value (PPV) of all (dark gray) and within-complex (red) gene pairs. When accounting for the expected ratio of SL:non-SL (1:1000), a SINaTRA score threshold of 0.95 yields a median PPV of 17% (a 170-fold increase over what is expected by chance). At 0.85, the PPV drops to 7%. PPV increases in within-complex gene pairs, suggesting that this may be a good initial filter for experimental validation.* D.) *At each SINaTRA score cut-off, we plot the number of experimentally identified SL pairs in that bin (red), as well as the number we expect to find at each level (gray).*

SINaTRA outperforms translation-free and non-network methods

After evaluating SINaTRA in *S. pombe* and *S. cerevisiae*, we compared its performance to those of models based on genetic homology and functional similarity. We show ROC curves of each previously discussed methods and compared it to that of SINaTRA (Figure 3.3B) and use the AUC as a summary performance statistic. We additionally compared the performance of SINaTRA to domain similarity, structural similarity, and information centrality (Figure 3.A.4). We found that SINaTRA had significantly higher AUC than any other method we considered ($p<2.2e-16$, DeLong's test, all comparisons).

We then estimated the PPV for all gene pairs at 20 SINaTRA score thresholds (Figure 3.3C); the ratio of SL:non-SL pairs was held at the expected ratio (1:1000 [77]). We found a significant improvement over chance (Odds ratio = 121.1, p = 2.72e-32, Fisher exact test). For example, at a SINaTRA score of 0.85, the PPV is approximately 7% — 70 times higher than expected by chance. It increases to 17% at a cut-off of 0.95, corresponding to a 170-fold increase over chance. In comparison, the untranslated method of SL prediction rises to a PPV of 17% at a cut-off of 0.65 and dips sharply at 0.70. No gene pairs receive a score higher than 0.70 in the untranslated model.

We also found that no model out of genetic homology, functional similarity, structural similarity, or bi-nodal information centrality had a gene pair score higher than 0.05; therefore, we first identified which cut-off would provide the highest PPV, and plotted each value as dotted lines in Figure 3.3C. We also provide a direct comparison between true and false positives and negatives for SINaTRA compared to homology in Figure 3.A.5. We found that, for all homologous pairs, the model achieves an OR of 144.9 ($p<2.2e-16$, Fisher's exact test), corresponding to an AUC of 0.60. In contrast, SINaTRA achieves an OR of 929.6 ($p<2.2e-16$, Fisher's exact test) and a corresponding of AUC = 0.91 (Figure 3.A.4) when using a SINaTRA

cutoff of 0.85 on this same subset of pairs (any pair where at least one gene is not in the network is given a SINaTRA score of 0).

When we expand our data to the 'whole genome,' comprising all possible pairs from the set of Homologene and network genes (Materials and Methods), the homology-based method attains a lower, but significant, OR (OR = 60.1, p<2.2e-16) and an AUC of 0.52. A similar expansion in SINaTRA yields an OR of 304.2 (p<2.2e-16) when considering gene pairs with SINaTRA scores ≥ 0.85 as SL (Figure 3.A.5).

We used Analysis of Variance (ANOVA) to evaluate the independent contributions of the methods when combined with SINaTRA. We found that genetic homology, protein similarity, and univariate connectivity contributed no significant improvement in performance over the SINaTRA-only model. This result held for genetic homology even when considering only the subset of ~2 million gene pairs that are homologous between *S. cerevisiae* and *S. pombe* (X^2 = 407.66, p = 0.64). Functional similarity, represented by gene ontology [GO], significantly improved the SINaTRA model (X^2 = 445.09, p<2.2e-16, ANOVA) (Table 3.A.1).

SINaTRA identifies missing synthetic lethality in *S. pombe*

We estimated the number of previously unidentified synthetic lethal pairs at 20 SINaTRA thresholds (Figure 3.3D). For example, at a SINaTRA ≥ 0.85, we expect to find 177 SL pairs but only 65 have previously been experimentally identified. 1,759 gene pairs have a score of 0.85 or greater in *S. pombe*, corresponding to an expected hit rate of 1 in 15.

Synthetic lethality is enriched in protein complexes

We identified all within-complex gene pairs in *S. pombe* (N = 5,806, Materials and Methods) and found 46 experimentally identified SL pairs. We found that the positive predictive value (PPV) is consistently higher in within-complex pairs, reaching 0.27 at a cut-off of 0.95 (Odds ratio = 148.4, p = 1.33e-37, Fisher exact test).

Translated models are robust to network completeness

One network property that varies significantly between species is density, defined as the fraction of edges that exist in the network compared to the total possible number of edges [78]. Density can be affected by at least two factors: network size (see Note 3.A.1) and network completeness. A complete network would have all known edges elucidated, so that every non-edge would be certain to indicate a non-interaction, rather than being either a non-interaction (true negative) or a lack of information on that interaction (false negative).

Although we cannot be certain of the underlying reason for the differences, the densities of the networks used in this dissertation do vary widely; *S. cerevisiae* has one of the highest (density = 0.004), while those of *S. pombe*, *M. musculus*, and *H. sapiens* are lower, with densities of approximately 0.003, 0.001, and 0.001, respectively. We tested the extent to which SINaTRA was sensitive to these differences by ablating the target network (*S. pombe*) to densities between 90% and 50% of the original network (*Materials and Methods*). The lowest density approximates that of the human and mouse PPI networks. Untranslated parameters achieve AUCs between 0.43 and 0.60 for all ablated graphs. We found that ablation by 10% decreased rank-normalized AUC from 0.86 to 0.83, and ablation by 50% dropped the AUC to 0.79 (Figure 3.A.6).

Prediction of synthetic lethality is not driven by node popularity

Higher degree nodes are more likely to be studied, and more popularly studied genes may also be more likely to have been tested for synthetic lethality. As a measure of this potential bias, we defined a normalized popularity measure (degree/popularity), where popularity is the number of times a particular gene appears in the BioGrid database. Although SINaTRA score is correlated with degree and, thus, popularity, it is not correlated with normalized popularity in either *S. cerevisiae* or *S. pombe* (Figure 3.A.7). Further, we found that the predictive

performance of SINaTRA is independent of each of the three measures (degree, popularity, and node popularity) according to ANOVA ($p < 0.0001$ for both comparisons).

Prediction of synthetic lethality in mice

We used the model trained on *S. cerevisiae* as the source species and *M. musculus* as the target. There is no comprehensive database of SL in mouse, and only nine mouse SL pairs are recorded in BioGrid. Of these, eight were predicted to be SL with a score ≥0.5; five had scores ≥0.70. SL prediction achieved an AUC of 0.937, significantly outperforming GO similarity (AUC = 0.687; p = 1.556e-11, DeLong's method).

DISCUSSION

In this chapter, we present a computational method, Species INdependent TRAnslation (SINaTRA), for predicting synthetic lethal (SL) relationships in any species with an available protein-protein interaction (PPI) network. Our approach uses SL data from *S. cerevisiae*, the most well-characterized organism for this interaction, to train a statistical model that identifies network connectivity profiles indicative of synthetic lethality. Once trained, the model can be applied to any other species for which PPI data exist. The model takes a feature vector of PPI network parameters for a gene pair as its input, and returns a probabilistic score between 0 and 1 that we deem the SINaTRA score. These scores represent the likelihood of an SL relationship between the two genes.

We validated our method by predicting which pairs are likely to be SL in *S. pombe*, another species for which a large number of SL pairs are known. We additionally tested on *M. musculus*, for which fewer pairs are known. Our approach significantly outperforms others we tested. Most notably, our method does not rely on any knowledge of gene structure, sequence, or function; instead, it uses only the connectivity patterns exhibited by synthetic lethal pairs of genes as they appear in a protein-protein interaction network. Future work may focus on integrating other sources of knowledge with the goal of improving predictive performance and understanding the role of connectivity under different functional conditions.

Previous interspecies methods of predicting synthetic lethality

Previous work on interspecies SL prediction has focused on the use of genetic homology [73]. We found that the method has fairly high predictive power between *S. cerevisiae* and *S. pombe* when considering only gene pairs with known homology (Figure 3.3B). Unfortunately, many genes have no known homology information and, because of this,

model performance suffers when considering all interspecies gene pairs. Genes with multiple homologues further complicate prediction, as they result in ambiguous predictions. In an effort to address some of these challenges with using established orthologs, we also implemented two additional methods: one using shared structural domains, and one derived from structural families. Neither method outperformed SINaTRA. The most successful comparison method was the number of shared functional annotations in the Gene Ontology (AUC = 0.81), which performed almost as well as SINaTRA (AUC = 0.86). We additionally found that the information contained in the functional annotations and SINaTRA was not redundant, suggesting that a model that combines connectivity profiles with functional annotations may yield improved performance.

Connectivity homology as a novel method for predicting synthetic lethality

We validated our connectivity-homology-centered approach in two species where SL has been experimentally explored (*S. cerevisiae* and *S. pombe*). We found that our approach, called SINaTRA, significantly outperformed published methods at predicting SL genes in the target species and we achieve precision up to 150 times higher than expected by chance. This precision increased to over 250 times higher than chance when using additional biological priors.

False positive rate in predictions of synthetic lethality

For very rare biological phenomena, it is essential to consider the false positive rate of any experimental or computational approach. An unbiased random selection of gene pairs would yield approximately one synthetic lethal pair for every 1,000 tested. If biased by biological priors, such as limiting the analysis to pairs of genes with products in the same protein complexes, this yield may increase 8-fold, to one out of every 125 pairs tested.

The SINaTRA score we present can also be used as a biological prior. In this case, it is the connectivity pattern of the pair of proteins that makes them more likely to participate in a

synthetic lethal interaction. For example, a score of 0.85 or greater would yield approximately 1 SL for every 10 pairs tested. Combined with other biological priors, the SINaTRA score can be a powerful tool for directing experimental exploration of synthetic lethality. Figure 3.3D illustrates this expected hit rate versus the number of experiments that would be necessary. These scores can be used to guide experimental exploration depending on the throughput and cost of the experimental approach.

Limitations

Our method for predicting SL relies on the availability of protein-protein interaction data. Due to the existence of high-throughput experimental techniques, such as tandem affinity purification and yeast two-hybrid, these are some of the most widely available -omic data. However, comprehensive networks are only available for a handful of species. Future work with this method may be but served by integrating other available data, such as genetic sequence or gene expression. These other data sources may help address the issue of context-specificity in our predictions.

In this study, we used 12 distinct graph theoretic parameters to describe each gene pair. The choice of these parameters was based on what was available and has been used in prior work, and is not an exhaustive list. Other methods for computing connectivity may be incorporated in future versions of the algorithm, such as spectral methods.

Overall, we believe this section has shown the utility of connectivity homology to the prediction of synthetic lethality between species, as long as both species have well-populated protein-protein interaction networks, and one of the species has been interrogated thoroughly for synthetic lethality. We find that even ~700 SL pairs are sufficient for constructing a successful model, as evidenced by *S. pombe*; however, at this time, *S. cerevisiae* remains the best source species, with approximately 13,000 SL pairs as of a 2013 database.

In the next chapter, we will apply our SINaTRA model built on *S. cerevisiae* data to predict synthetic lethality in human networks.

METHODS

Previous methods of modeling synthetic lethality: genetic homology, structural similarity, and functional similarity

We downloaded protein homology data from Homologene [71], protein structure data from SCOP [79,80], and GO data from Entrez [56,81]. We used PFam [75,76] data for protein domain similarity; IDs were mapped to Entrez gene IDs for *S. cerevisiae* and *S. pombe* using DAVID [82,83]. We calculated bi-nodal information centrality for each gene pair based on Kranthi *et al*. [74].

In order to create the homology-based model, we replicated a previous paper [73] that defined a gene pair as SL if its homologous pair in another species is SL. Gene pairs were defined as SL if the homologous pair in the source species was SL. In the case of multiple homologous pairs in the source species, gene pairs were classified as SL if at least one of the homologous pairs was known to be SL. Homology-based models use only genes with known homologs between the two species of interest. Whole-genome, homology-based models are the union of all genes in the homologous dataset with all genes that appear in our protein-protein interaction network. Genes with no known homologs are given a feature value of 0.

Protein similarity was defined using values between 0 (no match) and 4 (same class) according to SCOP annotations. Functional similarity was defined using GO process and function terms, excluding "*molecular_function*" and "*biological_process*." Gene pairs were assigned a value based on the number of overlapping GO terms assigned to each gene. Using PFam domain data, we used the size of PFam ID overlap (range: [0,8)) for within-species gene pairs. For SCOP-, GO-, and PFam-based models, we trained the logistic regression model on *S. cerevisiae* and applied it to *S. pombe*. The homology-based model was already "translated," and the model was trained and tested in *S. pombe* alone using logistic regression and five-fold cross-

validation. Information centrality does not require translation and was calculated in *S. pombe* alone; the model was constructed using logistic regression and tested with five-fold cross-validation.

Calculation of translated network parameters

In order to rank-normalize data for a given species, we calculated all individual single- and two-node parameters. Then, for each parameter, we ranked all calculated values from smallest to largest, resolving ties at random. We then divided all values by the total number of genes in the network (for single-node parameters) or the total number of gene pairs (for node-pair parameters). This resulted in all genes or gene pairs having all parameter values between 0 and 1.

In Figure 3.1.B, we mention three other methods of translation: Normalized, tied-rank normalized, and quantile normalized. Regular normalization of a parameter returns each value divided by the maximum value of that parameter, such that each value is between 0 and 1. Tied-rank normalization assigns median rank to all equal values, and then normalizes single-node parameters by the number of genes in the network, and node-pair parameters by the total number of pairs. To account for different-sized networks in quantile normalization [84], we upsampled parameter values.

Building connectivity-homology-based models of synthetic lethality

We generated PPI networks using data gathered from BioGrid [54]; each node represents a gene, while edges represent a physical interaction between gene protein products. We pruned all disconnected nodes to ensure one connected component.

BioGrid additionally provided SL data used in this investigation. *S. cerevisiae* had over 14,000 unique SL pairs and *S. pombe* have over 700, while *M. musculus* has 8 pairs. Gene pairs may have one of two classes: SL or non-SL. Because of the scarcity of SL pairs, pairs not explicitly labeled as SL are considered non-SL.

We used the NetworkX (version 1.8.1) package for Python [85] to calculate all network parameters except shared neighbours, shared non-neighbours, and shared 2nd-degree neighbours, which were elucidated from adjacency matrices for each network. All single-parameter classifiers employ logistic regression due to its high interpretability and simple nature. We implemented multi-parameter classifiers using random forests [86], which are accurate and efficient on large datasets, as well as resistant to over-fitting data. We used five-fold cross-validation in classifier construction, where training occurs with 80% of the data, and classifier evaluation uses the remaining 20%. Finally, to avoid positional bias in case of a single node having exceptionally high values, we shuffled the order in which each single-node parameter appears. We calculated parameter importance using the built-in function from Python's sklearn package [87].

Networks successfully predict within-species synthetic lethality

We predicted SL within a species using the network parameters defined in Table 1.2 without any normalization (raw) as features of the classifier, and experimental data from BioGrid [54] as the known classes. From these, we performed five-fold cross-validation by randomly selecting 1/5 of the data on which to train our classifier, and testing it on the remaining 4/5. We trained models using either logistic regression or random forest.

Translation of synthetic lethality between *S. cerevisiae* and *S. pombe*

To predict synthetic lethality, we trained classifiers on raw and translated parameters of our source species, using SL status downloaded from BioGrid as labels. We then applied the classifier to data from our target species. Here, *S. cerevisiae* is the source species, and we used its network parameters to train classifiers. *S. pombe* is the target species. Classifier inputs were vectors of network parameters.

SINaTRA outperforms translation-free and non-network methods

Synthetic lethality is expected to occur in 1/1000 gene pairs in diploid organisms [77]; therefore, the positive predictive value (PPV; the fraction of true positives out of all called positives) expected by chance is 0.001. We calculated the PPV on all *S. pombe* gene pairs, and on all gene pairs in the same complex. We selected 1000X the number of non-SL pairs as SL and bootstrapped the 99% CI of the PPV for both untranslated and SINaTRA-based predictions. To calculate PPV at each cutoff C, gene pairs with SINaTRA $\geq$ C were considered to be SL, while pairs with SINaTRA < C were considered non-SL.

Complex membership was identified by using the Entrez GO database and filtering all GO terms that contained the word "complex" and were in the "component" category. This amounted to 8,365 pairs, of which 5,806 appeared in our network. 46 of these were experimentally known SL pairs, leaving a ratio of approximately 3:400 SL:non-SL. We estimated that, because many SL pairs are unknown in *S. pombe*, the ratio of SL:non-SL in within-complex pairs will be approximately 1:50, and selected SL:non-SL pairs in
a ratio of 1:50 in order to estimate within-complex PPV. This simulation was performed 1,000 times to identify the 99th percentile CI.

We additionally plotted the PPV of SL prediction using genetic homology, structural similarity, functional similarity, and information centrality. The expected PPV of all of these were calculated using SL:non-SL gene pairs in ratios of 1:1000; because the cut-offs occurred in a range significantly smaller than [0,1], we selected the cut-off that would provide the optimal PPV for the given model (all pairs), then calculated the PPV when adjusting for SL:non-SL ratio. The PPV of genetic homology was calculated using only *S. pombe* pairs that have homologs in *S. cerevisiae*.

We identified the true and false positives and negatives for homology and whole-genome homology as follows: if the input score was >0 and the target species pair was SL, it was a true positive; if non-SL, it was a false positive. If the input score was 0 and the target species was non-SL, it was a true negative; else, if SL, it was a false negative. In whole-genome models, all node pairs with no homology information for at least one node were given a score of 0. Odds ratios were calculated using confusion matrices of form [[TP,FP],[FN,TN]] and Fisher's exact test.

For whole-genome SINaTRA methods, if the gene pair SINaTRA score ≥ given cutoff and the target species pair was SL, it was a true positive; else, if non-SL, it was a false positive. If the gene pair SINaTRA score < given cutoff and the target species pair was non-SL, it was a true negative; else, if SL, it was a true positive. In a whole-genome SINaTRA model, nodes that appeared in the Homologene database but not in the network were assigned SINaTRA scores of 0.

We identified the expected number of unidentified SL pairs in *S. pombe* by taking the PPV at each SINaTRA cutoff and multiplying it by the number of putative hits at that cutoff. We then transformed this cumulative plot into bins, such that for cutoff C, the number in that bin represents all expected pairs with C ≤ SINaTRA < C+0.05.

Translated models are robust to network completeness

We ablated the *S. pombe* network to 90, 80, 70, 60, and 50% of its original size by removing (100-N)% edges at random. We trained a random forest classifier on the complete *S. cerevisiae* network and tested it on the ablated *S. pombe* networks and measured classifier success again using AUROC.

Prediction of synthetic lethality is not driven by node popularity

We plotted the median SINaTRA score of genes in *S. cerevisiae*, *S. pombe*, and humans by the node's degree, popularity (the number of times it appeared in the BioGRID database), and normalized popularity (degree/popularity). We calculated the Spearman correlation coefficient for all plots, for all species.

Prediction of synthetic lethality in mice

We predicted SL pairs in mice as we did with *S. pombe*, using *S. cerevisiae* as the source species.

Statistical analyses and software

We calculated network parameters using the NetworkX version 1.8.1. We performed statistical analysis in R version 3.0.2. De Long's test for comparing ROC curves was implemented using the pROC library [88]. Scripts use Python version 2.7.5. Graphics were generated using Python's Matplotlib [89]. BioGrid release 3.2.104 was used in all analyses.

APPENDIX

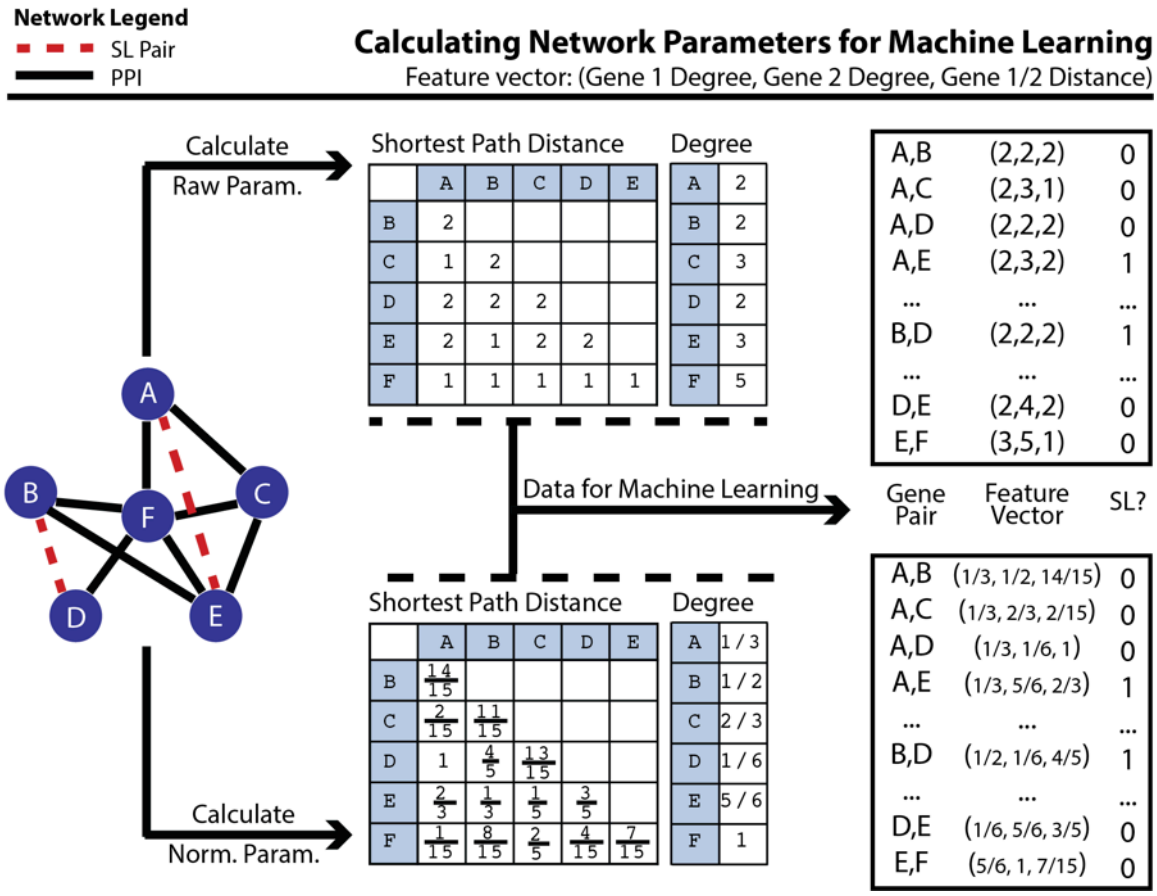


Figure 3.A.1: Calculating network parameters for machine learning
We illustrate the creation of network-based classifiers using untranslated data (top) and rank-normalized (translated) data (bottom).

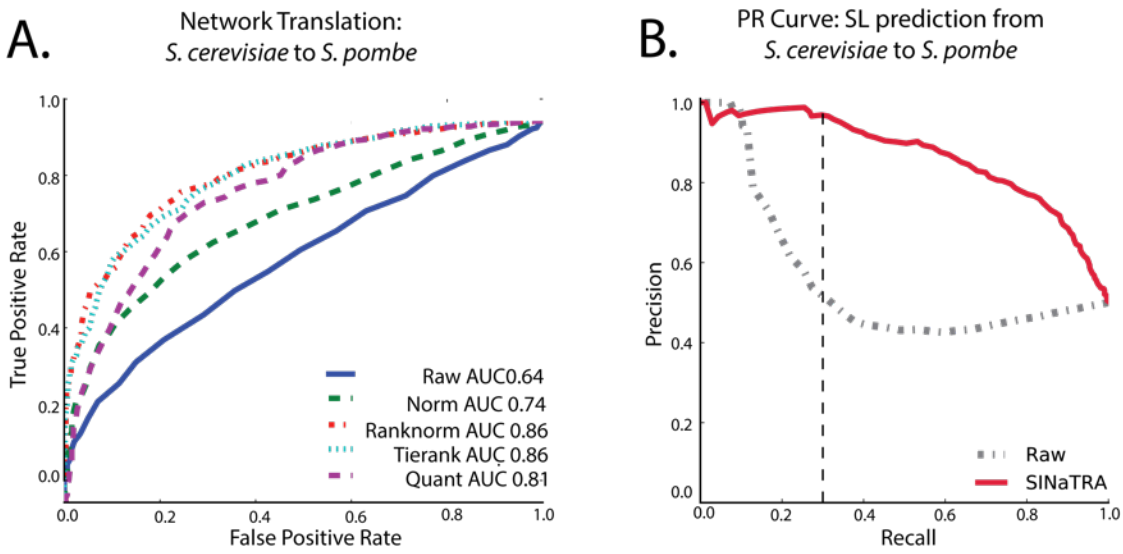


Figure 3.A.2: Prediction of synthetic lethality from *S. cerevisiae* to *S. pombe*
A.) Normalization method performance in SL prediction from S. cerevisiae *to* S. pombe. *Normalization methods are described in Materials and Methods. B.) Precision-recall curves for SINaTRA (red) and untranslated (blue).*

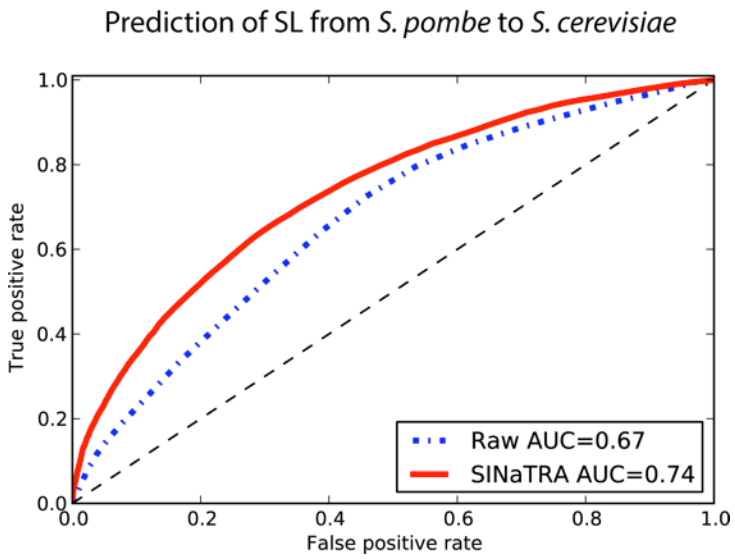


Figure 3.A.3: Prediction of synthetic lethality from *S. pombe* to *S. cerevisiae*
The black dotted line represents expected ROC by chance. Raw and SINaTRA ROC curves were significantly different (DeLong's test).

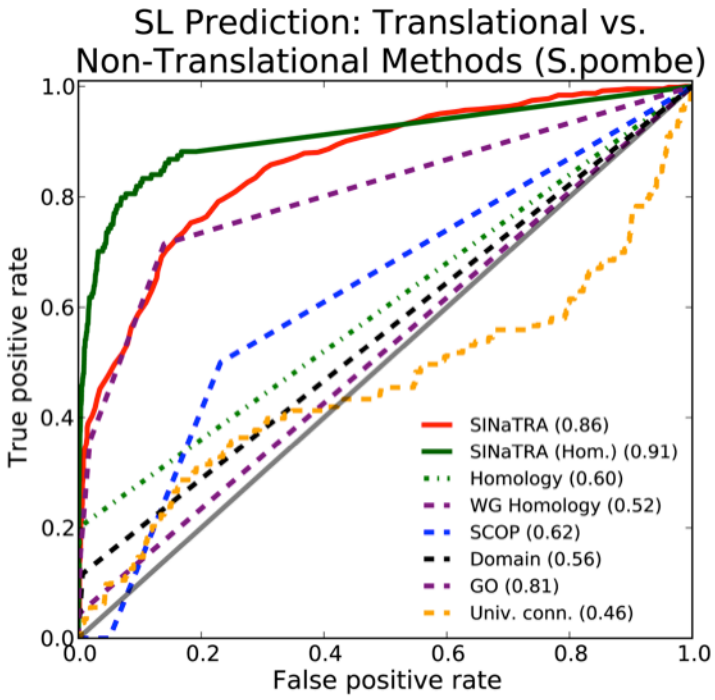


Figure 3.A.4: Prediction of synthetic lethality using translational and non-translational methods
We create classifiers based on genetic homology (AUC = 0.60), genetic homology extrapolated to the whole genome (WG Homology; AUC = 0.52), protein domain (PFam; AUC = 0.56), protein structure (SCOP; AUC = 0.62), bi-nodal information centrality (AUC = 0.46), and function (GO; AUC = 0.81), and compare these performances to SINaTRA (AUC = 0.86) and SINaTRA restricted to only pairs existing in the homology database (SINaTRA (Hom.); AUC = 0.91) when predicting SL in S. pombe.

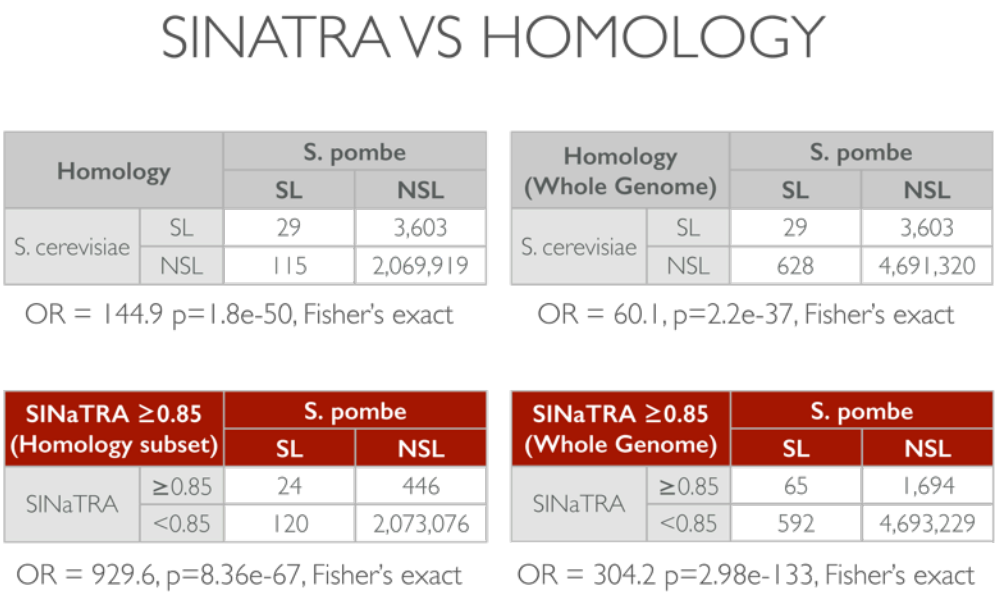


Figure 3.A.5: SINaTRA vs. homology

For each table, the upper left corner is true positives (TP); upper right is false positives (FP); bottom left is false negatives (FN); and bottom right is true negatives (TN). We found that the number of true positives, as well as the PPV, is significantly higher in SINaTRA-based methods than homology-based ones. See Materials and Methods for details.

Data	Individual AUC	Model+ SINaTRA	ChiSq	p-value
SINaTRA	0.8603	-	-	-
Genetic homology	0.519	0.8603	0.16962	0.6804
Genetic homology (homologs only)	0.5171	0.8801	0.21458	0.6432
Structural similarity	0.5016	0.8602	1.5494	0.2132
Functional similarity (binary)	0.7876	0.8981	407.66	<2.2e-16
Functional similarity (discrete)	0.8069	0.8958	445.09	<2.2e-16
Univariate connectivity	0.4463	0.8603	0.0014578	0.9695

Table 3.A.1: SINaTRA vs. other models of predicting SL

Columns 2–3 represent AUCs of models based on non-translational or non-network methods of predicting SL, and those methods plus SINaTRA. Columns 4–5 describe results of ANOVAs of nested general linear models of SINaTRA, then SINaTRA plus each of the methods. Only functional similarity provides an improved model when combined with SINaTRA.

Note 3.A.1: Density of biological networks

Here, we consider the density of biological networks of different species. The density of a network is defined as:

$$\rho_{NW} = \frac{E}{\frac{N(N-1)}{2}}$$

where E is the number of edges in the network and N is the number of nodes. If a network grows by M nodes and μ edges, it has $E_2 = E_1 + \mu$.

In order for a network to maintain its density, we have:

$$\rho_{NW_1} = \rho_{NW_2}$$

$$\frac{E_1}{\frac{N(N-1)}{2}} = \frac{E_2}{\frac{(N+M)(N+M-1)}{2}}$$

$$\frac{E_1}{N(N-1)} = \frac{E_1 + \mu}{(N+M)(N+M-1)}$$

$$\mu = E_1 \frac{M(M+2N-1)}{N(N-1)}$$

If we expect the network to grow by one node (*i.e.* $M = 1$), this becomes:

$$\mu = E_1 \frac{2}{N-1}$$

Given that $\rho_{NW} = \frac{E}{\frac{N(N-1)}{2}}$, this can be rearranged to:

$$\mu = N\rho_{NW_1}$$

If we consider the *S. cerevisiae* network, with $N = 5{,}810$, $M = 79{,}642$, and $\rho_{NW_1} = 0.004$, this means that adding one node will require $\mu = N\rho_{NW} = 27.4$ edges. However, the median degree in the network is 12.5, and the mode is 3. Given the duplication-divergence model of network evolution [57], each node is equally likely to be duplicated; therefore, it is likely that the next node added will decrease the density of the network. This suggests that the larger the network, the less dense it will be.

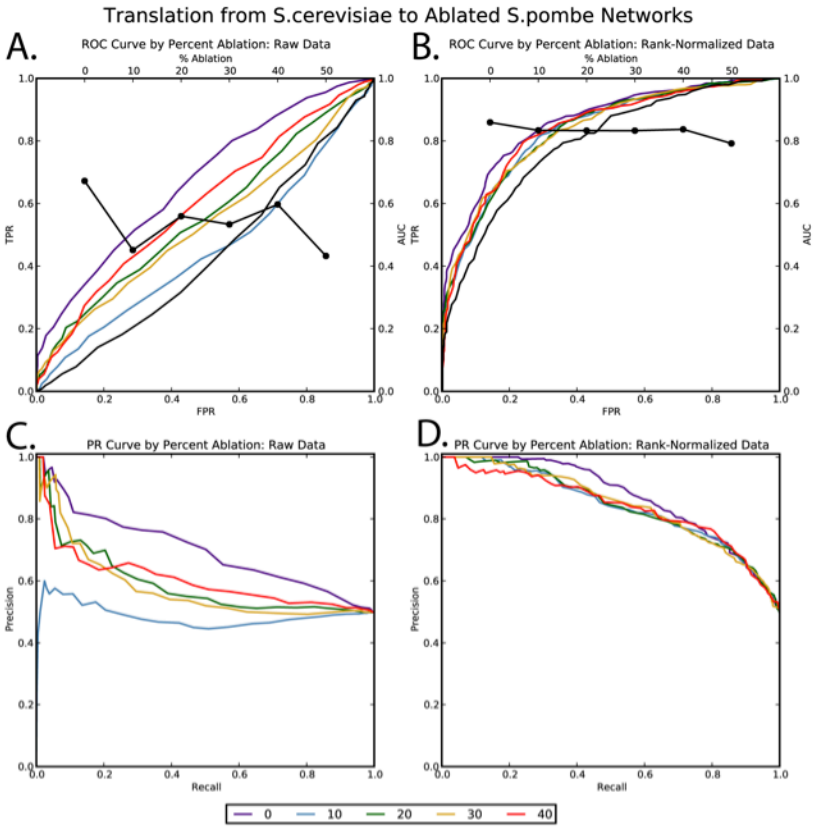


Figure 3.A.6: Network ablation and the prediction of synthetic lethality
A.) *SL prediction from full* S. cerevisiae *to ablated* S. pombe *networks using untranslated parameters. Black line represents AUC, while coloured lines represent ROC; red is highest ablation (50%), while violet is lowest (10%).* B.) *SL prediction from full S. cerevisiae to ablated* S. pombe *networks using SINaTRA. Black line represents AUC, while coloured lines represent ROC; red is highest ablation (50%), while violet is lowest (10%).* C.) *Precision-recall curves of SL prediction from full S. cerevisiae to ablated* S. pombe *networks using untranslated parameters.* D.) *Precision-recall curves of SL prediction from full S. cerevisiae to ablated* S. pombe *networks using SINaTRA.*

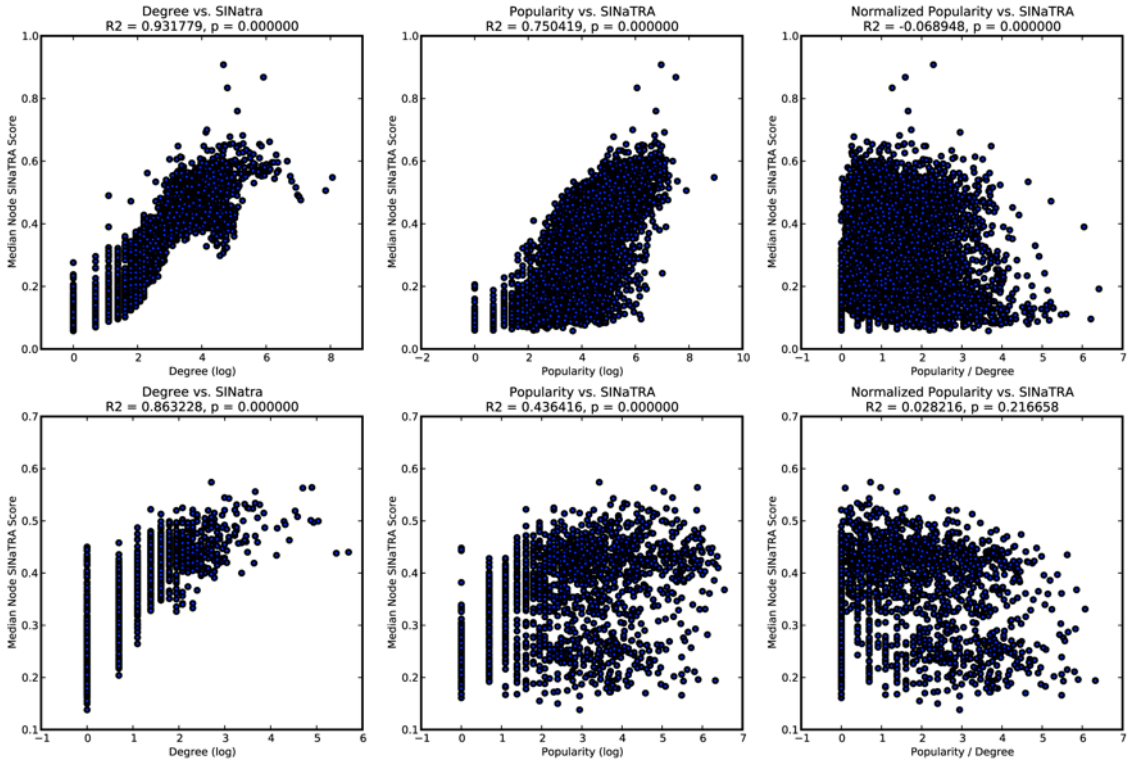


Figure 3.A.7: SINaTRA and node popularity

We plotted the median SINaTRA score of all genes for S. cerevisiae *(top) and* S. pombe *(bottom) vs. node degree (left), node popularity (center; the number of times it appears in the BioGrid database), and normalized popularity (right; popularity/degree). We found that, while SINaTRA score is correlated with the former two measures, it is not correlated with the latter, which gives a better approximation of research bias.*

ACKNOWLEDGEMENTS

This chapter is a reproduction, in part, of a publication in PLOS Computational Biology by Jacunski *et al*. We thank Brent R. Stockwell, Hossein Khiabanian, and Cameron Palmer for their valuable input and editorial suggestions.

CHAPTER 4 – INTERSPECIES MODELS OF SYNTHETIC LETHALITY IN HUMANS

INTRODUCTION

Synthetic lethality (SL) has been suggested as a powerful tool for studying drug action in humans; for example, it can guide the development of cancer combination therapy [67,68] and inform drug-drug interactions. Although SL has been studied extensively in yeast, few genome-wide studies have occurred in humans, and several factors impede a species-wide evaluation of SL. These include the ethical implications of studying SL directly, the inability to discern state-specific SL interactions from global ones in experimental cell lines (*e.g.* cancer), and – most significantly – the heavy experimental burden. Over 200 million assays would be required to determine the SL status of all human gene pairs in just a single cellular context. *In silico* methods are therefore necessary to guide the identification of SL in humans.

In the previous chapter, we showed that we are able to predict synthetic lethality in species with no known synthetic lethal pairs, given a species in which SL is well studied. The only necessary information is experimentally derived protein-protein interaction (PPI) networks for both the source and target species, and the SL data of the source species. Here, we use SINaTRA to predict SL in humans to assign each human gene pair a score between 0 and 1, indicating the likelihood that the two genes exhibit an SL relationship. As a post-processing step to enrich our predictions, we use databases of population genetic variation in humans to filter out likely false positives. Finally, we evaluate of the biomedical implications of our human SL gene pairs by discovering “hot zones” of putative SL pairs that suggest novel cancer combination therapies.

RESULTS

Prediction of synthetic lethality in humans

We applied the SL model trained on *S. cerevisiae* to human network parameters and generated a SINaTRA score between 0 and 1 for all human gene pairs; a higher score indicates greater evidence of SL according to our model. We next compiled a database of severe, tolerated, homozygous, deleterious co-mutations. These occur when at least one patient is homozygous for a deleterious mutation in both genes of a given pair in either of two datasets (1000 Genomes [90,91] and Sweden-Schizophrenia Population-Based Case-Control Exome Sequencing [dbGaP accession: phs000473.v1.p1 [92-94]]). We evaluated all gene pairs and found 450,010 that match these criteria (0.4% of all possible pairs). We found that, on average, the filtered gene pairs had significantly lower SINaTRA scores (median score = 0.116) versus all gene pair scores (median = 0.122; Mann Whitney U = 98,055,441,225.5, $p \leq 2.2e10\text{-}16$). We removed the filtered pairs from our predictions as likely non-SL pairs. Using a SINaTRA cutoff ≥ 0.85, we find the false discovery rate (FDR) from this filtering is 0.36% (61 false positives to 16,886 true positives).

In the interests of space, we provide a filtered list of 1,311 gene pairs with SINaTRA ≥ 0.95 in Table 4.A.4 as an embedded table, and in the supplementary results of Jacunski *et al*. as a CSV file [66], and provide the complete list of 109,358,780 gene pairs and SINaTRA scores as a database download at the Tatonettti laboratory website (URL: http://tatonettilab.org/resources.html).

Putative synthetic lethal pairs are more likely to be in the same pathway

Previous work has shown that SL pairs tend to be part of the same pathway [20,22,30]. We validated this in our predicted human SL pairs using KEGG annotations [95]. We found that gene pairs with SINaTRA scores ≥ 0.95, 0.90, and 0.80 were all significantly enriched for intra-

pathway interactions compared to pairs selected at random ($p<2.2e-16$, Fisher's exact test, all cutoffs). The ten highest-scoring gene pairs with the same pathway annotation are shown in Table 4.1.

Gene 1	Gene 2	SINaTRA Score	Pathway Name
KYNU	SMS	0.99	Tryptophan metabolism
KYNU	GSR	0.987	Tryptophan metabolism
SOS1	BCR	0.986	MAPK signaling pathway
MSH3	PMS2	0.986	Mismatch repair
RCOR1	REST	0.985	Huntington's disease
BIRC5	CASP9	0.985	Pathways in cancer
KYNU	NAGK	0.984	Tryptophan metabolism
POLR1B	POLR1A	0.98	Purine metabolism
RIPK1	RIPK3	0.98	Apoptosis
MAPK9	MAP2K7	0.98	MAPK signaling pathway

Table 4.1: The top ten highest-scoring within-pathway, putative SL gene pairs.

Protein complexes are significantly enriched for putative synthetic lethal pairs

A protein complex may be functional with one deleteriously mutated component, but present with a lethal phenotype when two such mutations occur [20]. Our results corroborate this pattern. We randomly selected 20 sets of mutually exclusive protein complexes with five subunits from the Comprehensive Resources of Mammalian Protein Complexes (CORUM) [96,97] and plotted the SINaTRA scores of all the associated genes as a heat map (Figure 4.1A). We observed that genes with their products in the same protein complex had significantly higher SINaTRA scores ($U = 3{,}425.5$, $p<2.2e-16$; Figure 4.1B). Additionally, within-complex pairs were significantly enriched for higher SINaTRA scores for complexes of size ≤ 10 proteins ($U = 3{,}114{,}511.5$, $p<0.0001$), and complexes of all sizes ($U = 295{,}820{,}010$, $p<0.0001$). Finally, as the size of a complex increases, the distributions of within-complex gene pair SINaTRA scores shifts to a leftward skew, echoing the distribution of gene pairs not in complexes. The proportion of

gene pairs that have products in the same complex is significantly higher than expected by chance ($p<0.0001$, Fisher's exact test, all SINaTRA score cutoffs) (Figure 4.1C).

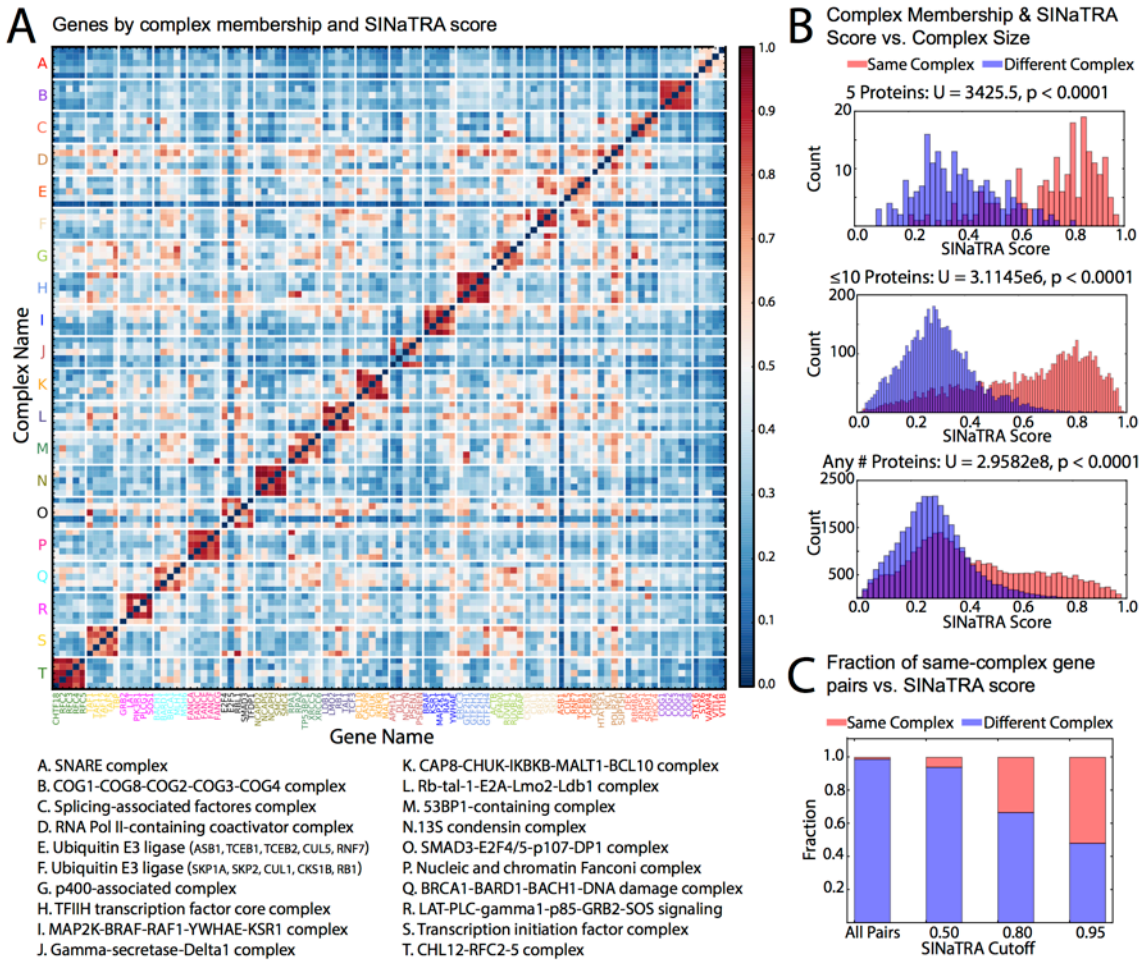


Figure 4.1: Protein complex subunits are more likely to be predicted synthetic lethal
A.) *We randomly selected 20 mutually exclusive groups of protein complexes that contained exactly five subunits; we mapped the corresponding gene pairs to SINaTRA scores, and plotted a heat map of the results. Data are not clustered and only one randomly sampling was performed. We observed that within-complex gene pairs have significantly higher SINaTRA scores ($p<0.0001$, Fisher's exact test).* B.) *We compared the SINaTRA scores of gene pairs with products in the same vs. different complexes for complexes with of 5 protein subunits (top), ≤10 proteins (middle), and any number (bottom). Although proteins in the same complex are always enriched for higher SINaTRA score, as complex size increases, complex membership becomes less indicative of two genes being SL.* C.) *We compared the fraction of gene pairs with products in the same vs. different complexes for three SINaTRA cutoffs (0.95, 0.80, 0.50) as well as for all gene pairs. A SINaTRA cutoff of 0.95 has approximately half of its pairs associated with the same complex; however, a decrease in the cutoff shifts this balance. This may indicate an increase in different mechanisms of SL in pairs with lower scores. "All Pairs" shows the expected proportion of in-complex pairs in our data.*

Prediction of synthetic lethality is not driven by node popularity

As in *S. cerevisiae* and *S. pombe*, we were concerned about research bias, as higher degree nodes are more likely to be studied, and more popularly studied genes may be more likely to have been tested for synthetic lethality. As a measure of this potential bias, we defined a

normalized popularity measure (degree/popularity), where popularity is the number of times a particular gene appears in the BioGrid database. We found that, as in *S. cerevisiae* and *S. pombe*, SINaTRA score is not correlated with normalized popularity in humans (Figure 4.A.1). We found that the predictive performance of SINaTRA is independent of each of the three measures (degree, popularity, and node popularity) according to ANOVA ($p < 0.0001$).

Context-specific synthetic lethality

Synthetic lethality can change between contexts [69]; a gene pair that is SL in a cancer cell may not be in healthy tissue. This may occur due to changes in protein expression, as well as activation or inactivation of protein pathways, which can alter context-specific PPIs [98].

S. cerevisiae and *S. pombe* are unicellular organisms; therefore, models based on these species will necessarily focus on high-level, context-free synthetic lethal predictions. As such, the initial predictions from SINaTRA present all pairs that have synthetic lethal potential in their global connectivity patterns.

In order to explore context-specific SL pairs, we identified all human gene pairs with SINaTRA score ≥0.85. We next created tissue- and cell-line-specific lists of SL pairs by removing a gene pair if that tissue is not known to express both gene products according to the Human Protein Atlas [99,100]. The proportion of SL pairs retained after filtering is illustrated in Figure 4.2A (tissue) and Figure 4.2B (cell); bars are color-coded by biological system. Although the number of proteins removed from the network is correlated with the number of SL pairs filtered from each given tissue or cell line (Figures 4.2C-D), we find that the number of filtered SL pairs is, at times, lower or higher than expected by chance (Table 4.A.1-2) (*Materials and Methods*). For example, rectal tissue has approximately half as many SL pairs filtered out (70) as expected (146; OR = 0.477,p = 1.6e-5, Fisher's exact test). In contrast, tissue of the small intestine has over twice as many SL pairs filtered (1653) than expected (826; OR = 2.11, p<2.2e-16, Fisher's exact test). Respiratory epithelial cells also have a high number of filtered SL pairs (O: 550, E: 280; OR = 2.00,p<2.2e-16).

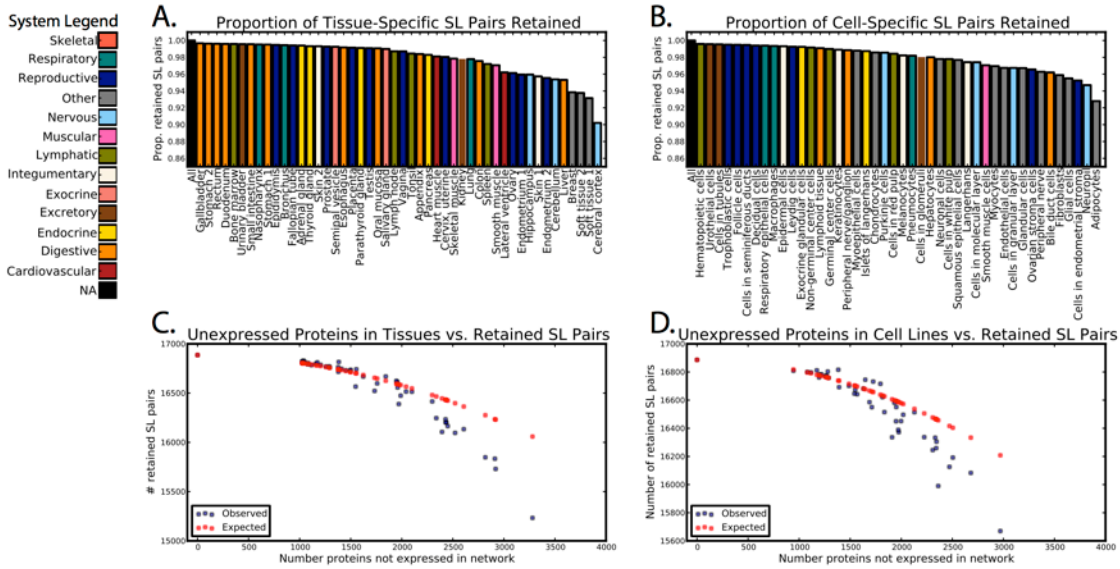


Figure 4.2: Tissue-Specific Synthetic Lethality
We identified all human gene pairs with SINaTRA≥0.85 and all tissue- and cell-line-specific SL pairs by filtering out all gene pairs where neither gene product is expressed in the tissue/cell-line. A.) *The proportion of retained SL pairs by tissue. Tissues are color-coded by the system to which they belong (legend: far left).* B.) *The proportion of retained SL pairs by cell type. Cells were associated with tissue and mapped to system. Cells occurring in multiple tissues from different systems are coded as "other."* C.) *The observed number of retained tissue-specific SL pairs (blue) versus the expected number (red; model described in Materials and Methods).* D.) *The observed (blue) vs. expected (red) number of retained cell-specific SL pairs. The presence of higher- or lower-than-expected numbers of retained SL pairs may indicate context-specific resistance or susceptibility to SL interactions.*

Comparisons with previously published methods

Recent work on human SL includes the Syn-Lethality database [101], which compiles experimentally identified human SL pairs, and the DAISY method [102], a computational method of identifying SL pairs. We found that the gene pairs from both datasets had significantly higher SINaTRA scores (Syn-Lethality: U = 12,265, p<2.2e-16; DAISY (VHL): U = 299, p = 5.86e-6; DAISY (cancer): U = 1992856, p<2.2e-16; Figure 4.3A). Compared to the median of untested pairs (0.122; 99% CI: [0.122,0.122]), DAISY's cancer predictions had a median score of 0.233 (99% CI: [0.225,0.243]); its VHL predictions had a median score of 0.255 (99% CI:[0.195,0.368]) and the Syn-Lethality dataset had a median score of 0.459 (99% CI: [0.397,0.514]).

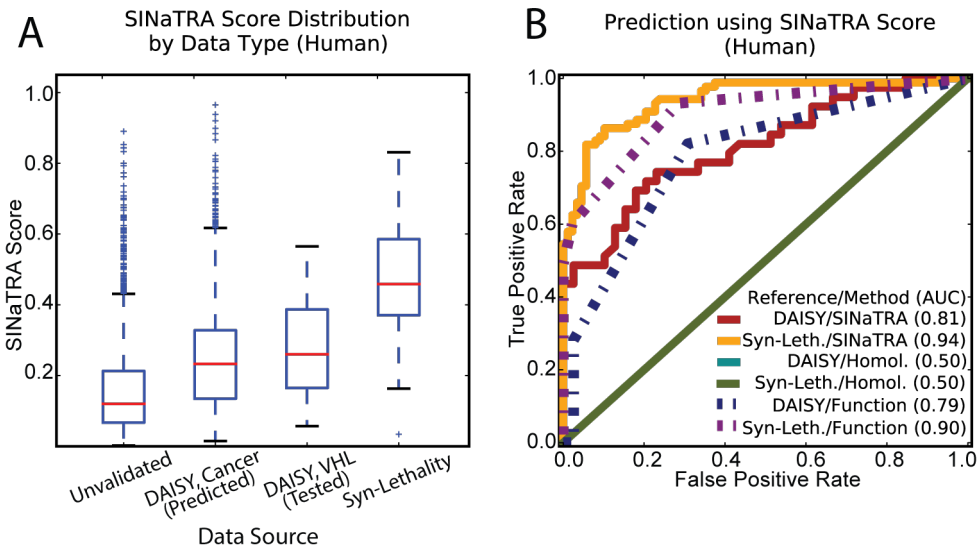


Figure 4.3: SINaTRA versus previously published methods
A.) *SINaTRA scores of all human predictions, as well as pairs predicted or found to be SL in two datasets: DAISY and Syn-Lethality.* B.) *We compare the predictive ability of SINaTRA score to identify genes belonging to DAISY (tested) and Syn-Lethality datasets compared to functional similarity and homology.*

From the Syn-Lethality database, we selected only SL gene pairs involving genetic deficiency, inactivation, or mutation. Of the 88 pairs matching these criteria, all were in our network, and we found 34 of these to have SINaTRA ≥0.5 ($p = 4.8e-11$, Fisher's exact test), and 11 with SINaTRA≥0.75 ($p = 0.0070$, Fisher's exact test). 2,816 gene pairs were predicted to be SL specifically in cancer using DAISY, and 2,576 were in our network; of those, we found that 151 had SINaTRA≥0.5 ($p = 7.5e-24$, Fisher's exact test), and 14 had SINaTRA≥0.75 ($p = 0.00096$, Fisher's exact test).

We observed that SINaTRA score could predict genes in both the DAISY and Syn-Lethality datasets with AUCs of 0.73 and 0.93, respectively. (Figure 4.3B). In turn, homology was not at all predictive in either dataset (AUC = 0.50 for both; no homology data present for the pairs), unlike functional annotations (AUC = 0.786, DAISY; AUC = 0.904, Syn-Lethality). We then considered the precision-recall curves of these data and found that SINaTRA in both datasets outperformed function in DAISY, while function in Syn-Lethality had similar performance to that of SINaTRA (Figure 4.4).

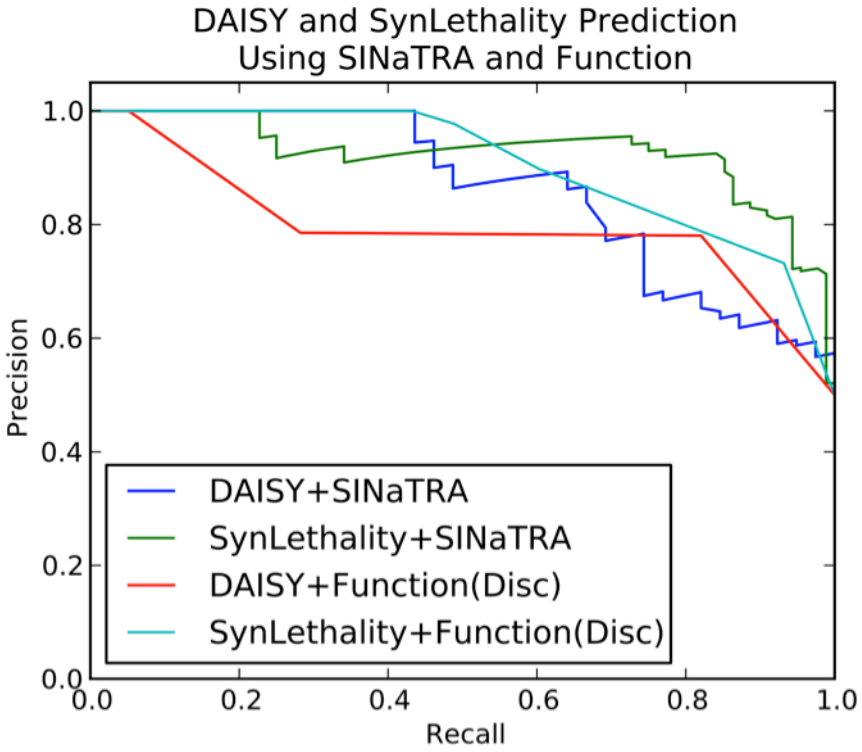


Figure 4.4: Precision-recall of SINaTRA, DAISY, and Syn-Lethality
Precision-recall curves for SINaTRA and functional homology's abilities to predict members of the DAISY and Syn-Lethality studies.

The landscape of human synthetic lethality

We identified 1,311 predicted SL gene pairs with SINaTRA≥0.95. These pairs contained 986 unique genes, of which 402 existed in only one pair (repetition count range: [1,26]; median: 2). From this list, we found 458 gene pairs that were associated with biological pathway data

from Reactome [103] (357 unique genes, of which 167 exist in only one pair; see Table 4.A.3). We present these gene pairs as a network of networks (Figure 4.5). Hexagonal nodes represent pathways, and edges connect pathways when SL pairs are predicted between-pathway (i.e. with one member in each). Within each hexagonal node is a pathway-specific synthetic lethal network, where genes are nodes, and edges appear where the genes have a SINaTRA score ≥ 0.95. We found that 334 (73%) of these interactions are within-pathway and 124 (27%) are between-pathway (OR = 3.69, $p<0.0001$, Fisher Exact Test).

Among the within-pathway SL pairs, we found that apoptosis, the immune system, and gene expression have low closeness centrality in their SL networks, which indicates high interconnectedness. The immune system has the highest number of associated SL gene pairs (101); the most central of these is RIPK1, with 15 connections. Several functions have no associated SL pairs, including extracellular matrix organization, metabolism of proteins, and reproduction. These functions may have little functional redundancy that allows for SL to occur. Of the between-pathway SL pairs, we found that each pair of pathways shares an average of 2.8 SL pairs. The immune system/signal transduction between-pathway pairs are the most numerous (11 pairs).

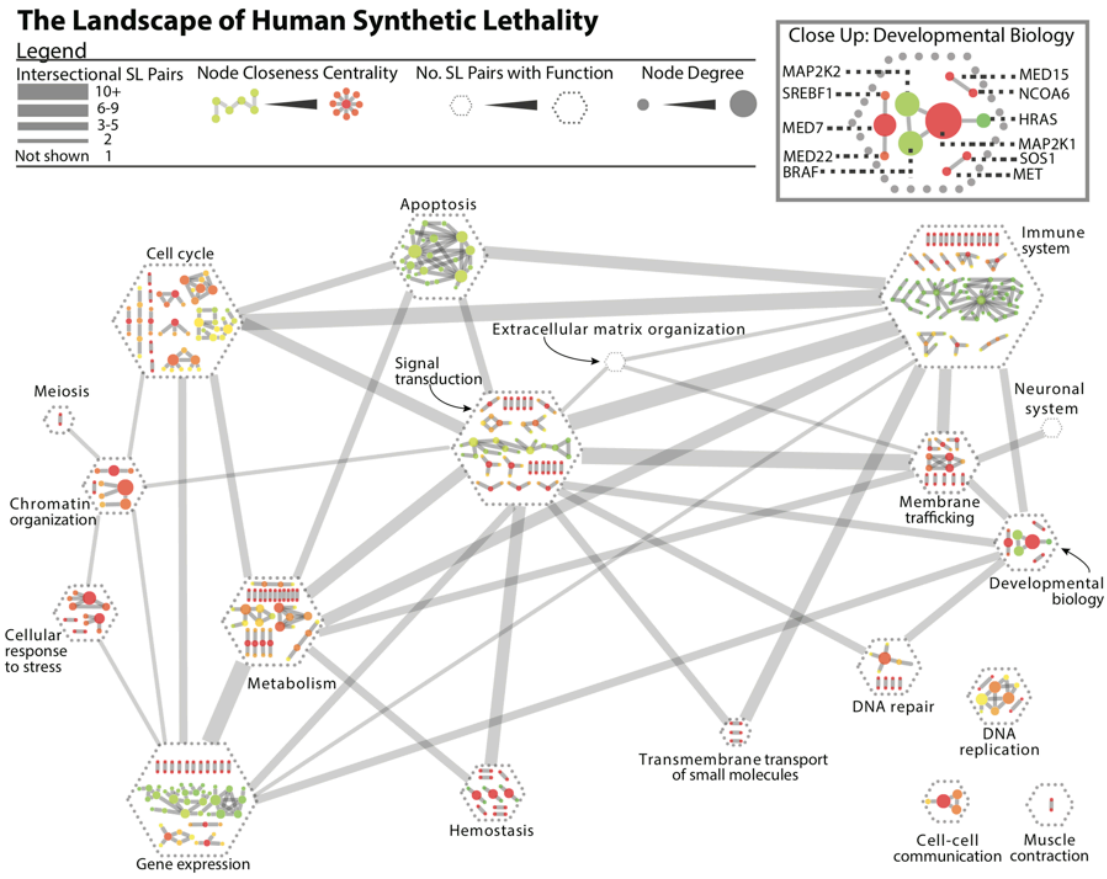


Figure 4.5: The landscape of human synthetic lethality
This network depicts all gene pairs with SINaTRA score ≥0.95 (1,229 SL pairs) that map to Reactome pathways (458 pairs). Here, each hexagon represents one high-level pathway designation in Reactome. Larger nodes indicate more SL pairs with that designation. Within the hexagonal nodes, we show the networks of synthetic lethality where both members have the same function in Reactome. Each node is a gene and an edge represents a predicted SL interaction. Gene nodes are weighted by degree and coloured by closeness centrality. In turn, weighted edges join hexagonal nodes if pathway-divergent pairs exist; that is, one member of the pair is of one pathway while the second member is of the other. Edges are weighted by the number of pathway-divergent gene pairs associated with both pathways.

Function-specific mechanisms of synthetic lethality

We grouped gene pairs into 17 high-level Reactome functional categories and clustered them by their parameter values (*Materials and Methods*). We found pathway-specific parameter enrichment exists in node-pair parameters (inverse shortest path, communicability, shared neighbours, and shared non-neighbours), but not in single-node parameters, as evidenced by the increase in variance of paired parameters versus single-node parameters (Figure 4.6). For example, the signal transduction pathway has higher values for node-pair parameters than other

functions and all SL pairs. In contrast, apoptosis, DNA repair, and DNA replication have node-pair signals that are closer to the mean of all of its within-function pairs than between functions.

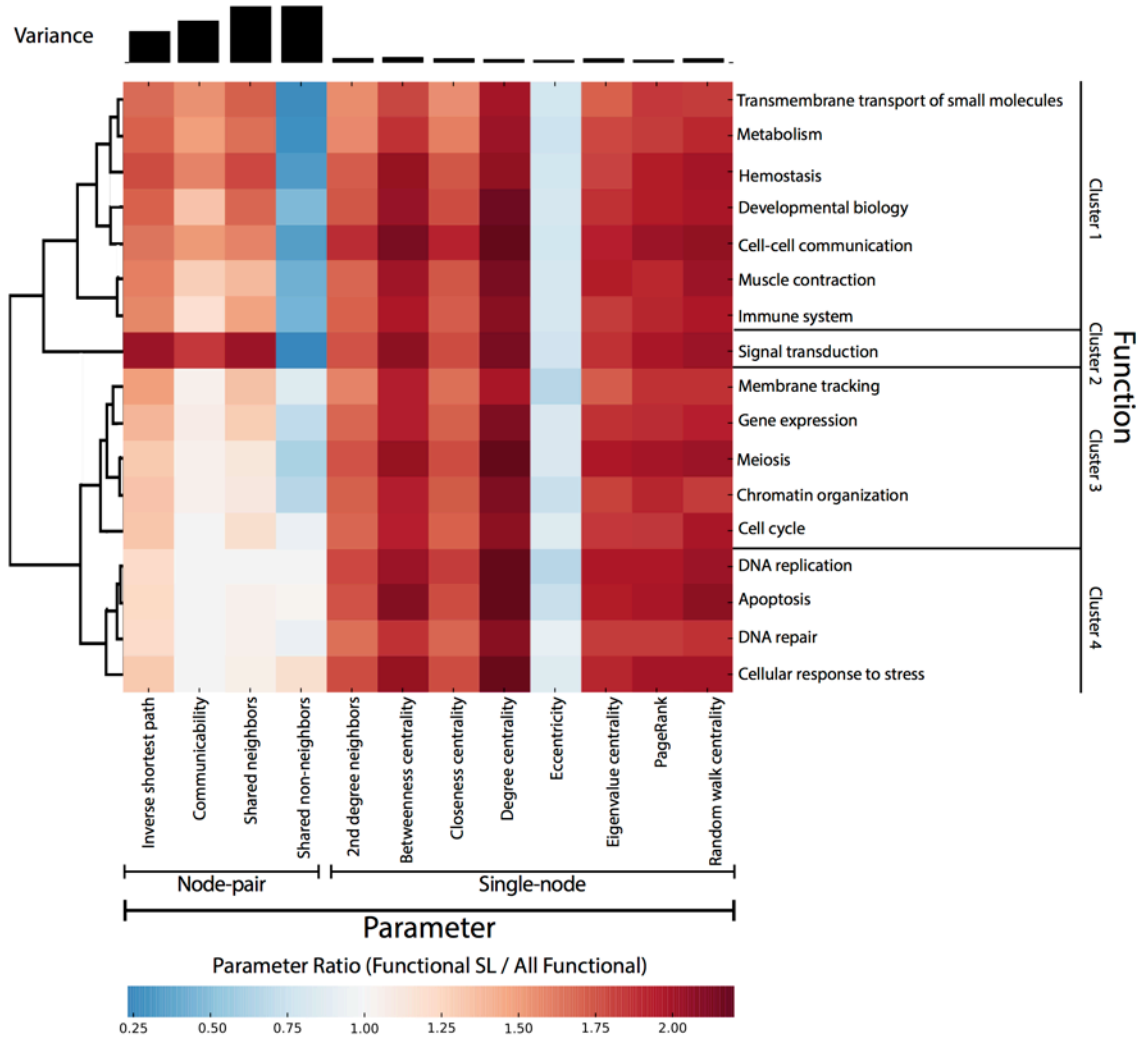


Figure 4.6: SINaTRA and functional signals of synthetic lethality
The heat map represents the ratio of median parameters for the SL pairs of a given function versus all pairs of a given function. For example, the SL pairs of Signal Transduction have values for inverse shortest path that are twice as great as the non-SL pairs of Signal Transduction. Rows are clustered by node-pair parameter values (see Table 1.1). Parameter variance is plotted above the heat map. Single-node parameters (see Table 1.1) are consistently altered in SL regardless of function. However, node-pair parameters differ between functions. This distinction suggests that network substructure may dictate SL mechanisms associated with a specific function.

We then annotated each putative SL gene pair from these 17 functional categories for three possible mechanisms: (1) complex, where the proteins products of the pair are known to form a

complex, (2) parallel, where the proteins function in the same pathway with no known direct or indirect interaction, and (3) other, for gene pairs that do not fit in (1) or (2). In total, there were 5,249 putative SL gene pairs for the 17 categories. Most of these pairs were in the same complex (56.2%, N = 2,950), followed by parallel (24.0%, N = 1,260) and other (19.8%, N = 1,039). We tested each function category for enrichments for particular mechanisms of SL. We found that each function has different proportions of putative mechanistic annotations (Figure 4.7).

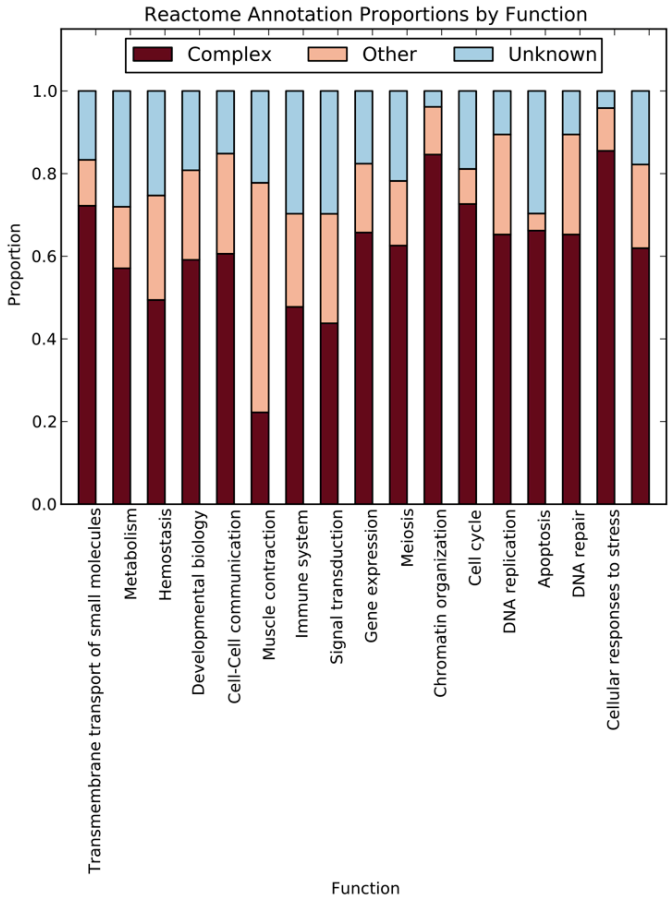


Figure 4.7: Reactome annotation proportions by function
Putative functional SL pairs were annotated using Reactome pathways and grouped into three sets: within-complex interaction, other interaction, and unknown. The fraction of SL pairs in each group is illustrated here by function.

We found that immune system (OR = 1.48, p = 0.000001) and signal transduction (OR = 1.42, p = 0.000894) were significantly enriched for SL genes that function in parallel, after multiple hypothesis correction (Table 4.2). We found four categories were enriched for SL genes

that were components in complexes: gene expression (OR = 1.38, p = 0.000298), meiosis (OR = 4.31, p = 0.046), chromatin organization (OR = 2.10, p = 0.008499), and DNA repair (OR = 4.76, p < 2.2e-16) (Table 4.2). Finally, we found that Cluster 1 (Figure 4.6), which includes transmembrane transport, metabolism, hemostasis, developmental biology, cell-cell communication, muscle contraction, and the immune system, is significantly enriched for SL genes that function in parallel (OR = 1.36, p = 0.00008).

Function	Complex (Count/OR)	Other (Count/OR)	Parallel (Count/OR)	
Transmembrane transport of small molecules	52/2.04**	8/0.5	12/0.63	*Cluster 1*
Metabolism	330/1.04	86/0.68**	162/1.27**	
Hemostasis	86/0.75	44/1.39	44/1.07	
Developmental biology	191/1.13	70/1.13	62/0.74**	
Cell-cell communication	20/1.2	8/1.3	5/0.56	
Muscle contraction	2/0.22**	5/5.08**	2/0.9	
Immune system	606/0.64*	286/1.25**	377/1.48*	
Signal transduction	352/0.55*	213/1.58*	239/1.42*	*Cluster 2*
Membrane trafficking	71/1.51**	18/0.81	19/0.67	*Cluster 3*
Gene expression	572/1.37*	143/0.71**	199/0.86	
Meiosis	22/4.31**	3/0.53	1/0.13**	
Chromatin organization	77/2.1**	9/0.37**	20/0.73	
Cell cycle	124/1.48**	46/1.31	20/0.36*	
DNA replication	96/1.54**	6/0.17*	43/1.35	*Cluster 4*
Apoptosis	124/1.48**	46/1.31	20/0.36*	
DNA repair	124/4.76*	15/0.46	6/0.13*	
Cellular responses to stress	101/1.27	33/1.03	29/0.68	

Table 4.2: Within-function enrichment of putative SL pairs based on gene product interactions
Complex describes all gene pairs that are within the same pathway. Other represents all pairs that have another described PPI. Parallel refers to all pairs with no known PPI between them. Interactions are determined using Reactome data.

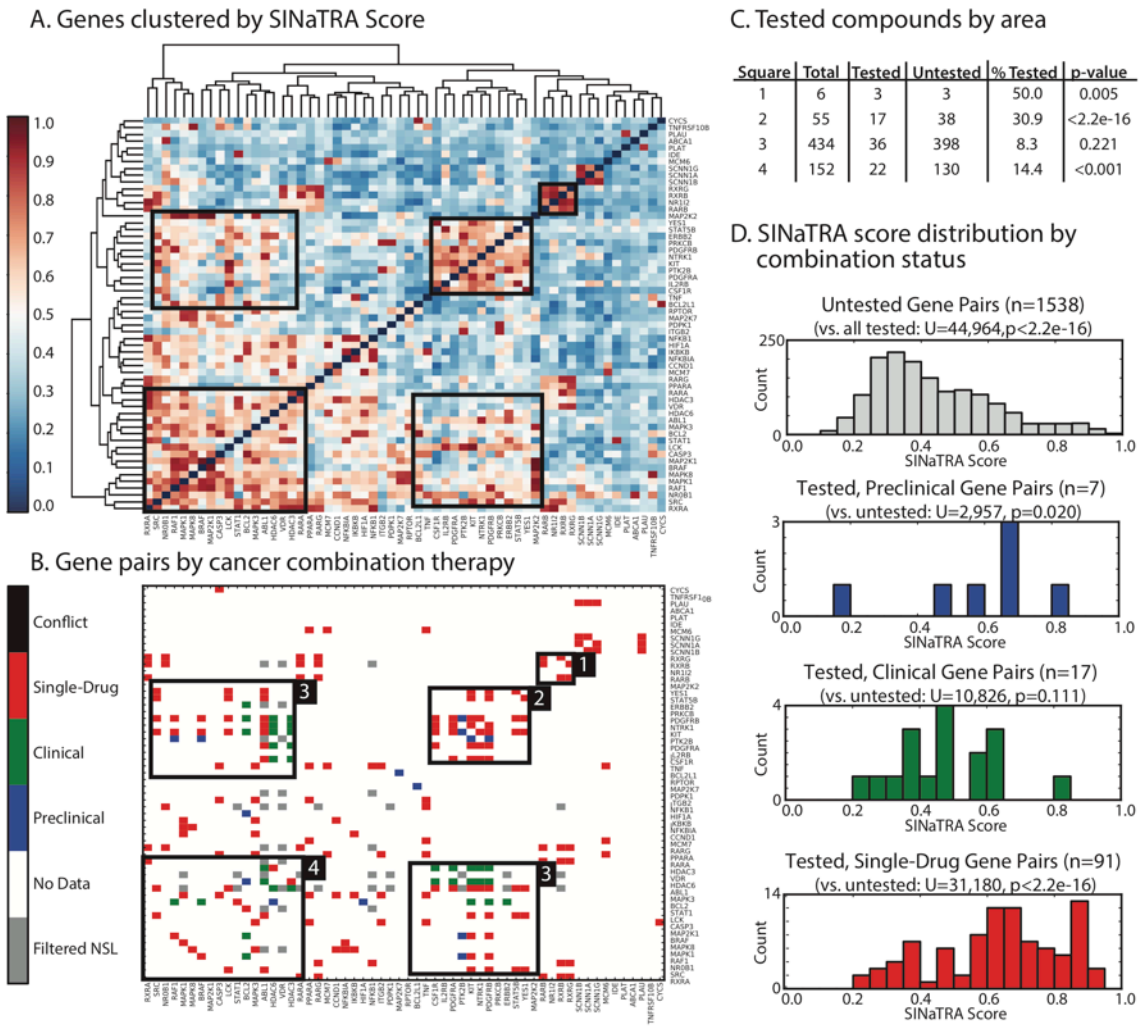
Putative synthetic lethal pairs suggest novel cancer therapies

We identified 58 unique genes from high-scoring gene pairs (SINaTRA≥0.85) where both members were targets of cancer therapies (68 unique drugs). These genes were clustered by SINaTRA score (Figure 4.8A) using hierarchical clustering; areas of high (red) and low (blue) SINaTRA scores are easily observed. We found that gene pairs that are targeted by drugs have

significantly higher SINaTRA scores than those that are not; median SINaTRA score increases significantly from pairs that are targeted by only one drug (median score = 0.156), to those targeted by two drugs (median score = 0.166), to those targeted by only one cancer drug (median score = 0.211), to those targeted by two cancer drugs (median score = 0.283) (Figure 4.A.2).

Next, we identified which of these gene pairs were filtered out through co-mutation analysis (gray), as well as those linked to single-drug therapies (red), drug combination therapies in the clinical pipeline (blue: preclinical; green: in clinical trials). These data were overlaid on the heat map (Figure 4.8B). We found that gene pairs targeted by cancer drugs have significantly higher SINaTRA scores than filtered pairs and pairs not under investigation (Figure 4.8D; $U = 44{,}964$, $p < 0.0001$, Mann-Whitney U test).

We also visually identified "hotspots" of drug combinations (black boxes, Figure 4.8A and 4.8B) that correspond to gene pairs with high SINaTRA scores (Figure 4.8C). We found that Area 1 alone contains genes related to gene expression ($p = 0.040$), transcription initiation from RNA polymerase II promoter ($p = 0.025$), and steroid hormone receptor activity ($p = 0.025$; Fisher's exact test with multiple hypothesis testing). In addition, Area 2 is associated with protein autophosphorylation ($OR = 39.1$, $p = 0.000613$; Fisher's exact test). Areas 3 and 4 are not significantly associated with any GO terms.



Square	Total	Tested	Untested	% Tested	p-value
1	6	3	3	50.0	0.005
2	55	17	38	30.9	<2.2e-16
3	434	36	398	8.3	0.221
4	152	22	130	14.4	<0.001

Figure 4.8: SINaTRA and drug combinations

A.) *Druggable gene pairs clustered by SINaTRA score. Sixty-two unique genes that participated in predicted SL interactions with SINaTRA scores >0.85, where both genes mapped to drugs in DCDB, were identified. All pairwise SINaTRA scores were computed and clustered by score. Areas of high- and low SINaTRA scores are clearly visible.* B.) *All possible gene pairs identified in Part A were mapped to DCDB, and gene pairs whose products are targeted by single drugs and combination therapies in the clinical pipeline were highlighted (pre-clinical, blue; clinical trials, green; single drug, red; gene pairs filtered out by genetic analysis, gray; filtered gene pairs associated with drugs, black [n = 0]). Areas enriched for drug combinations were highlighted in both parts A and B.* C.) *Enrichment of tested compounds in the four areas of interest were calculated using the Fisher Exact Test, and p-values were calculated. Areas 1, 2 and 4 were significantly enriched*. D.) *Distributions of SINaTRA score by drug type.*

DISCUSSION

In this chapter, we expand on a computational method, Species INdependent TRAnslation (SINaTRA), for predicting synthetic lethal (SL) relationships in any species with an available protein-protein interaction (PPI) network. Here, we use SL data from *S. cerevisiae* – the most well characterized organism for this interaction – to predict SL in humans.

Possible mechanisms of synthetic lethality

Several mechanisms of synthetic lethality have previously been proposed [21]; these include within complex, parallel pathways, and essential linear pathways. Connectivity parameters provide hints to the mechanisms driving a particular gene pair to SL. Our data suggest that function-specific network substructures are different, and may be related to trends of SL mechanism within a function. For example, metabolism has a much higher proportion of 'unknown' pathway annotations than does apoptosis (Figure 4.7). This suggests that putative metabolic SL pairs act through parallel pathways, while apoptotic pairs may act through within-complex mechanisms. Further, gene pairs in apoptotic pathways are farther apart and have lower communicability than gene pairs in metabolic pathways, which may also change the proportion of SL pairs that have that functional annotation.

We also observe that a fraction of the predicted SL pairs had between-pathway interactions, where members of an SL pair do not share any single function (Figure 4.5). The respective gene products may act at an interface between two related functions; the putative SL pair may be a false positive; or – most interestingly – one (or both) genes have previously unidentified functions that cause their SL behavior. One such example is the putative SL pair, BAIAP2 (insulin receptor signaling; UniProt DB) and ALDH7A1 (protection from oxidative stress; UniProt DB) (SINaTRA score: 0.957). Oxidative stress is associated with insulin resistance

[104], and knocking out both of these genes may mimic or exacerbate insulin resistance, leading to complications and adverse events.

Context-specific synthetic lethality

Biological contexts, such as tissue type and disease state, can influence synthetic lethal interactions [69]. In translating SL between species, certain factors must be kept in mind; for example, *S. cerevisiae* is a unicellular organism, whereas humans are not. Thus, a gene pair that is SL in one human cell type may not be SL in another. Although this can provide a tremendous therapeutic boon when two drugs targeting two gene products mimic an SL interaction in cancer cells and leave healthy cells unaffected, it also complicates using SL patterns between species of varying complexity.

At this time, cellular and tissue specificity are not captured by the SINaTRA model. However, we can customize our predictions for a given cell or tissue by pruning away any predicted genes that are known not to be expressed in the given context. We used the Protein Atlas [99] to perform this customization and found that certain tissues and cell types had significantly more or fewer SL pairs removed using this method. These deviations may suggest tissue or cell types that are particularly robust, or susceptible, to SL interactions. For example, respiratory epithelial cells and endothelial cells have many more SL pairs filtered out than expected by chance; this suggests that the tissues are not as susceptible to SL reactions. These trends require further investigation, as they may have significant implications for human health.

Predicted synthetic lethal pairs in humans inform cancer polypharmacology

Leveraging synthetic lethal relationships specific to cancer cells has been a strategy in drug discovery for nearly a decade. Therefore, we applied our predictions of synthetic lethality to the study of pharmacology. We found that many cancer combination therapies currently in the clinical pipeline target genes with high SINaTRA scores, suggesting that they use mechanisms of

synthetic lethality as their modes of action. Clustering reveals hotspots of high SINaTRA scores that are significantly enriched for combination therapies under investigation. Importantly, our algorithm was able to identify these without any *a priori* knowledge of the drug combination. Gene pairs found in these hotspots that have not been previously investigated may be promising leads for novel polyphamacological treatments, and we will consider these in the following chapter.

In summary, the methodology presented in this chapter can help to inform a wide variety of studies in human health by utilizing information gathered in model species. In particular, the differential mechanistic analysis that highlights how biological functions may be targeted using synthetic lethality and the "hot spots" of drug synergy highlighted by our cancer therapy analysis indicate promising areas for novel therapeutics.

METHODS

Prediction of synthetic lethality in humans

After establishing the success of parameter translation, we applied the rank-normalized inter-species classifier to human gene pairs.

In order to filter human predictions for false positives, we obtained the VCF files from two studies and annotated them for patients homozygous for significantly deleterious mutations (high impact, resulting in nonsense mutation, early stop, or loss of start). We then identified gene pairs where both genes were simultaneously significantly deleteriously mutated in at least 1 patient but no more than 5% of patients in one study, and filtered these out as confirmed non-SL pairs (N = 405,010).

We compared the SINaTRA scores of the 'confirmed non-SL' pairs to all SINaTRA scores by randomly selecting an equal number of the remaining pairs and applying the Mann-Whitney U test.

We chose high-confidence SL predictions to be those which our classifier assigned SINaTRA scores of >0.95 that were not filtered out by our genetic screen.

Putative synthetic lethal pairs are more likely to be in the same pathway

We identified all putative SL pairs with SINaTRA scores >0.95, 0.90, and 0.80; these groups consisted of 1,224, 6,366, and 32,290 gene pairs, respectively. For all cut-offs, we mapped the genes to their respective pathways using the KEGG database. We compared the number of putative SL gene pairs with the same pathway to the number expected in a group of that size at random. Significance was assessed using the Fisher exact test.

Protein complexes are significantly enriched for putative synthetic lethal pairs

We identified all complexes from the CORUM mammalian protein complex database where all members of the complex mapped unambiguously to one Entrez gene ID. We then randomly selected 20 mutually exclusive complexes composed of five proteins each, and identified the SINaTRA scores for all pairwise combinations of the genes associated with these products. We plotted the SINaTRA scores as a heat map. To test significance, we randomly selected the same number of inter-complex gene pairs as there were intra-complex gene pairs, and applied the Mann-Whitney U test.

We additionally investigated whether this trend of significance would hold for all protein complexes that were composed of ≤10 proteins from our filtered list, and for all protein complexes in our filtered list. Significance was tested using the same methodology and the Mann-Whitney U test.

Prediction of synthetic lethality is not driven by node popularity

As with *S. cerevisiae* and *S. pombe* in the previous chapter, we plotted the median SINaTRA score of genes in humans versus the node's degree, popularity (the number of times it appeared in the BioGRID database), and normalized popularity ($\frac{\text{degree}}{\text{popularity}}$). We calculated the Spearman correlation coefficient for all plots.

Context-specific synthetic lethality

Protein expression data in tissues was downloaded from the Protein Atlas. ENS identification codes were mapped to Entrez gene IDs, and putative SL pairs at each SINaTRA cutoff were determined to be non-SL in context if both proteins were not detected in the tissue of choice. We identified all gene pairs with SINaTRA≥0.85. For each tissue and cell line, we removed a gene pair from the context-specific SL pair list if both genes' products were found not to be expressed in the given context. The SL pairs that were not filtered out by this method were

considered the retained SL pairs. We calculated the number of expected retained gene pairs as follows:

$$\left(1 - \frac{\#removed\ pairs}{total\ human\ pairs}\right) * N$$

where N is the total, unfiltered number of gene pairs that are SL at the chosen cutoff.

Comparisons with previously published methods

SL predictions from the Syn-Lethality and DAISY papers were mapped to their Entrez gene terms, and we found the SINaTRA score of each pair. Significance compared to random SINaTRA pairs was evaluated using the Mann-Whitney U test. We constructed classifiers for DAISY and Syn-Lethality using SINaTRA scores as the features and status in the given dataset as the class. We compared this with homology and functional similarity (GO).

We next tested the ability of three methods (SINaTRA, functional similarity, homology) to predict membership in the DAISY and Syn-Lethality datasets. We used only pairs from the tested VHL predictions from DAISY. We selected an equal number of gene pairs belonging in the dataset (positive examples) and not in the dataset (negative examples), and identified the SINaTRA scores, homology-based SL status from *S. cerevisiae*, and discrete within-species functional similarity score for each. These scores were used in calculation of the ROC curve and precision-recall curves.

The landscape of human synthetic lethality

In order to graphically explore the landscape of human synthetic lethality, we identified all gene pairs with SINaTRA scores ≥0.95. These were mapped to the Reactome database, using the highest terms in the hierarchy: apoptosis; binding and uptake of ligands by scavenger receptors; cell cycle; cell-cell communication; cellular response to stress; chromatin organization; circadian clock; developmental biology; disease; DNA repair; DNA replication; extracellular matrix

organization; gene expression; hemostasis; membrane trafficking; metabolism; metabolism of proteins; muscle contraction; neuronal system; organelle biogenesis and maintenance; reproduction; signal transduction; and transmembrane transport of small molecules. Of the 1,229 gene pairs with SINaTRA scores ≥0.95, there were 458 with both members mapped to a Reactome label.

SL pairs were represented in pathway-specific networks visualized in Cytoscape [61], where both genes were part of the same pathway. Genes are nodes, and two nodes are connected if their SINaTRA score is ≥0.95. Nodes are coloured by closeness centrality, and their size depends on node degree. Pathway-specific networks are designated by hexagons, which are joined to each other with edges weighted by the number of inter-pathway SL pairs that exist; that is, gene pairs with mutually exclusive pathway designations.

Function-specific mechanisms of synthetic lethality

We identified all gene pairs of the functions from the previous section, as well as an SL subset (SINaTRA score ≥0.85). We then found the median value of all node-pair and single-node parameters and plotted a heat map of the ratio of SL to all gene parameters. Because of the low variance between single-node parameters, we clustered each function by the node-pair parameters.

We next annotated all SL pairs with Reactome pathways into three groups: complex, parallel, and other. Two genes were annotated with “complex” if their protein products were known to participate in a protein complex together. Two genes were annotated with “parallel” if they had the same functional annotation but no direct interaction according to Reactome. Finally, two genes were annotated as other if they did not fit these either the “complex” or “parallel” definitions. For each functional category we tested if the gene pairs were enriched for parallel or complex annotations using a Fisher’s exact test.

Mapping drugs to gene product targets

We first mapped all gene pairs with SL score > 0.85 to drugs in the Drug Combination Database (DCDB) [105], such that both genes in a pair mapped to a cancer drug that targeted their products. Cancer drugs were identified from DCDB as those with indications containing the terms *cancer*, *leukemia*, *carcinoma*, *myeloma*, *tumour*, *sarcoma*, *lymphoma*, or *neoplasm*. From these gene pairs, we identified all unique genes among the pairs. We found a list of 52 unique genes from a list of 381 pairs.

Putative human synthetic lethal pairs are predictive of investigative cancer therapy

Using the aforementioned list of genes, we identified the SINaTRA score for all pairwise combinations of genes. We plotted these as a heat map, clustering the rows and columns by SINaTRA score. We then found all known single-drug and cancer combination therapies in experimental and clinical pipelines using DCDB, and overlaid these data on the clustered heat map to visually identify clusters of therapies and their correspondence to SINaTRA score. We additionally inspected all pairs of genes that were filtered out using our co-mutation analysis, and confirmed that none of them were also targets of cancer drugs. We performed a Mann-Whitney U test on distributions of SINaTRA scores for non-tested and filtered gene pairs vs. gene pairs associated with drugs, vs. single-drug gene pairs, vs. drug combinations in preclinical testing, and vs. drug combinations in clinical testing.

In order to identify GO enrichment, we tested the GO terms of within-box genes compared to all remaining genes from Figure 4.8. Statistical testing was performed using Fisher's exact test.

Statistical analyses and software

We calculated network parameters using the NetworkX version 1.8.1. We performed statistical analysis in R version 3.0.2. De Long's test for comparing ROC curves was implemented using the pROC library [88]. Scripts use Python version 2.7.5. Graphics were generated using Python's Matplotlib [89]. BioGrid [54] release 3.2.104 was used in all analyses.

APPENDIX

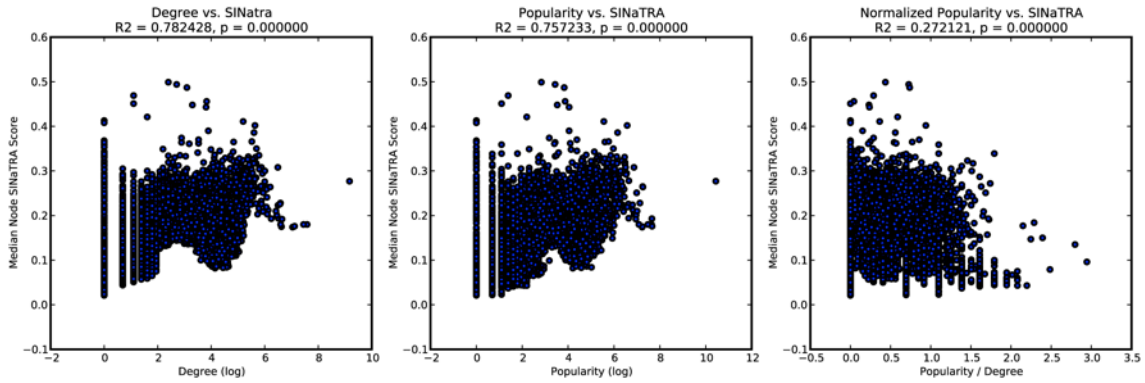


Figure 4.A.1: SINaTRA and node popularity
We plotted the median SINaTRA score of all human genes vs. node degree (left), node popularity (center; the number of times it appears in the BioGrid database), and normalized popularity (right; popularity/degree). We found that, while SINaTRA score is correlated with the former two measures, it is not correlated with the latter, which gives a better approximation of research bias.

Tissue	Removed	Remaining	Expected Removed	Expected Remaining	OR	p-value
All	0	16886	0	16886	-	-
Gallbladder	102	16784	95	16790	1.074068	0.668224
Stomach 2	272	16614	294	16591	0.923889	0.351466
Rectum	70	16816	146	16739	0.477257	0
Duodenum	1038	15848	610	16275	1.747487	0
Bone marrow	93	16793	89	16796	1.04513	0.823629
Urinary bladder	780	16106	440	16445	1.81004	0
Small intestine	1653	15233	826	16059	2.109725	0
Nasopharynx	330	16556	294	16591	1.124822	0.157236
Stomach 1	412	16474	304	16581	1.364066	0.000051
Epididymis	69	16817	81	16804	0.851193	0.327724
Bronchus	684	16202	456	16429	1.521016	0
Fallopian tube	752	16134	522	16363	1.461061	0
Adrenal gland	88	16798	103	16782	0.853555	0.277611
Thyroid gland	139	16747	152	16733	0.913709	0.444875
Skin 2	99	16787	111	16774	0.891201	0.407329
Prostate	56	16830	82	16803	0.681831	0.02676
Seminal vesicle	320	16566	184	16701	1.753303	0
Esophagus	684	16202	453	16432	1.531368	0
Placenta	371	16515	320	16565	1.162885	0.054549
Parathyroid gland	640	16246	420	16465	1.544351	0
Testis	790	16096	488	16397	1.649125	0
Oral mucosa	372	16514	338	16547	1.102791	0.210652
Salivary gland	217	16669	261	16624	0.829173	0.042776
Lymph node	80	16806	89	16796	0.898342	0.489041
Vagina	157	16729	171	16714	0.917305	0.438012
Tonsil	652	16234	453	16432	1.456848	0
Appendix	288	16598	237	16648	1.218851	0.027761
Pancreas	150	16736	172	16713	0.870895	0.218634
Heart muscle	144	16742	185	16700	0.776426	0.023182
Cervix, uterine	123	16763	145	16740	0.847112	0.177957
Skeletal muscle	66	16820	84	16801	0.784827	0.141824
Kidney	170	16716	146	16739	1.165986	0.19354
Lung	126	16760	128	16757	0.984199	0.899985
Colon	364	16522	231	16654	1.588347	0
Spleen	721	16165	461	16424	1.58905	0
Smooth muscle	116	16770	121	16764	0.958335	0.745068
Lateral ventricle	75	16811	81	16804	0.92554	0.631389
Ovary	497	16389	298	16587	1.687934	0
Endometrium 1	1156	15730	654	16231	1.823882	0
Hippocampus	1052	15834	651	16234	1.656798	0
Skin 1	471	16415	405	16480	1.167568	0.026015
Endometrium 2	82	16804	81	16804	1.012346	1
Cerebellum	63	16823	80	16805	0.786657	0.155456
Liver	154	16732	161	16724	0.956064	0.692438
Breast	115	16771	127	16758	0.90481	0.439831
Soft tissue 2	262	16624	291	16594	0.898719	0.214081
Soft tissue 1	74	16812	105	16780	0.70342	0.020262
Cerebral cortex	218	16668	202	16683	1.080179	0.461455

Table 4.A.1: Tissue-specific synthetic lethality

The number of edges removed in each tissue-specific context compared to the expected number removed. OR and p-values are calculated using Fisher's exact test.

Cell	Removed	Remaining	Expected Removed	Expected Remaining	OR	p-value
All	0	16886	0	16886	-	-
Hematopoietic cells	435	16451	306	16579	1.432629	0.000002
Urothelial cells	125	16761	109	16776	1.147815	0.325124
Cells in tubules	244	16642	188	16697	1.302162	0.007668
Trophoblastic cells	1216	15670	677	16208	1.857827	0
Follicle cells	641	16245	409	16476	1.589523	0
Cells in seminiferous ducts	803	16083	552	16333	1.477323	0
Decidual cells	335	16551	292	16593	1.150172	0.090358
Respiratory epithelial cells	550	16336	280	16605	1.996631	0
Macrophages	436	16450	291	16594	1.511397	0
Epidermal cells	265	16621	253	16632	1.048124	0.626256
Leydig cells	89	16797	95	16790	0.936452	0.658358
Exocrine glandular cells	77	16809	68	16817	1.132892	0.50568
Non-germinal center cells	372	16514	348	16537	1.070454	0.386259
Lymphoid tissue	240	16646	182	16703	1.323197	0.005176
Germinal center cells	102	16784	114	16771	0.894044	0.41385
Keratinocytes	549	16337	380	16505	1.459594	0
Peripheral nerve/ganglion	116	16770	121	16764	0.958335	0.745068
Myoepithelial cells	128	16758	127	16758	1.007874	1
Islets of langerhans	694	16192	482	16403	1.458597	0
Chondrocytes	86	16800	89	16796	0.966062	0.820717
Purkinje cells	170	16716	247	16638	0.685048	0.000145
Cells in red pulp	553	16333	417	16468	1.3371	0.000011
Melanocytes	760	16126	469	16416	1.649611	0
Pneumocytes	70	16816	146	16739	0.477257	0
Cells in glomeruli	336	16550	225	16660	1.503259	0.000003
Hepatocytes	213	16673	184	16701	1.159553	0.157391
Neuronal cells	185	16701	170	16715	1.089148	0.455122
Cells in white pulp	154	16732	228	16657	0.672411	0.000137
Squamous epithelial cells	110	16776	123	16762	0.893563	0.393819
Langerhans	300	16586	219	16666	1.37647	0.000392
Cells in molecular layer	514	16372	299	16586	1.741534	0
Smooth muscle cells	204	16682	204	16681	0.99994	1
Myocytes	371	16515	258	16627	1.447736	0.000006
Endothelial cells	896	15990	428	16457	2.154599	0
Cells in granular layer	141	16745	208	16677	0.675132	0.000306
Glandular cells	581	16305	422	16463	1.390119	0
Ovarian stroma cells	627	16259	421	16464	1.508089	0
Peripheral nerve	194	16692	148	16737	1.314345	0.014339
Bile duct cells	305	16581	287	16598	1.063807	0.480902
Fibroblasts	103	16783	110	16775	0.935917	0.631306
Glial cells	497	16389	298	16587	1.687934	0
Cells in endometrial stroma	390	16496	313	16572	1.251747	0.003742
Neuropil	83	16803	125	16760	0.662301	0.003467
Adipocytes	74	16812	105	16780	0.70342	0.020262

Table 4.A.2: Cell-specific synthetic lethality

The number of edges removed in each cell-specific context compared to the expected number removed. OR and p-values are calculated using Fisher's exact test.

ABCD1	CDC45	GTF2F1	MED22	POLR1A	SYK
ABCD3	CDC7	GTF2F2	MED25	POLR1B	TAB2
ABL1	CDCA8	GTF2H1	MED30	POLR1D	TAF1
ACVR1B	CFLAR	GTF3C2	MED7	POLR2E	TAF13
ADSL	CHMP4A	GTF3C3	MET	POLR2F	TAF1D
ADSS	CLINT1	GTF3C4	MIS12	POLR2G	TAF6
ALDH7A1	CLSPN	GTF3C5	MNAT1	POLR2H	TALDO1
ANAPC1	COPE	HARS	MOB1A	POLR3C	TAX1BP1
ANAPC10	CPSF1	HERC2	MTA1	POLR3D	TCEA1
ANAPC11	CRKL	HGS	MTA3	POLR3F	TEC
ANAPC2	CTSA	HIRA	MVD	PPP2CB	TGFBR1
ANAPC4	CXCR4	HRAS	NAPA	PPP2R1B	THRA
ANAPC5	CYLD	IDH1	NCAPD2	PPP2R5A	TICAM1
AP1B1	DAPK1	IL1R1	NCAPG	PPP2R5C	TIRAP
AP2A1	DCK	IL6ST	NCOA1	PPP2R5D	TLR4
APAF1	DCP2	INCENP	NCOA2	PRKCI	TNF
ARCN1	DLAT	ING4	NCOA6	PRKCQ	TNFAIP3
ARHGEF1	DMAP1	ING5	NCOR1	PRKCZ	TNFRSF10B
ARHGEF6	DNMT3A	INSR	NDC80	PSME2	TNFRSF1A
ARHGEF7	DR1	IRAK4	NDUFS2	PTTG1	TOLLIP
ASH2L	DSN1	IRF7	NDUFS3	PXN	TPM2
ASS1	DVL1	IRS1	NDUFS6	RACGAP1	TPM3
ATG12	DVL3	IRS2	NDUFS8	RAD17	TRADD
ATG5	DZIP3	ITK	NDUFV1	RAD9A	TRAF3
AURKB	E2F4	KAT6A	NDUFV2	RAE1	TRIM25
AXIN1	ECHS1	KIF23	NFKB2	RANGAP1	TRIM37
BAIAP2	EDC4	KIF3A	NFYA	RAP1A	TSC22D3
BAK1	EED	KIFAP3	NFYB	RBCK1	UBE2B
BAX	EHMT2	KIT	NME1	RFC1	UBE2C
BCAP31	EPN1	KLC1	NME2	RFC3	UBE2R2
BCAR1	ERBB3	KLC2	NOD1	RHEB	UBE2S
BCL2L1	ERCC1	KLC4	NOD2	RIPK1	UBE2V1
BCR	ERCC2	KMT2D	NOS3	RIPK2	UBE3A
BID	ERCC3	KYNU	NPEPPS	RIPK3	UGDH
BIRC2	ERCC4	LATS1	NR1I2	RPS6KA3	USP8
BIRC3	EXOC1	LATS2	NSL1	RRN3	VAMP2
BIRC5	EXOC4	LCK	NT5C2	SARS	VAMP8
BMPR1A	EXOC8	LEF1	NUBP2	SEC24A	VAPA
BRAF	EXOSC1	LEO1	NUF2	SEC24C	VAPB
BRCA2	EXOSC6	LSM3	NUP98	SEC61A1	VAV2
CABIN1	FADD	LSM5	ORC3	SEC61B	VAV3
CAMK2A	FANCC	LSM6	ORC4	SGK1	VPS25
CARD11	FANCG	MAD2L1	ORC5	SHC1	VPS36
CASC5	FANCL	MALT1	ORC6	SMAD6	VPS4A
CASP10	FAS	MAP2K1	PAK2	SMAD7	WAPAL
CASP2	FBXO5	MAP2K2	PARD3	SMC2	WARS
CASP3	FZR1	MAP2K4	PARD6A	SMS	WAS
CASP7	GATAD2B	MAP2K6	PARD6B	SNAP23	WDR5
CASP8	GCDH	MAP2K7	PARK2	SNF8	WIPF1
CASP9	GNA12	MAP3K14	PDGFRB	SOCS1	WWP1
CAV1	GOLGA2	MAPK10	PDHB	SOCS3	XIAP
CBLB	GORASP1	MAPK8	PDPK1	SOS1	XPC
CCDC101	GOT1	MAPK9	PDS5A	SREBF1	XPO5
CCNA1	GSK3A	MBIP	PFAS	STAG1	YEATS4
CCNB1	GSR	MCM10	PGD	STAG2	ZAP70
CCNT1	GSS	ME1	PHF1	STAM	ZFYVE9
CD44	GTF2A1	MEAF6	PIK3C3	STAM2	ZWINT
CDC16	GTF2B	MED12	PIK3CA	STAT3	
CDC23	GTF2E1	MED15	PIK3R4	STAT5B	
CDC25C	GTF2E2	MED16	PLCG2	STX4	

Table 4.A.3: List of genes in the "Landscape of Synthetic Lethality"

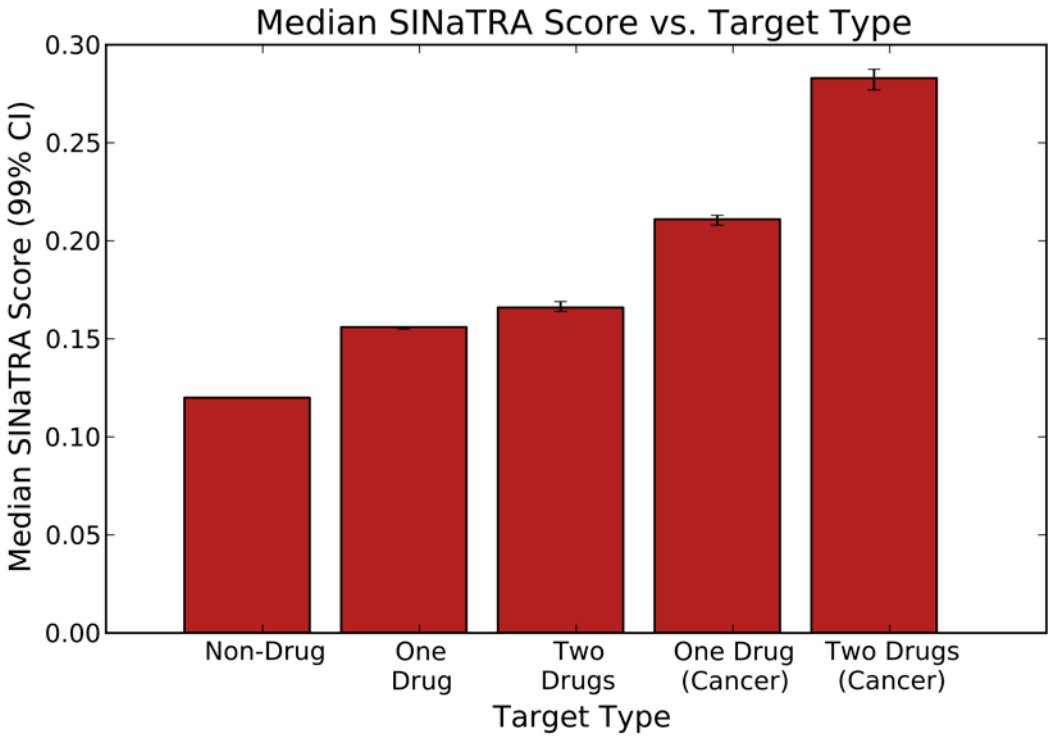


Figure 4.A.2: Median SINaTRA scores of drug targets

We observed that gene pairs targeted by drugs are significantly enriched in SINaTRA score, and the median scores increase from genes that contain only one non-cancer drug target, to those that are affected by two non-cancer drug targets, to those that contain one cancer drug target, to those that contain two. The differences are significant for all comparisons.

Table 4.A.3: List of human gene pairs with SINaTRA ≥ 0.95 (p.107-116)

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
SMS	6611	KYNU	8942	0.990
MAP2K1	5604	BRAP	8315	0.988
PAWR	5074	SNX6	58533	0.988
GSR	2936	KYNU	8942	0.987
MSH3	4437	PMS2	5395	0.986
PPP2R5D	5528	TPD52L2	7165	0.986
PPIB	5479	RBMS1	5937	0.986
BCR	613	SOS1	6654	0.986
USP33	23032	USP28	57646	0.986
RCOR1	23186	REST	5978	0.985
MRPL2	51069	MRPL9	65005	0.985
BIRC5	332	CASP9	842	0.985
MSH3	4437	RAD9A	5883	0.984
NAGK	55577	KYNU	8942	0.984
OPTN	10133	RIPK1	8737	0.983
TWF2	11344	CAPN2	824	0.982
TNIP1	10318	TAX1BP1	8887	0.982
MARK2	2011	PARD6A	50855	0.982
PACSIN3	29763	WIPF1	7456	0.982
PPIB	5479	SUGP1	57794	0.982
VAPB	9217	SEC22B	9554	0.982
C11orf58	10944	KYNU	8942	0.982
RTF1	23168	WDR61	80349	0.982
GSE1	23199	MTA3	57504	0.981
SFXN3	81855	SFXN1	94081	0.981
KMT2A	4297	AFF1	4299	0.981
POLR3C	10623	GTF3C5	9328	0.981
GATAD2B	57459	MTA3	57504	0.981
SNX6	58533	SHMT1	6470	0.981
BIRC5	332	CDCA8	55143	0.981
MAPK9	5601	MAP2K7	5609	0.980
RAD17	5884	RFC3	5983	0.980
HRAS	3265	BRAP	8315	0.980
TPD52	7163	TPD52L2	7165	0.980
NSF	4905	STX7	8417	0.980
TOPBP1	11073	ATRIP	84126	0.980
POLR1A	25885	POLR1B	84172	0.980
RIPK3	11035	RIPK1	8737	0.980
GSPT1	2935	RDX	5962	0.980
CASP7	840	CASP10	843	0.979
CBX1	10951	DNMT3A	1788	0.979
TIMM44	10469	PPIB	5479	0.979
TNK2	10188	BCAR1	9564	0.979
HEXIM1	10614	AFF1	4299	0.979
ATG7	10533	SNX6	58533	0.979
APAF1	317	CASP9	842	0.979
C11orf58	10944	SMS	6611	0.979
FAS	355	RIPK1	8737	0.979
JUNB	3726	FOSL1	8061	0.979
MALT1	10892	USP2	9099	0.979
PACSIN2	11252	WIPF1	7456	0.979
DDOST	1650	AUP1	550	0.979
RBMS1	5937	SCP2	6342	0.979
STX7	8417	SCO2	9997	0.979
RBMS1	5937	STX7	8417	0.979
HEXIM1	10614	BRD4	23476	0.979
TOR1AIP1	26092	RIF1	55183	0.979
GDI2	2665	RAB1A	5861	0.978
GABPA	2551	SP3	6670	0.978
CASP10	843	RIPK1	8737	0.978
DNM1L	10059	CYHR1	50626	0.978
UBQLN2	29978	PFDN2	5202	0.978
SEL1L	6400	DERL1	79139	0.978
RBCK1	10616	NOD2	64127	0.977
ARPC1B	10095	PPP2R5D	5528	0.977
PPP1R2	5504	NAE1	8883	0.977
MALT1	10892	PRKCQ	5588	0.977
ZMYND8	23613	INTS1	26173	0.977
POLR3F	10621	POLR3C	10623	0.977
POLR1A	25885	POLR2H	5437	0.977
VAV2	7410	CD44	960	0.977
USP21	27005	USP2	9099	0.977

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
MKRN3	7681	RNF7	9616	0.977
E2F6	1876	RYBP	23429	0.976
MET	4233	SOS1	6654	0.976
CARD11	84433	USP2	9099	0.976
BAIAP2	10458	CDC25C	995	0.976
UBE2J1	51465	AUP1	550	0.976
NME1	4830	NME2	4831	0.976
ADSS	159	DCK	1633	0.976
USP8	9101	KIF23	9493	0.976
PPIB	5479	ZC3H11A	9877	0.976
MRPL9	65005	MRPL44	65080	0.976
TNF	7124	TNFRSF1B	7133	0.976
DLAT	1737	PDHB	5162	0.976
TRIP4	9325	MED13	9969	0.976
TAF1	6872	TAF7	6879	0.976
STX4	6810	SNAP23	8773	0.976
PPP2R1B	5519	PPP2R5D	5528	0.976
MED28	80306	MED7	9443	0.976
TICAM1	148022	TRAF5	7188	0.975
HK1	3098	CAPN2	824	0.975
MAP2K1	5604	WNK1	65125	0.975
DNMT3A	1788	POLR2H	5437	0.975
APEH	327	SNX6	58533	0.975
INSR	3643	SOCS3	9021	0.975
WARS	7453	XPNPEP1	7511	0.975
TPRKB	51002	OSGEP	55644	0.975
UBE2V2	7336	TRIM5	85363	0.975
ADSS	159	PTMS	5763	0.975
IKZF1	10320	GATA1	2623	0.975
MAP2K1	5604	LAMTOR3	8649	0.975
MARK3	4140	CDC25C	995	0.975
TPD52L2	7165	AARSD1	80755	0.975
ERLIN2	11160	FLOT2	2319	0.975
ATG4B	23192	ULK1	8408	0.975
STX7	8417	VAMP8	8673	0.975
DNM1L	10059	PSMG3	84262	0.975
GATA1	2623	SMARCD1	6602	0.974
RIPK1	8737	USP2	9099	0.974
RBCK1	10616	RIPK1	8737	0.974
PIK3R4	30849	PIK3C3	5289	0.974
LCK	3932	BCAR1	9564	0.974
BAX	581	CASP9	842	0.974
TOPBP1	11073	PMS2	5395	0.974
BAK1	578	BAX	581	0.974
DNM1L	10059	APEH	327	0.974
GSR	2936	PEPD	5184	0.974
DNMT3A	1788	EHMT1	79813	0.974
ORC4	5000	MCM10	55388	0.974
BRD4	23476	RFC3	5983	0.974
INTS6	26512	SEM1	7979	0.974
STX4	6810	VAMP8	8673	0.974
PEPD	5184	KYNU	8942	0.973
UBE2V2	7336	MKRN3	7681	0.973
GPS2	2874	THRA	7067	0.973
OPTN	10133	TNIP1	10318	0.973
ORC5	5001	CDC45	8318	0.973
RCOR1	23186	SNAI1	6615	0.973
XPNPEP1	7511	API5	8539	0.973
PDPK1	5170	PRKCQ	5588	0.973
RBCK1	10616	USP21	27005	0.973
RNF31	55072	NOD2	64127	0.973
RBMS1	5937	RRBP1	6238	0.973
CEBPG	1054	JUNB	3726	0.973
TRIP4	9325	MED23	9439	0.973
TAF7	6879	SETD7	80854	0.973
CABIN1	23523	HIRA	7290	0.973
OSGEP	55644	ZPR1	8882	0.973
IDH1	3417	PTMS	5763	0.973
SH3GLB2	56904	SH3GL1	6455	0.973
EHMT2	10919	DNMT3A	1788	0.973
DNM1L	10059	TBC1D15	64786	0.973
CSF1R	1436	SOS1	6654	0.973

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
MCM10	55388	CDC45	8318	0.972
NSF	4905	NAPA	8775	0.972
VPS36	51028	VPS25	84313	0.972
ARAF	369	MAP2K1	5604	0.972
ARHGEF7	8874	ARHGEF6	9459	0.972
XPNPEP1	7511	GTF3C4	9329	0.972
TTC9C	283237	VPS4B	9525	0.972
PAWR	5074	CARS	833	0.972
TES	26136	OSGEP	55644	0.972
RTF1	23168	TCEA1	6917	0.972
BIRC5	332	INCENP	3619	0.972
ECHS1	1892	GSS	2937	0.972
ORC3	23595	CDC45	8318	0.972
NFYA	4800	NFYB	4801	0.972
PPP2R5A	5525	PPP2R5C	5527	0.972
TNIP1	10318	TNF	7124	0.972
RBMS1	5937	ZC3H11A	9877	0.972
TAF6	6878	SETD7	80854	0.972
TAF1	6872	TAF13	6884	0.972
STX7	8417	SNX3	8724	0.972
SCFD1	23256	SNAP23	8773	0.972
AP2A1	160	EPN1	29924	0.972
STAG1	10274	WAPL	23063	0.972
NDC80	10403	ZWINT	11130	0.972
PEPD	5184	CTSA	5476	0.972
DCK	1633	THOP1	7064	0.972
WDR5	11091	KAT6A	7994	0.972
PACSIN2	11252	UGP2	7360	0.971
RAB1A	5861	RAB11B	9230	0.971
INSR	3643	SOCS1	8651	0.971
DCK	1633	PTMS	5763	0.971
BID	637	CASP2	835	0.971
ARAP1	116985	INPP5D	3635	0.971
SIAH2	6478	SKI	6497	0.971
INPP5D	3635	BCR	613	0.971
GORASP2	26003	TBCD	6904	0.971
SEC61B	10952	ASNA1	439	0.971
MFAP1	4236	SNIP1	79753	0.971
VAPB	9217	VAPA	9218	0.971
ATG12	9140	ATG5	9474	0.971
ARPC3	10094	CALU	813	0.971
ERP44	23071	PAK2	5062	0.971
OSGEP	55644	LAGE3	8270	0.971
CPSF1	29894	GTF3C3	9330	0.971
PEPD	5184	CASP7	840	0.971
TBCB	1155	ERP44	23071	0.971
LEO1	123169	DNMT3A	1788	0.971
TRAIP	10293	TNFRSF1B	7133	0.971
TRAF5	7188	RIPK1	8737	0.971
UBE2J1	51465	DERL1	79139	0.971
CBX1	10951	CHD1L	9557	0.971
MNAT1	4331	USP2	9099	0.971
TNK2	10188	AMPH	273	0.971
BACH1	571	BRCA2	675	0.971
APC2	10297	ANAPC10	10393	0.971
UBE2J1	51465	YOD1	55432	0.971
VAMP8	8673	NAPA	8775	0.971
OSGEP	55644	TPD52L2	7165	0.971
GLRX3	10539	IDH1	3417	0.971
CRKL	1399	CBLB	868	0.970
STN1	79991	MED27	9442	0.970
TOPBP1	11073	BACH1	571	0.970
S100A16	140576	VAPB	9217	0.970
CSTF2	1478	GET4	51608	0.970
NFYB	4801	CNTN2	6900	0.970
ANAPC10	10393	PTTG1	9232	0.970
TWF2	11344	OGFOD1	55239	0.970
RAD9A	5883	CLSPN	63967	0.970
POLR2G	5436	RECQL5	9400	0.970
TGFB1I1	7041	TRAF4	9618	0.970
BRCA2	675	SEM1	7979	0.970
ATG3	64422	ATG12	9140	0.970

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
VAV2	7410	SOCS1	8651	0.970
STAG1	10274	PDS5A	23244	0.970
TIMM44	10469	RBMS1	5937	0.970
SOS1	6654	LRRK1	79705	0.970
BRD4	23476	AFF1	4299	0.970
TOPBP1	11073	RAD9A	5883	0.970
NME1	4830	WDR1	9948	0.970
WAPL	23063	PDS5A	23244	0.970
POLR2G	5436	MED28	80306	0.970
ADSL	158	API5	8539	0.970
ADSL	158	UBQLN2	29978	0.970
PHF1	5252	HIST1H3E	8353	0.970
ZAP70	7535	CBLB	868	0.970
HEXIM1	10614	CCNT1	904	0.970
CHAF1B	8208	HIST1H3E	8353	0.970
BAP1	8314	HAT1	8520	0.970
NDUFS8	4728	NDUFV2	4729	0.969
NCDN	23154	PPP1R2	5504	0.969
BID	637	FADD	8772	0.969
IRS1	3667	PIK3CA	5290	0.969
HIC1	3090	EED	8726	0.969
MARK2	2011	HDAC7	51564	0.969
ERLIN2	11160	INSIG1	3638	0.969
PLIN3	10226	ADSL	158	0.969
VAV2	7410	CBLB	868	0.969
RNF31	55072	TRIM25	7706	0.969
TMED9	54732	SEC22B	9554	0.969
NDUFA7	4701	STX7	8417	0.969
MAP2K7	5609	MAP2K4	6416	0.969
ARPC1A	10552	TPD52L2	7165	0.969
ANAPC11	51529	PTTG1	9232	0.969
TAL1	6886	TCF4	6925	0.969
PLIN3	10226	IPO11	51194	0.969
AFF1	4299	CCNT1	904	0.969
FERMT2	10979	ERP44	23071	0.969
UBQLN2	29978	RPA2	6118	0.969
DNM1L	10059	SEC24A	10802	0.969
TNK2	10188	PDGFRB	5159	0.969
NOD2	64127	RIPK1	8737	0.969
NDUFS2	4720	NDUFV2	4729	0.969
UBQLN2	29978	TPD52L2	7165	0.969
MPRIP	23164	RACGAP1	29127	0.969
NCOA6	23054	KMT2B	9757	0.969
GSPT2	23708	RDX	5962	0.969
CASP2	835	CASP7	840	0.969
TCEA1	6917	WDR61	80349	0.969
TAF6	6878	TAF13	6884	0.969
VPS29	51699	TBCD	6904	0.969
NFKB2	4791	MAP3K14	9020	0.969
ASS1	445	ATG3	64422	0.969
MAFG	4097	BACH1	571	0.968
VAMP8	8673	VAMP3	9341	0.968
SSSCA1	10534	VPS4B	9525	0.968
SMARCD1	6602	SMARCD2	6603	0.968
LEO1	123169	WDR61	80349	0.968
ANKRD28	23243	PPP6C	5537	0.968
BIRC3	330	BIRC5	332	0.968
LSM6	11157	LSM5	23658	0.968
FBXO5	26271	CDC23	8697	0.968
MED15	51586	TRIP4	9325	0.968
DNM2	1785	PACSIN3	29763	0.968
NUBP2	10101	PFDN2	5202	0.968
RBCK1	10616	TRIM25	7706	0.968
SAE1	10055	ADSL	158	0.968
MARK2	2011	USP21	27005	0.968
TICAM1	148022	RIPK1	8737	0.968
EXOSC2	23404	AICDA	57379	0.968
PLIN3	10226	RPRD1A	55197	0.968
PFDN2	5202	VBP1	7411	0.968
RAB8A	4218	SCP2	6342	0.968
FAS	355	BID	637	0.968
INTS1	26173	POLR2H	5437	0.968

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
SBDS	51119	STX7	8417	0.968
SERTAD1	29950	CCND2	894	0.968
LPP	4026	TPRKB	51002	0.968
PPP2CB	5516	PPP2R5D	5528	0.968
GTF3C2	2976	GTF3C5	9328	0.968
PRKCQ	5588	CARD11	84433	0.968
INTS9	55756	SEM1	7979	0.968
ANKFY1	51479	ROCK1	6093	0.968
POLR3C	10623	GTF3C2	2976	0.968
CHRAC1	54108	PTMS	5763	0.968
ATG7	10533	ATG3	64422	0.968
ZWINT	11130	NUF2	83540	0.968
USP33	23032	USP21	27005	0.968
UBE2O	63893	TRAF5	7188	0.968
SSSCA1	10534	UBQLN2	29978	0.968
PAWR	5074	SHMT1	6470	0.968
FERMT2	10979	DFFA	1676	0.968
SURF4	6836	SEPT7	989	0.968
ADSS	159	HSPE1	3336	0.968
KRT85	3891	USP8	9101	0.968
SCP2	6342	STX7	8417	0.967
SEC24A	10802	STMN2	11075	0.967
USP28	57646	UCHL3	7347	0.967
LPP	4026	PPP2R5D	5528	0.967
ASS1	445	CTSA	5476	0.967
FANCG	2189	BRCA2	675	0.967
CASP7	840	CASP8	841	0.967
TIRAP	114609	TLR4	7099	0.967
ETF1	2107	GSPT1	2935	0.967
MRPL50	54534	MRPL44	65080	0.967
MRPL10	124995	MRPL41	64975	0.967
CASP10	843	MAP3K14	9020	0.967
CASP2	835	CASP10	843	0.967
DIAPH1	1729	KLC1	3831	0.967
DNM1L	10059	VPS26A	9559	0.967
IPO11	51194	TPD52L2	7165	0.967
DMAP1	55929	YEATS4	8089	0.967
FAS	355	TRADD	8717	0.967
RYBP	23429	PCGF2	7703	0.967
ACBD3	64746	TBCD	6904	0.967
MALT1	10892	CARD11	84433	0.967
YAF2	10138	PCGF2	7703	0.967
RAD9A	5883	RAD17	5884	0.967
RIC8A	60626	WDR61	80349	0.967
INSR	3643	STAT5B	6777	0.967
PARD6A	50855	MARK4	57787	0.967
SNX6	58533	SNX1	6642	0.967
ADSS	159	WDR1	9948	0.967
ORC3	23595	ORC5	5001	0.967
MSH3	4437	SLX4	84464	0.967
FBXO5	26271	UBE2S	27338	0.967
APEH	327	PAWR	5074	0.967
NIPBL	25836	SP100	6672	0.967
RNF31	55072	RIPK2	8767	0.967
THY1	7070	SCO2	9997	0.967
FERMT2	10979	ALDH7A1	501	0.967
AP1B1	162	CLINT1	9685	0.967
RIPK1	8737	TAX1BP1	8887	0.967
TWF2	11344	PDIA4	9601	0.966
MRPL24	79590	MRPL45	84311	0.966
TYMS	7298	VPS4B	9525	0.966
RNF31	55072	TNF	7124	0.966
CBLC	23624	ZAP70	7535	0.966
WDR4	10785	ADSL	158	0.966
GTF2A1	2957	TAF1	6872	0.966
TAF1	6872	CCNT1	904	0.966
DRAP1	10589	TAF9B	51616	0.966
MAPK9	5601	MAP2K4	6416	0.966
PAFAH1B2	5049	PPP5C	5536	0.966
SCFD1	23256	NSF	4905	0.966
SCAF4	57466	COA7	65260	0.966
GSS	2937	PPP5C	5536	0.966

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
RASSF1	11186	LATS1	9113	0.966
DNMT3A	1788	RTF1	23168	0.966
SPEN	23013	RUNX1T1	862	0.966
SCFD1	23256	MRPL40	64976	0.966
VAV3	10451	INSR	3643	0.966
ERCC4	2072	SLX4	84464	0.966
TRAIP	10293	RNF114	55905	0.966
METTL1	4234	IPO11	51194	0.966
ERBB3	2065	ZAP70	7535	0.966
CEBPG	1054	FOSL1	8061	0.966
VAMP2	6844	VAPB	9217	0.966
RAP1GDS1	5910	NAE1	8883	0.966
SMAD6	4091	TSC22D1	8848	0.966
HRAS	3265	PIK3CA	5290	0.966
MRPL3	11222	MRPL42	28977	0.966
ORC3	23595	MCM10	55388	0.966
PFDN2	5202	TPD52L2	7165	0.966
FGFR1OP2	26127	ZRANB1	54764	0.966
TRIM33	51592	LDB1	8861	0.966
DNM2	1785	AMPH	273	0.966
CEBPD	1052	LEF1	51176	0.966
SYK	6850	CBLB	868	0.966
PCGF3	10336	BCOR	54880	0.966
BAIAP2	10458	KLC4	89953	0.966
PAWR	5074	SNX2	6643	0.966
USP28	57646	USP8	9101	0.966
WDR82	80335	ASH2L	9070	0.966
SRPRB	58477	RRBP1	6238	0.966
NUBP2	10101	DSTN	11034	0.966
YOD1	55432	TRIM54	57159	0.966
PACSIN3	29763	ITSN1	6453	0.966
STX4	6810	VAPB	9217	0.966
VPS29	51699	SHMT1	6470	0.966
LSM6	11157	LSM3	27258	0.966
TRIM23	373	USP2	9099	0.966
CFLAR	8837	MAP3K14	9020	0.966
MAP2K1	5604	MAP2K2	5605	0.965
SHMT1	6470	CARS	833	0.965
KIT	3815	CBLB	868	0.965
SMG1	23049	GSPT1	2935	0.965
SWAP70	23075	PPP2R5C	5527	0.965
UFM1	51569	OSGEP	55644	0.965
RPS6KA3	6197	NPEPPS	9520	0.965
GSS	2937	GINS3	64785	0.965
SMAD6	4091	ACVR1B	91	0.965
GARS	2617	UGDH	7358	0.965
GTF3C4	9329	GTF3C3	9330	0.965
SARS	6301	TBCD	6904	0.965
RAB8A	4218	RAB11B	9230	0.965
SEL1L	6400	SYVN1	84447	0.965
TDP2	51567	TRAF5	7188	0.965
DPP3	10072	ASS1	445	0.965
IDH1	3417	NME1	4830	0.965
SNF8	11267	ACBD3	64746	0.965
ADSL	158	WARS	7453	0.965
TWF2	11344	HK1	3098	0.965
CHMP5	51510	CHMP1B	57132	0.965
CASP10	843	TNFRSF10A	8797	0.965
GPS2	2874	TBL1XR1	79718	0.965
ALDH7A1	501	PGD	5226	0.965
REST	5978	CDYL	9425	0.965
FAS	355	CFLAR	8837	0.965
SH3GLB2	56904	AARSD1	80755	0.965
WWP1	11059	SMAD6	4091	0.965
PLIN3	10226	ANKMY2	57037	0.965
OSGEP	55644	P3H1	64175	0.965
RFC1	5981	RFC3	5983	0.965
INPP5D	3635	LRRK1	79705	0.965
TNFRSF1B	7133	TRADD	8717	0.965
KMT2A	4297	CCNT1	904	0.965
OSGEP	55644	NIF3L1	60491	0.965
ATG3	64422	ATG5	9474	0.965

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
MRPL3	11222	MRPL4	51073	0.965
PFDN2	5202	PFDN5	5204	0.965
PARD6A	50855	PARD6B	84612	0.965
AMBRA1	55626	ULK1	8408	0.965
ARAF	369	WNK1	65125	0.965
RIPK2	8767	TRAF4	9618	0.965
PAWR	5074	VPS29	51699	0.964
ADSL	158	NPLOC4	55666	0.964
INTS1	26173	CPSF3L	54973	0.964
PFDN1	5201	PFDN2	5202	0.964
CYLD	1540	RIPK1	8737	0.964
BIRC2	329	RIPK1	8737	0.964
ITK	3702	WAS	7454	0.964
NDUFS2	4720	NDUFS3	4722	0.964
METTL1	4234	TPD52L2	7165	0.964
WARS	7453	GTF3C4	9329	0.964
ING4	51147	JADE1	79960	0.964
ZWINT	11130	DSN1	79980	0.964
SRPRB	58477	RBMS1	5937	0.964
UFM1	51569	UBE2V2	7336	0.964
NCOA6	23054	MED15	51586	0.964
VAMP8	8673	SNAP23	8773	0.964
PEPD	5184	GDA	9615	0.964
PPP5C	5536	GINS3	64785	0.964
ATG4B	23192	ATG12	9140	0.964
PQBP1	10084	ZC3H11A	9877	0.964
RAB1A	5861	RAB7A	7879	0.964
NCOA1	8648	TRIP4	9325	0.964
RAB4A	5867	RABEP2	79874	0.964
KIT	3815	TEC	7006	0.964
CD2AP	23607	CBLB	868	0.964
RYBP	23429	BCOR	54880	0.964
LATS2	26524	AJUBA	84962	0.964
ADSS	159	IDH1	3417	0.964
C11orf58	10944	NAGK	55577	0.964
BRD4	23476	MED14	9282	0.964
MGA	23269	PCGF6	84108	0.964
MAP3K11	4296	MAP2K7	5609	0.964
ERCC1	2067	TAF7	6879	0.964
PAFAH1B2	5049	NAE1	8883	0.964
MRPL38	64978	MRPL11	65003	0.963
FBXO5	26271	ANAPC11	51529	0.963
EHD4	30844	GTF3C4	9329	0.963
TRIM5	85363	USP2	9099	0.963
RBCK1	10616	UBE2S	27338	0.963
NUBP2	10101	UBA6	55236	0.963
PFDN4	5203	VBP1	7411	0.963
MAPK10	5602	MAP2K4	6416	0.963
RYBP	23429	TFDP1	7027	0.963
PLIN3	10226	WDR4	10785	0.963
OS9	10956	UBE2J1	51465	0.963
RPRD1B	58490	STAT5B	6777	0.963
NDUFA7	4701	ZC3H11A	9877	0.963
MAT2B	27430	SNX1	6642	0.963
WDR61	80349	CTR9	9646	0.963
HES1	3280	FANCL	55120	0.963
ME1	4199	UGDH	7358	0.963
SUGP1	57794	STX7	8417	0.963
FADD	8772	TNFRSF10A	8797	0.963
PLIN3	10226	ANP32A	8125	0.963
HMG20A	10363	MTA3	57504	0.963
PPP1R8	5511	EED	8726	0.963
EPOR	2057	STAT5A	6776	0.963
RIPK1	8737	CFLAR	8837	0.963
CTPS2	56474	WARS	7453	0.963
DPP3	10072	PEPD	5184	0.963
USP33	23032	TRIM63	84676	0.963
SBDS	51119	RBMS1	5937	0.963
GDI2	2665	RAB11B	9230	0.963
GTF3C5	9328	GTF3C3	9330	0.963
ERLIN2	11160	DERL1	79139	0.963
PACSIN3	29763	WAS	7454	0.963

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
MRPL3	11222	MRPL9	65005	0.963
TAF6	6878	TAF7	6879	0.963
RBMS1	5937	SNX3	8724	0.963
SNX3	8724	ZC3H11A	9877	0.963
NUP98	4928	RAE1	8480	0.963
LGALS1	3956	PSMG1	8624	0.963
OLA1	29789	RAP1GDS1	5910	0.963
BCL2L1	598	TP53BP2	7159	0.963
MCL1	4170	BAX	581	0.963
KMT2D	8085	ASH2L	9070	0.963
PHF1	5252	EED	8726	0.963
SAE1	10055	WDR4	10785	0.962
ZRANB1	54764	STRIP1	85369	0.962
GTF2E2	2961	GTF2F2	2963	0.962
PDCD10	11235	HK1	3098	0.962
PDCD10	11235	ERP44	23071	0.962
NDUFA7	4701	SCO2	9997	0.962
UBE2S	27338	ANAPC11	51529	0.962
ERLIN2	11160	UFD1L	7353	0.962
GTF2B	2959	TCEA1	6917	0.962
MRPL13	28998	MRPL44	65080	0.962
AUP1	550	SYVN1	84447	0.962
PAK2	5062	UBE2R2	54926	0.962
HDLBP	3069	VPS36	51028	0.962
METTL1	4234	NPLOC4	55666	0.962
CSNK1D	1453	PPP1R14A	94274	0.962
TACC3	10460	SSSCA1	10534	0.962
MED4	29079	TRIP4	9325	0.962
TOMM40	10452	TOMM22	56993	0.962
TRPC4AP	26133	RIPK1	8737	0.962
FKBP9	11328	API5	8539	0.962
FANCC	2176	FANCL	55120	0.962
NSF	4905	SNAP23	8773	0.962
SUGP1	57794	ZC3H11A	9877	0.962
ZC3H15	55854	THOC2	57187	0.962
PQBP1	10084	RRBP1	6238	0.962
RAB7A	7879	RAB11A	8766	0.962
VAMP2	6844	SEC22B	9554	0.962
KLC1	3831	KLC4	89953	0.962
AGFG1	3267	TPRKB	51002	0.962
AP2A1	160	DAB2	1601	0.962
NUDC	10726	UGP2	7360	0.962
TRAF3	7187	CBLB	868	0.962
GTF2H1	2965	MNAT1	4331	0.962
VAV3	10451	PDGFRB	5159	0.962
TWF2	11344	ASNS	440	0.962
ATF6	22926	NFYA	4800	0.962
PHF8	23133	ASH2L	9070	0.962
TRIM37	4591	TRAF3	7187	0.962
POLR3C	10623	GTF3C4	9329	0.962
GTF2F2	2963	POLR2H	5437	0.962
TRAF5	7188	MAP3K14	9020	0.962
PTPN12	5782	UGP2	7360	0.962
ADSL	158	NT5C2	22978	0.962
MED15	51586	MED28	80306	0.962
MRPL38	64978	MRPL44	65080	0.962
TOMM22	56993	LAMTOR3	8649	0.962
TAF9B	51616	TAF1	6872	0.962
TAF7	6879	CCNT1	904	0.962
PHLPP1	23239	WDR20	91833	0.962
MRPL37	51253	MRPL45	84311	0.961
MECP2	4204	SKI	6497	0.961
PQBP1	10084	NDUFA7	4701	0.961
CABIN1	23523	AMPH	273	0.961
TSC22D3	1831	SGK1	6446	0.961
COMMD1	150684	RELB	5971	0.961
CXCR4	7852	SOCS3	9021	0.961
SNF8	11267	VPS36	51028	0.961
HGS	9146	ZFYVE9	9372	0.961
ARPC5	10092	TPRKB	51002	0.961
MRPL2	51069	MRPL23	6150	0.961
MAP2K2	5605	BRAF	673	0.961

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
NDUFS3	4722	NDUFV2	4729	0.961
CAV1	857	CD44	960	0.961
POLR1D	51082	POLR1B	84172	0.961
SBDS	51119	SNX3	8724	0.961
DCP2	167227	EDC4	23644	0.961
DERL2	51009	UBE2J1	51465	0.961
SEC61B	10952	SEC61A1	29927	0.961
PARD6A	50855	PRKCI	5584	0.961
AP1M2	10053	SCNN1B	6338	0.961
TICAM1	148022	TLR4	7099	0.961
MRPL40	64976	MRPL24	79590	0.961
S100A16	140576	SUCLA2	8803	0.961
CBX1	10951	MECP2	4204	0.961
TRIM32	22954	TRIM5	85363	0.961
TAL1	6886	LDB1	8861	0.961
E2F6	1876	TFDP1	7027	0.961
H2AFZ	3015	YEATS4	8089	0.961
PTPRS	5802	PPFIA1	8500	0.961
BIRC3	330	DZIP3	9666	0.961
DR1	1810	METTL1	4234	0.961
COPZ1	22818	TLE3	7090	0.961
COPZ1	22818	KLC1	3831	0.961
VPS29	51699	SNX6	58533	0.961
GNA12	2768	GSK3A	2931	0.961
UBE2J1	51465	SEL1L	6400	0.961
TAF9B	51616	TAF13	6884	0.961
TRAPPC8	22878	TRAPPC9	83696	0.961
OPTN	10133	TAX1BP1	8887	0.960
USP33	23032	TRIM54	57159	0.960
BCCIP	56647	KYNU	8942	0.960
RIPK3	11035	FADD	8772	0.960
CTDP1	9150	TRIP4	9325	0.960
POLR1A	25885	TAF1D	79101	0.960
ADSS	159	CHRAC1	54108	0.960
VPS36	51028	WDR12	55759	0.960
FAS	355	CASP10	843	0.960
UBE2V1	7335	RIPK1	8737	0.960
MRPL42	28977	MRPL37	51253	0.960
ATG4B	23192	ATG3	64422	0.960
VPS36	51028	VPS29	51699	0.960
BAIAP2	10458	MARK2	2011	0.960
DCK	1633	PLS3	5358	0.960
POLR2J	5439	MED9	55090	0.960
MARK3	4140	HDAC7	51564	0.960
TRIM23	373	RNF126	55658	0.960
MAFG	4097	ATF3	467	0.960
ADSL	158	XPNPEP1	7511	0.960
GTF2F2	2963	TCEA1	6917	0.960
BIRC2	329	NOD2	64127	0.960
GTF2B	2959	NCOA1	8648	0.960
MRPL2	51069	MRPL37	51253	0.960
GTF3C5	9328	GTF3C4	9329	0.960
RPA3	6119	AARSD1	80755	0.960
NDUFS2	4720	NDUFS6	4726	0.960
MED28	80306	TRIP4	9325	0.960
ABL1	25	CBLB	868	0.960
DR1	1810	MBIP	51562	0.960
ZPR1	8882	VPS4B	9525	0.960
ARAF	369	BRAF	673	0.960
DVL1	1855	AXIN1	8312	0.960
RRN3	54700	POLR1B	84172	0.960
TICAM1	148022	RNF216	54476	0.960
RXRG	6258	AJUBA	84962	0.960
TRAF1	7185	TRADD	8717	0.960
ING4	51147	MEAF6	64769	0.960
RNF7	9616	CDC34	997	0.960
LMO2	4005	TAL1	6886	0.960
NUDC	10726	PAPOLA	10914	0.960
STN1	79991	MED23	9439	0.960
CASP9	842	CASP10	843	0.960
UBA6	55236	OGFOD1	55239	0.960
NRDC	4898	PPP3CA	5530	0.960

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
WDR4	10785	UBQLN2	29978	0.960
SREBF1	6720	MED7	9443	0.960
MED28	80306	MED10	84246	0.960
PFDN2	5202	PFDN4	5203	0.960
TRIM54	57159	UCHL3	7347	0.959
PIK3C3	5289	BECN1	8678	0.959
GTF2E1	2960	SH3GL1	6455	0.959
UBE4B	10277	VPS4B	9525	0.959
ORC4	5000	ORC5	5001	0.959
ANAPC1	64682	PTTG1	9232	0.959
DERL1	79139	CD3D	915	0.959
ATP6V1A	523	RAB1A	5861	0.959
RUNX1	861	CBFB	865	0.959
MAD2L1	4085	CDC16	8881	0.959
ECHS1	1892	RAP1GDS1	5910	0.959
POLR2J	5439	MED28	80306	0.959
GTF2B	2959	POLR2H	5437	0.959
NUBP2	10101	PSME2	5721	0.959
ASF1B	55723	HIST1H3E	8353	0.959
USP28	57646	BAP1	8314	0.959
FBXO5	26271	CDC16	8881	0.959
TRAIP	10293	TRAF5	7188	0.959
TAF1	6872	TAF2	6873	0.959
POLR3C	10623	GTF3C3	9330	0.959
SGF29	112869	MBIP	51562	0.959
MED28	80306	CCNC	892	0.959
RIPK1	8737	RIPK2	8767	0.959
PLIN3	10226	CTPS2	56474	0.959
PQBP1	10084	RBMS1	5937	0.959
SMARCD2	6603	BAZ1B	9031	0.959
TAB2	23118	RIPK1	8737	0.959
USP21	27005	MARK4	57787	0.959
KIF3A	11127	KIFAP3	22920	0.959
ORC5	5001	CDC7	8317	0.959
GTF2F2	2963	POLR2G	5436	0.959
USP28	57646	STAM	8027	0.959
DFFA	1676	PFAS	5198	0.959
RIPK1	8737	FADD	8772	0.959
GOLGA2	2801	GORASP1	64689	0.959
STX5	6811	GOSR1	9527	0.959
RCOR1	23186	MTA3	57504	0.959
BID	637	TNFRSF1A	7132	0.959
ALDH7A1	501	PDIA4	9601	0.959
DCK	1633	TALDO1	6888	0.959
COPZ1	22818	COPG2	26958	0.959
PARK2	5071	CAMK2A	815	0.959
USP33	23032	CCP110	9738	0.959
MAP2K1	5604	BRAF	673	0.959
BCL2L11	10018	MCL1	4170	0.959
WDR12	55759	TBCD	6904	0.959
BID	637	TNFRSF10B	8795	0.959
CASP8	841	CASP9	842	0.959
ASS1	445	PEPD	5184	0.959
UBA6	55236	XPNPEP1	7511	0.959
PAWR	5074	PPM1G	5496	0.959
NUDC	10726	VPS26A	9559	0.959
UBE4B	10277	SBDS	51119	0.959
PSMG3	84262	PSMG1	8624	0.959
FAS	355	FADD	8772	0.959
PAICS	10606	SWAP70	23075	0.959
UBA6	55236	UCHL3	7347	0.959
RUFY1	80230	TELO2	9894	0.959
MAPK7	5598	SGK1	6446	0.959
MED6	10001	STN1	79991	0.959
CSNK1D	1453	FHL1	2273	0.959
EXOC1	55763	EXOC4	60412	0.959
WDR5	11091	WDR61	80349	0.959
MED22	6837	MED7	9443	0.959
STX4	6810	VAMP2	6844	0.958
STX4	6810	NAPA	8775	0.958
RBCK1	10616	USP2	9099	0.958
MRPL42	28977	MRPL44	65080	0.958

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
RNF31	55072	TRAF1	7185	0.958
NUBP2	10101	XPNPEP1	7511	0.958
NDUFS3	4722	NDUFV1	4723	0.958
UBE2V2	7336	ZPR1	8882	0.958
WDR82	80335	KMT2D	8085	0.958
HIC1	3090	PHF1	5252	0.958
MCL1	4170	BID	637	0.958
GSPT2	23708	GSPT1	2935	0.958
NME1	4830	PLS3	5358	0.958
MRPL3	11222	MRPL10	124995	0.958
ADSS	159	TALDO1	6888	0.958
INSR	3643	IRS2	8660	0.958
NUBP2	10101	OGFOD1	55239	0.958
MYO6	4646	TAX1BP1	8887	0.958
DFFA	1676	PPP5C	5536	0.958
TNF	7124	TRADD	8717	0.958
ORC6	23594	CDC45	8318	0.958
GORASP2	26003	ACBD3	64746	0.958
UBQLN2	29978	USP34	9736	0.958
PQBP1	10084	SCO2	9997	0.958
ABCD1	215	ABCD3	5825	0.958
TNFRSF1A	7132	RIPK2	8767	0.958
EXOC8	149371	EXOC4	60412	0.958
GABARAPL2	11345	ULK1	8408	0.958
GTF2H1	2965	TCEA1	6917	0.958
RCOR1	23186	TAL1	6886	0.958
PAFAH1B2	5049	RAP1GDS1	5910	0.958
PPP3CA	5530	P3H1	64175	0.958
HERC2	8924	CCP110	9738	0.958
STAT6	6778	NCOA1	8648	0.958
MAP3K2	10746	MAP2K7	5609	0.958
TNPO3	23534	IPO9	55705	0.958
DR1	1810	IPO11	51194	0.958
PNP	4860	RAB1A	5861	0.958
E2F6	1876	PCGF6	84108	0.958
PHLPP1	23239	WDR48	57599	0.958
PFDN2	5202	PPP2CB	5516	0.958
FEN1	2237	ELAC2	60528	0.958
FKBP9	11328	UBA6	55236	0.958
AHR	196	GTF2F2	2963	0.958
EXOC1	55763	DST	667	0.958
PLIN3	10226	SRP9	6726	0.957
BAIAP2	10458	PAK2	5062	0.957
CD2AP	23607	BCAR1	9564	0.957
TFDP1	7027	PCGF6	84108	0.957
NOD1	10392	RIPK2	8767	0.957
ERCC1	2067	SLX4	84464	0.957
STRN4	29888	PPP2R5C	5527	0.957
SNX6	58533	FAM129B	64855	0.957
PFDN5	5204	VBP1	7411	0.957
TNF	7124	RIPK1	8737	0.957
LLGL1	3996	PRKCI	5584	0.957
SMAD7	4092	LEF1	51176	0.957
TRAF3	7187	RIPK1	8737	0.957
TRAF1	7185	RIPK1	8737	0.957
RACGAP1	29127	KIF23	9493	0.957
PLS3	5358	PTMS	5763	0.957
GTF2H1	2965	XPC	7508	0.957
POLR2H	5437	POLR3D	661	0.957
ULK1	8408	RB1CC1	9821	0.957
SMAD6	4091	BMPR1A	657	0.957
HSPE1	3336	WDR1	9948	0.957
TTC9C	283237	UBE2V2	7336	0.957
MAP3K4	4216	MAP2K7	5609	0.957
TNK2	10188	YES1	7525	0.957
HARS	3035	ELAC2	60528	0.957
STAM	8027	HGS	9146	0.957
THRA	7067	MED12	9968	0.957
TRIM37	4591	FXR2	9513	0.957
NDUFA7	4701	RRBP1	6238	0.957
WDR4	10785	P3H1	64175	0.957
BAIAP2	10458	ALDH7A1	501	0.957

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
MRPL4	51073	MRPL1	65008	0.957
VPS36	51028	VPS26A	9559	0.957
PPP1R2	5504	RAP1GDS1	5910	0.957
POLR2D	5433	MED28	80306	0.957
ERCC3	2071	MNAT1	4331	0.957
TIRAP	114609	IRAK4	51135	0.957
MRPL37	51253	MRPL41	64975	0.957
ERO1A	30001	PDIA4	9601	0.957
EPOR	2057	INPP5D	3635	0.957
SBDS	51119	SRPRB	58477	0.957
TNFRSF1A	7132	RIPK1	8737	0.957
SULT1A1	6817	TPM2	7169	0.957
VAMP2	6844	SNAP23	8773	0.957
IFIT5	24138	LONP1	9361	0.957
CNOT7	29883	CDC7	8317	0.957
GABARAPL2	11345	ATG4B	23192	0.957
CCNB1	891	CDC25C	995	0.957
CASP3	836	CASP10	843	0.957
GORASP2	26003	GTF2A1	2957	0.957
IKZF1	10320	SIN3B	23309	0.957
RPRD1A	55197	NPLOC4	55666	0.957
OSGEP	55644	RPRD1B	58490	0.957
VPS4B	9525	USP34	9736	0.957
MNAT1	4331	TRIM5	85363	0.957
SMC2	10592	NCAPD2	9918	0.957
MAP3K4	4216	TRAF4	9618	0.957
GPS1	2873	IRF5	3663	0.957
RIPK2	8767	CFLAR	8837	0.957
MCM10	55388	CDC7	8317	0.956
PHC2	1912	PCGF2	7703	0.956
RASSF1	11186	RASSF5	83593	0.956
MED14	9282	QKI	9444	0.956
KIF1BP	26128	SH3GL1	6455	0.956
DERL2	51009	SYVN1	84447	0.956
TNFAIP3	7128	RIPK1	8737	0.956
FADD	8772	CFLAR	8837	0.956
MAP3K20	51776	RPS6KA5	9252	0.956
ATG7	10533	TBCD	6904	0.956
USP21	27005	UCHL3	7347	0.956
PDGFRB	5159	SNX2	6643	0.956
NUBP2	10101	FKBP9	11328	0.956
ADSS	159	CHMP4A	29082	0.956
GTF2F1	2962	TAF1	6872	0.956
BCR	613	LRRK1	79705	0.956
APC2	10297	ANAPC11	51529	0.956
UBE2C	11065	ANAPC5	51433	0.956
PELO	53918	ABCD3	5825	0.956
TAB2	23118	RIPK2	8767	0.956
STAM2	10254	HGS	9146	0.956
PIK3CA	5290	ARHGEF1	9138	0.956
TRIM37	4591	TRAF5	7188	0.956
CHERP	10523	DHX8	1659	0.956
MRPL41	64975	MRPL44	65080	0.956
TRIM39	56658	USP2	9099	0.956
COPE	11316	COPG2	26958	0.956
PDCD10	11235	FKBP9	11328	0.956
GINS3	64785	GINS4	84296	0.956
MTPN	136319	RAB1A	5861	0.956
ANAPC4	29945	MAD2L1	4085	0.956
GSR	2936	MVD	4597	0.956
SAE1	10055	TPD52L2	7165	0.956
NDC80	10403	MIS12	79003	0.956
MNAT1	4331	TRIM39	56658	0.956
UBE2J1	51465	SYVN1	84447	0.956
E2F4	1874	ASH2L	9070	0.956
SAE1	10055	LPP	4026	0.956
UBQLN2	29978	UBL7	84993	0.956
POLR1A	25885	POLR1D	51082	0.956
SLC25A10	1468	AMBRA1	55626	0.956
SH3GLB1	51100	PPP2R5D	5528	0.956
SNX1	6642	UGDH	7358	0.956
SBDS	51119	SCO2	9997	0.956

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
EDC4	23644	NMT1	4836	0.956
BID	637	CASP3	836	0.956
INTS1	26173	ASUN	55726	0.956
MARK2	2011	MARK3	4140	0.956
STRN3	29966	PTPA	5524	0.956
GTF3C2	2976	GTF3C4	9329	0.956
PSAP	5660	SURF4	6836	0.956
FZR1	51343	PTTG1	9232	0.956
IDH1	3417	TALDO1	6888	0.956
SNX6	58533	SNX4	8723	0.956
UBE2S	27338	ANAPC2	29882	0.956
KPNA6	23633	KPNA3	3839	0.956
UBE2E2	7325	MKRN3	7681	0.956
HSPA4L	22824	APEH	327	0.956
UBAC1	10422	ADRM1	11047	0.956
AGFG1	3267	LAP3	51056	0.956
BIRC2	329	CASP9	842	0.956
MRPL2	51069	MRPL40	64976	0.956
TRIM54	57159	USP8	9101	0.956
NACC1	112939	BCOR	54880	0.956
ERO1A	30001	P3H1	64175	0.956
APEH	327	CARS	833	0.956
UBE2C	11065	CDC23	8697	0.955
PTMS	5763	WDR1	9948	0.955
NCOA2	10499	NR1I2	8856	0.955
LEO1	123169	TCEA1	6917	0.955
WDR12	55759	TSG101	7251	0.955
IL1R1	3554	TOLLIP	54472	0.955
BIRC3	330	RNF31	55072	0.955
PDCD10	11235	PPP2R1B	5519	0.955
BCAP31	10134	DERL1	79139	0.955
ACTR3	10096	TPD52L2	7165	0.955
ERCC1	2067	ERCC4	2072	0.955
GABARAPL2	11345	ATG3	64422	0.955
RBFOX2	23543	TOLLIP	54472	0.955
MYCN	4613	NTRK1	4914	0.955
NGFR	4804	TRAF3	7187	0.955
WDR4	10785	ERO1A	30001	0.955
MAP3K11	4296	MAP2K4	6416	0.955
TROVE2	6738	UGP2	7360	0.955
SRPRB	58477	SCO2	9997	0.955
SMC2	10592	NCAPG	64151	0.955
POLR2F	5435	MED30	90390	0.955
TRIP4	9325	MED27	9442	0.955
PTPN12	5782	BCAR1	9564	0.955
FAS	355	TNF	7124	0.955
LSM2	57819	DHX16	8449	0.955
RYBP	23429	PCGF1	84759	0.955
MED16	10025	STN1	79991	0.955
SCP2	6342	SCO2	9997	0.955
ERBB3	2065	IL6ST	3572	0.955
VPS35	55737	SNX6	58533	0.955
RBCK1	10616	TNF	7124	0.955
TTC9C	283237	UBL7	84993	0.955
INTS1	26173	INTS9	55756	0.955
NUBP2	10101	ATP6V1F	9296	0.955
STMN2	11075	STAM	8027	0.955
TAL1	6886	RUNX1	861	0.955
MARK2	2011	MARK4	57787	0.955
EHD1	10938	API5	8539	0.955
HRAS	3265	RIN1	9610	0.955
NAGK	55577	BCCIP	56647	0.955
BAG1	573	SH3GL1	6455	0.955
CASP8	841	RIPK1	8737	0.955
GSS	2937	PAFAH1B2	5049	0.955
NUBP2	10101	PRDX5	25824	0.955
BAIAP2	10458	KLC1	3831	0.955
JUNB	3726	ATF4	468	0.955
POLR3C	10623	POLR3D	661	0.955
MED28	80306	MED12	9968	0.955
MYO1E	4643	CAPNS1	826	0.955
ID3	3399	TCF4	6925	0.955

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
BCAP31	10134	CASP8	841	0.955
CAPN2	824	PDIA4	9601	0.955
MRPL4	51073	MRPL24	79590	0.955
BRCA2	675	BRCC3	79184	0.955
NDUFS2	4720	NDUFV1	4723	0.955
USP21	27005	TRIM54	57159	0.955
PQBP1	10084	SRPRB	58477	0.955
NUBP2	10101	NUDCD2	134492	0.955
MECP2	4204	SP3	6670	0.955
RLIM	51132	SIAH1	6477	0.955
BRCC3	79184	SPAG9	9043	0.955
NUBP2	10101	UBQLN2	29978	0.955
CDV3	55573	STX7	8417	0.955
TRIM54	57159	JOSD1	9929	0.955
HARS	3035	SHMT1	6470	0.955
WASHC4	23325	CD2AP	23607	0.955
EIF4E	1977	ASS1	445	0.955
PBRM1	55193	CHD7	55636	0.955
PARD6A	50855	RPAP3	79657	0.955
AGFG1	3267	P3H1	64175	0.955
UBE2V1	7335	ATP6V1F	9296	0.955
ASUN	55726	INTS3	65123	0.955
TDP2	51567	TRAF3	7187	0.955
SARS	6301	ACBD3	64746	0.955
AP1M2	10053	LOXL4	84171	0.955
CFLAR	8837	DEDD	9191	0.955
PIK3CA	5290	IRS2	8660	0.954
SEC24A	10802	SEC24C	9632	0.954
TIMM44	10469	SBDS	51119	0.954
MAP4K1	11184	GRAP2	9402	0.954
BAK1	578	BCL2L1	598	0.954
UBQLN2	29978	STAM	8027	0.954
TNF	7124	SHARPIN	81858	0.954
VAMP2	6844	VAMP8	8673	0.954
CHORDC1	26973	NAGK	55577	0.954
FAS	355	TNFRSF1A	7132	0.954
PJA1	64219	UCHL3	7347	0.954
FGFR1OP2	26127	STRIP1	85369	0.954
H3F3A	3020	SNX6	58533	0.954
AMBRA1	55626	RPTOR	57521	0.954
TGFBR1	7046	CD44	960	0.954
BPTF	2186	SMARCA1	6594	0.954
GTF2B	2959	ATF4	468	0.954
RAB11A	8766	RAB11B	9230	0.954
FLOT1	10211	FLOT2	2319	0.954
SEPT9	10801	SUCLG2	8801	0.954
PRDX5	25824	IGBP1	3476	0.954
RELB	5971	BCL3	602	0.954
UGP2	7360	ZYX	7791	0.954
ZMYND8	23613	INTS3	65123	0.954
LEO1	123169	CCNT1	904	0.954
PDLIM5	10611	NPLOC4	55666	0.954
HK1	3098	GTF3C4	9329	0.954
SCFD1	23256	MRPL13	28998	0.954
ANAPC5	51433	TRIM33	51592	0.954
TRIP4	9325	MED17	9440	0.954
BZW2	28969	PFAS	5198	0.954
MED15	51586	QKI	9444	0.954
AMPH	273	ITSN1	6453	0.954
CASP2	835	CASP8	841	0.954
ECHS1	1892	GINS3	64785	0.954
NUBP2	10101	PLIN3	10226	0.954
RDX	5962	SNX2	6643	0.954
CRKL	1399	EPOR	2057	0.954
DAPK1	1612	FADD	8772	0.954
OPTN	10133	TRAF3	7187	0.954
MRPS28	28957	MRPL42	28977	0.954
TTC9C	283237	UBE2V1	7335	0.954
NDUFA7	4701	VDAC3	7419	0.954
TES	26136	ZYX	7791	0.954
HSPBP1	23640	ERO1A	30001	0.954
PPP1R12A	4659	P3H1	64175	0.954

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
RNF38	152006	RNF114	55905	0.954
CEBPD	1052	JUNB	3726	0.954
SH3GL1	6455	CALR	811	0.954
SCAF4	57466	VAPB	9217	0.954
MRPL2	51069	MRPL14	64928	0.954
WDR5	11091	WDR82	80335	0.954
SRGAP2	23380	EXOC1	55763	0.954
UCHL3	7347	ATP6V1F	9296	0.954
RPRD1A	55197	ANKMY2	57037	0.954
XIAP	331	BIRC5	332	0.954
HYOU1	10525	BCAR1	9564	0.954
IRF7	3665	TNFAIP3	7128	0.954
RELB	5971	DPF2	5977	0.954
NDC80	10403	KNL1	57082	0.954
BIRC3	330	CASP9	842	0.954
VPS4B	9525	VPS26A	9559	0.954
PARD6A	50855	PARD3	56288	0.954
USP28	57646	CLSPN	63967	0.954
HSPE1	3336	IDH1	3417	0.954
USP28	57646	USP5	8078	0.954
VDAC3	7419	VAPA	9218	0.954
MTPN	136319	ATP6V1A	523	0.953
ATG7	10533	MAT2B	27430	0.953
BRAF	673	BRAP	8315	0.953
MRPL9	65005	MRPL19	9801	0.953
TRIM37	4591	NGFR	4804	0.953
RDX	5962	ANP32E	81611	0.953
ERLIN2	11160	SYVN1	84447	0.953
RNF38	152006	DZIP3	9666	0.953
WDR4	10785	OGT	8473	0.953
STAM2	10254	USP8	9101	0.953
TGFBR1	7046	ARHGEF6	9459	0.953
CEBPD	1052	GATA1	2623	0.953
ATG7	10533	SNX1	6642	0.953
HARS	3035	TTC1	7265	0.953
PPP2R5D	5528	PPFIA1	8500	0.953
RBCK1	10616	SHARPIN	81858	0.953
GTF2E1	2960	SH3GLB2	56904	0.953
PQBP1	10084	VDAC3	7419	0.953
STAM	8027	USP8	9101	0.953
SEC24A	10802	DSTN	11034	0.953
SEC24A	10802	UBE2B	7320	0.953
USP33	23032	NEURL4	84461	0.953
TPRKB	51002	IPO11	51194	0.953
CPSF1	29894	FIP1L1	81608	0.953
MRPS28	28957	NDUFA7	4701	0.953
PPP1R12A	4659	PFDN5	5204	0.953
SATB1	6304	TAL1	6886	0.953
DCK	1633	CHMP4A	29082	0.953
THRA	7067	MED25	81857	0.953
UBE3A	7337	HERC2	8924	0.953
LATS2	26524	MOB1A	55233	0.953
PNKP	11284	LIG3	3980	0.953
DMWD	1762	PHLPP1	23239	0.953
MNAT1	4331	MTA1	9112	0.953
KMT2A	4297	ASH2L	9070	0.953
XIAP	331	RIPK1	8737	0.953
PARD6A	50855	PRKCZ	5590	0.953
INTS6	26512	POLR2G	5436	0.953
ERP44	23071	PFAS	5198	0.953
MNAT1	4331	MKRN3	7681	0.953
KAT6A	7994	RUNX1	861	0.953
PLIN3	10226	NPLOC4	55666	0.953
ID1	3397	TCF4	6925	0.953
DVL1	1855	DVL3	1857	0.953
SIN3B	23309	REST	5978	0.953
BID	637	CASP8	841	0.953
IRS1	3667	PLCG2	5336	0.953
EPN1	29924	EHD2	30846	0.953
MAP1S	55201	RASSF5	83593	0.953
NOD2	64127	SHARPIN	81858	0.953
TNFRSF1A	7132	BCL10	8915	0.953

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
FZR1	51343	CCNA1	8900	0.953
AMPH	273	ITSN2	50618	0.953
ARR3	407	STAM	8027	0.953
SAE1	10055	OSGEP	55644	0.953
BIRC3	330	TNFRSF1A	7132	0.953
UBR1	197131	ATP6V1A	523	0.953
PXN	5829	BCAR1	9564	0.953
MRPL37	51253	MRPL24	79590	0.953
FBXO5	26271	ANAPC4	29945	0.953
TWF2	11344	CAPZA2	830	0.953
GOT1	2805	IDH1	3417	0.953
VPS36	51028	ACBD3	64746	0.953
PIH1D1	55011	TELO2	9894	0.953
SHMT1	6470	TTC1	7265	0.953
AGFG1	3267	SH3GLB1	51100	0.952
PDCD10	11235	PFAS	5198	0.952
PDLIM5	10611	IPO11	51194	0.952
SHC1	6464	CBLB	868	0.952
ZWINT	11130	MIS12	79003	0.952
GTF2F2	2963	MED29	55588	0.952
PACSIN2	11252	DNM1	1759	0.952
USP4	7375	BRAP	8315	0.952
KLC1	3831	TLE3	7090	0.952
BRCA2	675	BRE	9577	0.952
SULT1A1	6817	UCHL3	7347	0.952
AUP1	550	SEL1L	6400	0.952
EXOSC6	118460	EXOSC1	51013	0.952
PPP5C	5536	RAP1GDS1	5910	0.952
WDR4	10785	NPLOC4	55666	0.952
OPTN	10133	TNF	7124	0.952
PAK4	10298	BAIAP2	10458	0.952
C11orf58	10944	SNCAIP	9627	0.952
GDI2	2665	RAB11A	8766	0.952
FKBP9	11328	LIMD1	8994	0.952
TPM2	7169	TPM3	7170	0.952
HARS	3035	XPO5	57510	0.952
TOM1L2	146691	SULT1A1	6817	0.952
MCL1	4170	BAK1	578	0.952
LATS2	26524	SNAI1	6615	0.952
PFAS	5198	RAP1GDS1	5910	0.952
POLR2D	5433	RECQL5	9400	0.952
NUBP2	10101	UBE2V2	7336	0.952
GTF2E1	2960	SARS	6301	0.952
SHMT2	6472	SNX1	6642	0.952
BID	637	CASP10	843	0.952
HARS	3035	PFDN1	5201	0.952
TCEA1	6917	CTR9	9646	0.952
POLR2G	5436	INTS3	65123	0.952
CRKL	1399	BCAR1	9564	0.952
AKAP11	11215	PRKAR1A	5573	0.952
DMAP1	55929	SMARCAD1	56916	0.952
IPO11	51194	CALU	813	0.952
KDM6A	7403	KMT2D	8085	0.952
KAT6A	7994	ING5	84289	0.952
PPP2R5D	5528	RANGAP1	5905	0.952
NDUFV1	4723	NDUFV2	4729	0.952
MAPK8	5599	BID	637	0.952
SCP2	6342	VDAC3	7419	0.952
MRPL2	51069	MRPL24	79590	0.952
FKBP9	11328	UBR7	55148	0.952
RBCK1	10616	RNF31	55072	0.952
NDC80	10403	NSL1	25936	0.952
MARK2	2011	MAP3K3	4215	0.952
MEAF6	64769	KAT6A	7994	0.952
GGA2	23062	IGF2R	3482	0.952
TOMM40	10452	SBDS	51119	0.952
PRKCI	5584	PARD6B	84612	0.952
MRPS18B	28973	MRPS35	60488	0.952
ANAPC5	51433	SMARCAD1	56916	0.952
FAS	355	CAV1	857	0.952
TRAF1	7185	MAP3K14	9020	0.952
CRKL	1399	ERBB3	2065	0.952

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
BRD7	29117	SMARCD1	6602	0.952
MAP2K6	5608	MAP2K4	6416	0.952
APEH	327	SHMT1	6470	0.952
IRF7	3665	FADD	8772	0.952
EPOR	2057	IRS2	8660	0.952
BCL2L1	598	BID	637	0.952
BIRC5	332	DIABLO	56616	0.952
NPHP1	4867	BCAR1	9564	0.952
NFYC	4802	TAF6	6878	0.952
GTF2B	2959	NCOR1	9611	0.952
APC2	10297	CDC16	8881	0.952
MCRS1	10445	NFRKB	4798	0.952
GCDH	2639	NOS3	4846	0.952
GSPT1	2935	SNX2	6643	0.952
LATS2	26524	LATS1	9113	0.952
TNK2	10188	SIAH2	6478	0.951
MED29	55588	MED28	80306	0.951
SMAD5	4090	RUNX3	864	0.951
WAS	7454	WIPF1	7456	0.951
GTF2F2	2963	POLR2E	5434	0.951
TRIP6	7205	BCAR1	9564	0.951
STAG1	10274	STAG2	10735	0.951
TBCB	1155	MYO1E	4643	0.951
USP28	57646	PJA1	64219	0.951
RIPK1	8737	TNFRSF10B	8795	0.951
DNM1L	10059	FIS1	51024	0.951
RAP1A	5906	RHEB	6009	0.951
GTF2A1	2957	TAF6	6878	0.951
IKZF1	10320	CHD3	1107	0.951
ERCC2	2068	GTF2H1	2965	0.951
CASP10	843	DEDD	9191	0.951
TNFRSF10B	8795	TNFRSF10A	8797	0.951
UFM1	51569	ZPR1	8882	0.951
NUBP2	10101	TWF2	11344	0.951
USF1	7391	ASH2L	9070	0.951
ANXA6	309	THOP1	7064	0.951
IDH1	3417	EFHD2	79180	0.951
FANCC	2176	HES1	3280	0.951
NUBP2	10101	SEC24A	10802	0.951
MAST1	22983	LONP1	9361	0.951
TXNDC5	81567	BZW1	9689	0.951
HRAS	3265	MAP2K1	5604	0.951
ME1	4199	TBCD	6904	0.951
RABGAP1	23637	RAB7A	7879	0.951
KLC2	64837	KLC4	89953	0.951
COPE	11316	ARCN1	372	0.951
UFD1L	7353	USP13	8975	0.951
ECHS1	1892	PFAS	5198	0.951
H2AFZ	3015	DMAP1	55929	0.951
MED16	10025	THRA	7067	0.951
LPP	4026	IPO11	51194	0.951
STAT3	6774	CD44	960	0.951
VPS4A	27183	CHMP4A	29082	0.951
ERCC1	2067	TAF6	6878	0.951
DIAPH1	1729	COPZ1	22818	0.951
RAB1B	81876	RAB11B	9230	0.951
POLR2D	5433	INTS3	65123	0.951
AP2B1	163	CLINT1	9685	0.951
ANAPC10	10393	UBE2C	11065	0.951
NUBP2	10101	API5	8539	0.951
PCGF3	10336	BMI1	648	0.951
UBE2D4	51619	RNF7	9616	0.951
THOC3	84321	THOC1	9984	0.951
KIF1BP	26128	SH3GLB1	51100	0.951
MRPS28	28957	MRPL41	64975	0.951
MRPS35	60488	MRPS5	64969	0.951
HARS	3035	ACBD3	64746	0.951
SEPT9	10801	SEPT11	55752	0.951
CYLD	1540	USP13	8975	0.951
AP2A1	160	AMPH	273	0.951
NDC80	10403	AURKB	9212	0.951
PPP3CA	5530	VBP1	7411	0.951

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
SOCS1	8651	IRS2	8660	0.951
TRAF1	7185	CFLAR	8837	0.951
ATP6V1F	9296	VPS26A	9559	0.950
VDAC3	7419	VAPB	9217	0.950
CSNK1D	1453	DVL1	1855	0.950
FBXO5	26271	ANAPC5	51433	0.950
TRIM33	51592	CDC16	8881	0.950
RAB1A	5861	RAB11A	8766	0.950
NUBP2	10101	UBE2R2	54926	0.950
TRIM33	51592	CDC23	8697	0.950
YAF2	10138	BCOR	54880	0.950
BACH1	571	MAFK	7975	0.950
GTF2E1	2960	TCEA1	6917	0.950
MYD88	4615	TRAF3	7187	0.950
MCRS1	10445	TERT	7015	0.950
HTRA2	27429	BIRC3	330	0.950
TAF2	6873	TAF6	6878	0.950
GSS	2937	ASNS	440	0.950
TRIM33	51592	ANAPC1	64682	0.950
GEMIN5	25929	SRP9	6726	0.950
CBLC	23624	MET	4233	0.950
OSGEP	55644	BAG1	573	0.950
RICTOR	253260	MLST8	64223	0.950
SUGP1	57794	SNX3	8724	0.950
ARPC5	10092	ARPC3	10094	0.950
LSM4	25804	LSM3	27258	0.950
NGFR	4804	TRAF5	7188	0.950
UBE2D4	51619	TRIM25	7706	0.950
KIF1BP	26128	OSGEP	55644	0.950
NIF3L1	60491	RPA3	6119	0.950
UBE2B	7320	ATP6V1F	9296	0.950
GLRX3	10539	EFHD2	79180	0.950
SMARCAD1	56916	CDC23	8697	0.950
STN1	79991	MED17	9440	0.950
PDCD10	11235	UBR7	55148	0.950
UBE2V1	7335	TRIM5	85363	0.950
DPY30	84661	ASH2L	9070	0.950
NUBP2	10101	UBE2B	7320	0.950
UBE2V1	7335	UCHL3	7347	0.950
ATP6V1A	523	USP5	8078	0.950
INTS1	26173	PPP2R1B	5519	0.950
CBX1	10951	H3F3A	3020	0.950
MRPL38	64978	MRPL45	84311	0.950
SWAP70	23075	OSBP	5007	0.950
QKI	9444	MED13	9969	0.950
MKRN3	7681	TRIM5	85363	0.950
GTF2B	2959	POLR2E	5434	0.950
AGFG1	3267	LPP	4026	0.950
FANCD2	2177	FANCL	55120	0.950
PHLPP1	23239	USP46	64854	0.950
RBCK1	10616	TRAF1	7185	0.950
FANCC	2176	FANCM	57697	0.950
CD2AP	23607	RABEP2	79874	0.950
WDR4	10785	PPME1	51400	0.950
PEX19	5824	ABCD3	5825	0.950
POLR3H	171568	POLR1D	51082	0.950
HCFC1	3054	ASH2L	9070	0.950
AP1M2	10053	EPN1	29924	0.950
POLR1A	25885	POLR1E	64425	0.950
DDX39A	10212	SMS	6611	0.950
ARCN1	372	KLC1	3831	0.950
RDX	5962	EZR	7430	0.950
KANSL1	284058	EXOC1	55763	0.950
SAP30BP	29115	STRN4	29888	0.950
MRPS16	51021	MRPS9	64965	0.950
SULT1A1	6817	TPM4	7171	0.950
PNKP	11284	XRCC4	7518	0.950
SEC24A	10802	USP34	9736	0.950
CHEK2	11200	PPP2R5C	5527	0.950
NAGK	55577	LDHAL6B	92483	0.950
H3F3A	3020	ASF1B	55723	0.950
SULT1A1	6817	UBE2B	7320	0.950

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
WDR33	55339	FIP1L1	81608	0.950
MCM2	4171	MCM10	55388	0.950
SNX6	58533	ANP32E	81611	0.950
ASH2L	9070	KMT2B	9757	0.950
RAD9A	5883	DNAJC7	7266	0.950
BRD4	23476	CTDP1	9150	0.950
ING3	54556	YEATS4	8089	0.950
STN1	79991	MED13	9969	0.950

Gene 1		Gene 2		SINaTRA
Symbol	GeneID	Symbol	GeneID	
STMN2	11075	SULT1A1	6817	0.950
TAF6	6878	NCOR1	9611	0.950
ATG4B	23192	AMBRA1	55626	0.950
MED28	80306	MED25	81857	0.950
USP46	64854	WDR20	91833	0.950
HSPBP1	23640	OGT	8473	0.950
SEC61B	10952	SGTA	6449	0.950

ACKNOWLEDGEMENTS

This chapter is also a reproduction, in part, of a publication in PLOS Computational Biology by Jacunski *et al*. I thank Brent R. Stockwell, Hossein Khiabanian, and Cameron Palmer for their valuable input and editorial suggestions. Furthermore, I apologize for the repetitive nature of this acknowledgement. I hope you know I'm never gonna let you down like this again. Nor will I make you cry, say goodbye, tell a lie, or hurt you.

CHAPTER 5 – SYNTHETIC LETHALITY AND DRUG SYNERGY

INTRODUCTION

Drug combination therapy (DCT) has been used to treat infectious diseases and cancers [106,107]. In many cases, DCT involves drug synergy, where the individual components' effects are magnified when co-administered. DCT offers a number of benefits compared to single-drug therapies, including fewer and less severe side effects [108] and increased ability to combat the development of drug resistance [37,106,109]. Given the rising prevalence of drug resistance in cancer [110] and the high cost and attrition rates in the development of new drugs [111], identifying novel DCTs is particularly important.

Synthetic lethality (SL) has been suggested as a guide for identifying potential DCTs [21,68,112]. A gene pair is dubbed synthetic lethal when knocking out either gene causes no adverse effect in a cell, but knocking out both leads to cell death [22]. Although SL has been extensively studied in yeast [23,29], finding human synthetic lethal pairs has an extremely high experimental burden, especially given the potential for cell-, disease-, and tissue-specific SL pairs [69]; identifying all SL pairs in a single biological context would take at least 200 million pairwise tests. Therefore, most research to date regarding human SL has focused on computational models [102], including our previous work [66].

Although the identification of synergistic drug pairs has important therapeutic benefits, it is an ongoing problem in experimental biology. First, many methods for calculating synergy exist. Researchers may use effect-based strategies such as highest single agent or Bliss independence [46], or dose-effect-based strategies, such as Loewe additivity [39]. Here, we will focus on Bliss

independence, one of the most common measures. Drug combinations are evaluated using Excess Over Bliss (EOB), which is calculated using the effect of individual components.

Even after selecting a method, there is no established method to quantitatively assess the significance of interactions. For example, when considering areas of potential synergy or antagonism using EOB, analysis still has a tendency towards qualitative observation rather than quantitative assessment. It is typical to describe trends towards either synergy or antagonism when considering EOB, rather than defining areas of specific, statistically significant synergy. Therefore, an Excess Over Bliss score of 0.0001 can be as indicative of synergy as a score of 0.01 or 0.9.

Furthermore, EOB calculation typically does not fully account for variance in either control or dosed experiments. Specifically, the expected effect of a drug combination is calculated using a probabilistic model based on the effect of each drug alone. Single-drug effects are typically calculated by comparing measured levels of ATP of treated samples and calculating percent viability of each sample based on positive and negative control. In addition, EOB itself is reported as a median value with some standard error. This does not allow researchers to fully appreciate the variance between control samples or replicates.

Here, we develop DAVISS (Data-driven Assessment of Variability In Synergy Scores), a novel statistical method to measure the significance of drug synergy. We first fit dose-response curves to cell count data, including control wells. We use the distribution of control counts around the curve to generate a background distribution of EOB for comparison to our experimental results. We use the fitted curves to calculate the percent effect of each single-drug dose on cellular growth and compute EOB for each drug pair at each concentration combination. In order to assess the statistical significance of concentration-specific drug synergy, we perform

outlier testing using the control distribution; furthermore, we can also identify combination-wide synergy by comparing the distribution of all EOB scores for a particular drug pair to the control distribution.

We apply DAVISS to test predictions of human synthetic lethality we generated using SINaTRA (Species-INdependent TRAnslation) [66], as described in Chapter 4. Here, we identify five cancer-associated genes (*CSF1R*, *ERBB2*, *KIT*, *PTK2B*, *STAT5B*) where all possible pairs are associated with high SINaTRA scores; these are predicted to be synthetic lethal. We then identify another cancer-associated gene (*PDE10A*) that has very low SINaTRA scores with these five genes. These are our control, non-synthetic-lethal pairs. We map each gene to drugs with high target specificity and test all predicted SL combinations for drug synergy in four human cancer cell lines, and all predicted non-SL combinations in three human cancer cell lines.

We identify 3/10 predicted SL pairs associated with significant, consistent drug synergy over four cell lines (Amuvatinib/PF-431396, BLZ945/PF-431396, BLZ945/Mubritinib), compared to 0/5 predicted non-SL pairs in three cell lines. Our hit rate greatly exceeds the expected rate of SL (0.1% [77]). We also find that putative SL pairs are enriched for synergy at specific concentrations compared to predicted non-SL pairs. Finally, we identify three novel, cell-specific drug combinations: Amuvatinib/ Mubritinib and BLZ945/Mubritinib in CAL148, and BLZ945/PF-431396 in HS606T. These results suggest the high utility of DAVISS as a method of assessing the significance of drug synergy, and of SINaTRA as a viable guide for finding novel DCTs.

RESULTS

Previous work suggests areas of possible drug synergy

In the previous chapter, we described SINaTRA (Species-INdependent TRAnslation) [66], a machine-learning algorithm that allows us to predict human synthetic lethal pairs using *S. cerevisiae* experimental data and both yeast and human protein-protein interaction networks. We used this model to predict the likelihood of synthetic lethality for over 100 million gene pairs, which we reported as SINaTRA scores ranging from 0 (non-SL) to 1 (very likely SL). We identified 52 genes associated with cancer drugs and clustered them by SINaTRA score, and found that some regions of high SINaTRA score were significantly associated with a large number of single-drug and drug combination cancer therapies. From these results, we hypothesized that SINaTRA can be used to identify novel synergistic drug pairs operating through a mechanism of synthetic lethality.

As a proof of concept, we selected five genes of interest (*CSF1R*, *ERBB2*, *KIT*, *PTK2B*, *STAT5B*) from a series of cancer-drug-associated, predicted human SL pairs ("original gene set"; see Figure 4.8 in the previous chapter [66] & *Materials and Methods*). These genes have previously been associated with cancer, either generally [113,114] or with specific subtypes [115-118]. The SINaTRA scores of all possible pairs of the genes of interest, as well as the appropriate drug combinations, are found in Table 5.1A. The associated drugs are selective inhibitors for all genes of interest [119-121] except PF-431396, which has a reported IC50 of 2nM and 11nM in PTK2 and PTK2B, respectively [122].

	Gene 1		Drug 1	Gene 2		Drug 2	SINaTRA
A.	**ERBB2**	2064	Mubritinib	**PTK2B**	2185	PF-431396	0.88
	CSF1R	1436	BLZ945	**KIT**	3815	Amuvatinib	0.79
	PTK2B	2185	PF-431396	**STAT5B**	6777	CAS285986-31-4	0.71
	ERBB2	2064	Mubritinib	**KIT**	3815	Amuvatinib	0.7
Predicted SL	**KIT**	3815	Amuvatinib	**PTK2B**	2185	PF-431396	0.69
	KIT	3815	Amuvatinib	**STAT5B**	6777	CAS285986-31-4	0.68
	CSF1R	1436	BLZ945	**PTK2B**	2185	PF-431396	0.67
	ERBB2	2064	Mubritinib	**STAT5B**	6777	CAS285986-31-4	0.67
	CSF1R	1436	BLZ945	**ERBB2**	2064	Mubritinib	0.53
	CSF1R	1436	BLZ945	**STAT5B**	6777	CAS 285986-31-4	0.442
B.	**STAT5B**	6777	CAS 285986-31-4	**PDE10A**	10846	PF-2545920	0.063
	CSF1R	1436	BLZ945	**PDE10A**	10846	PF-2545920	0.058
	ERBB2	2064	Mubritinib	**PDE10A**	10846	PF-2545920	0.043
Pred. non-SL	**PTK2B**	2185	PF-431396	**PDE10A**	10846	PF-2545920	0.04
	KIT	3815	Amuvatinib	**PDE10A**	10846	PF-2545920	0.038

Table 5.1: Selected predicted SL and non-SL pairs and their drugs
A.) *All pair combinations of our genes of interest (CSF1R, ERBB2, KIT, PTK2B, STAT5B) and their associated drugs and SINaTRA scores.* B.) *Our five genes of interest and our selected negative control gene, PDE10A, with associated drugs and SINaTRA scores.*

Gene-drug database provides negative controls

We found that the median SINaTRA score for all combinations of genes from the original set was 0.407. This is significantly higher than the median of all gene pairs in the human network (0.122; $p<2.2e-16$, Mann-Whitney U). The lowest SINaTRA score from the original gene set is 0.12, which is in the 49.5th percentile of all scores. We concluded that any possible gene pair from the original set was too likely to be SL to be considered a good negative control for our experiments.

Therefore, we broadened our search to the Drug-Gene Interaction Database [123,124]. We identified 394 cancer-therapy-associated genes (*Materials and Methods*) and clustered them by SINaTRA score. We observed that our genes of interest remain close together (Figure 5.A.1). Of the filtered genes, we selected PDE10A, which has a SINaTRA score of 0.063 or lower (≤23rd

percentile) with all of our original genes of interest, and which is selectively inhibited by PF-2545920 [125] (Table 5.1B).

Dose-response curves provide background information

In order to account for experimental background, we fit drug curves using calculated viability and included control wells ([Drug]=0μM) (see *Materials and Methods* and Figures 5.A.2-3). We found that drug curves within a cell line had a consistent starting raw count (median: 0.1%; Table 5.A.1). This meant that percentage inhibitions computed for drug combinations would not be affected by large deviations between the individual components' "no drug" points.

Simulated analysis illustrates DAVISS result output

We manually constructed three datasets and a background distribution to illustrate the output of DAVISS (*Materials and Methods*). First, we consider the standard method of exploring drug synergy (Figure 5.1A), where the median EOB for all replicates of a drug combination is plotted as a heat map, with the dose of one drug on the x-axis, and the other on the y-axis. Drug interaction is assessed qualitatively, with higher scores indicating synergy and lower ones representing antagonism.

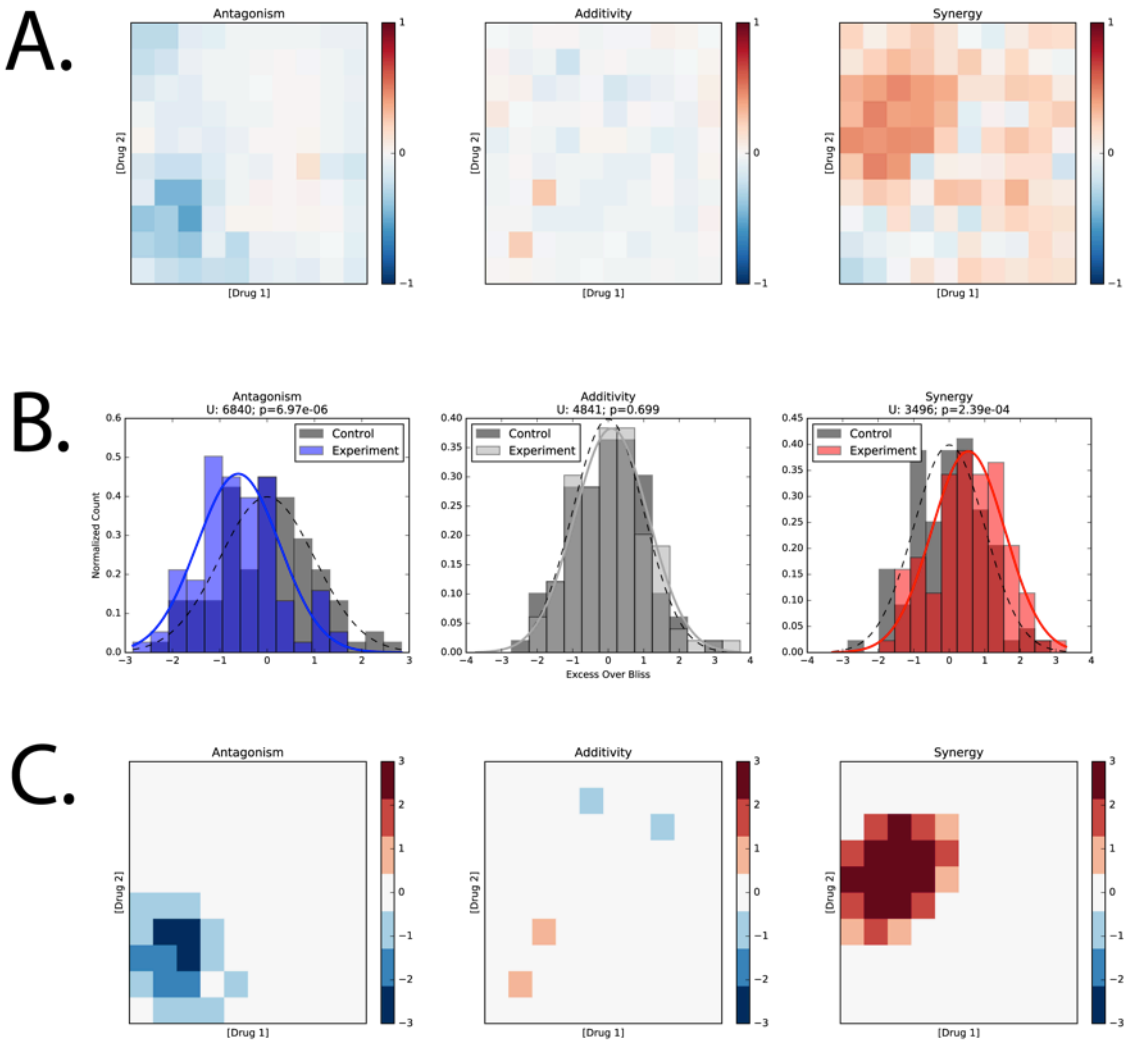


Figure 5.1: Simulated illustration of DAVISS output
A.) *Simulated antagonism (left), additivity (middle), and synergy (right) using Excess Over Bliss (EOB). Drug interactions are assessed visually, rather than statistically.* B.) *Simulated EOB for three drug combinations. The dark gray distribution represents the EOB distribution of control wells in all panels. A significant leftward divergence from the control indicates a trend towards antagonism for the combinatorial experiment (left, blue). Insignificantly different distributions indicate no synergy or antagonism (center, light gray). A significant rightward divergence from the control distribution indicates a combination-wide trend towards synergy (right, red). Significance is assessed using Mann-Whitney U.* C.) *Concentration-specific synergy tests. A higher positive value means more replicates are significantly synergistic (dark red), and a more negative value means more replicates are significantly antagonistic (dark blue). A value of 0 corresponds to no significant replicates, and is represented by light gray. Here, we show antagonism (left; three concentration combinations are significantly antagonistic in all three replicates, four combinations have two significant replicates, and 11 have one), additivity (middle; we consider additivity not just as no significant replicates whatsoever, but also as an equal number of significantly antagonistic and synergistic concentrations), and synergy (right; 10 concentration combinations with three significant replicates each, seven combinations with two significant replicates, and four with one significant replicate) assessed for individual concentrations in a drug combination. The x-axis represents the concentrations of Drug 1, while the y-axis represents those of Drug 2.*

We calculate the overall synergy of a combination by testing these EOB values against a simulated control distribution and illustrate the potential outcomes of such a test in Figure 5.1B. A number of EOB scores significantly lower than the null distribution indicates combination-wide trend towards antagonism (Figure 5.1B, left), while high EOB scores indicate a trend towards synergy (Figure 5.1B, right). Insignificant divergence from the null indicates additivity (Figure 5.1B, center). Significance is measured using the Mann-Whitney U test [64]. In this work, we use the terms "general synergy" and "general antagonism" to denote combinations that exhibit significant synergy or antagonism according to this test.

We can also test the significance of a particular replicate's EOB score by comparing it to the null distribution, which allows us to quantitatively assess the specific drug interactions, providing greater depth to the EOB heat maps that are typically used to report synergy. We illustrate such test results in Figure 5.1C. We represent antagonism in the leftmost panel, additivity in the center, and synergy in the right. We use "specific synergy" and "specific antagonism" to denote significant results according to this test for a particular drug combination.

Combining EOB, concentration-specific significance, and synergistic trends illustrates drug synergy

In order to evaluate the synergy within each drug combination in each cell line, we can combine the EOB heat map, concentration-specific assessment of synergy significance, and combination-wide trend towards synergy (Figure 5.2).

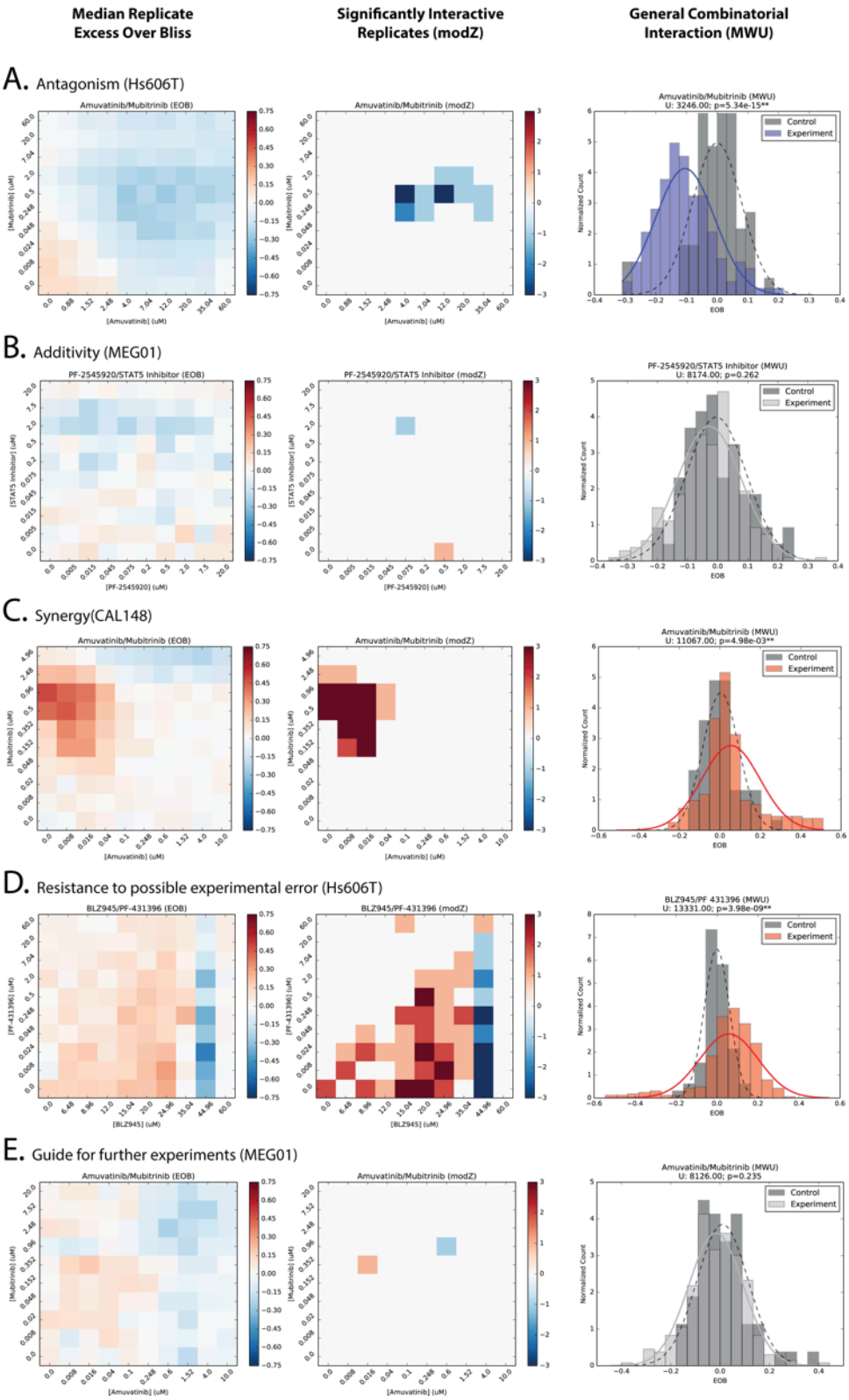


Figure 5.2: Experimental examples of drug interactions
A.) *Clear synergy* B.) *Additivity* C.) *Synergy* D.) *Trend towards additivity (MWU), but area of significant synergy that could be explored (modZ)* E.) *Clearly significant trend synergy (MWU), while maintaining robustness to possible experimental error (modZ and EOB heat map)*

In Figures 5.2A-C, we show antagonism, additivity, and synergy in three experimental examples. Figure 5.2A shows specific and general antagonism in PF-2545920/CAS 285986-31-4 in Hep-3B217; there are many replicates with specific antagonism (center), and the combination is also generally synergistic (right). Figure 5.2B illustrates additivity as shown in PF-2545920/PF-431396 in MEG01; although the EOB heat map (left) shows some areas of potential synergy and antagonism, there are only four significant replicates that are significantly antagonism (center), and the combination can be considered additive. Finally, we show synergy with Amuvatinib/Mubritinib in CAL148 in Figure 5.2C; there is both very high specific synergy (center), and a significant trend to synergy for the entire combination (right).

Some examples illustrate the utility of our method. For example, in Figure 5.2D, we see a column of antagonism when looking at the significant synergy of Mubritinib/PF-431396 in Hs606T. Because of the rigid pattern, this looks like possible experimental error, perhaps due to a faulty multipipette at one level of dosage. However, in spite of this, our method still manages to pick up general synergy across the entire combination (right).

In addition, we also show how the method can be used to inform further experiments. In Figure 5.2E, we see general antagonism in BLZ945/Mubritinib in Hep-3B217. However, there is clear synergy at one specific concentration (center); therefore, further experiments could focus on exploring more concentration combinations around that location to find consistent, general synergy. Full data are available in the supplementary materials; Figures 5.A.4-5.A.7 are predicted SL pairs, while Figures 5.A.8-5.A.10 are predicted non-SL.

We summarize the results for all combinations in all cell lines in Table 5.2. When we assess general synergy, we find that 3/10 predicted SL pairs exhibit only synergy or additivity,

compared to 0/5 non-SL pairs (non-significant; p=0.2637, one-sided Fisher's Exact Test). When considering specific synergy, 2/10 predicted SL pairs exhibit only synergy or additivity in specific concentrations, compared to 0/5 predicted non-SL ones (non-significant; p=0.4286, one-sided Fisher's Exact Test). Furthermore, 4/5 predicted non-SL combinations exhibit only antagonism and additivity, compared to 0/10 predicted SL pairs (p=0.0037, one-sided Fisher's Exact Test). Finally, we find significantly more putative SL drug combinations exhibit significant, specific synergy compared to putative non-SL combinations (14/40 vs. 1/15; p=0.0326, one-sided Fisher's Exact Test).

We then evaluate the concordance of general and specific synergy for each cell line and drug combination and find that 3/10 predicted SL pairs exhibit only concordant synergy or additivity, compared to 0/5 in predicted non-SL pairs (non-significant; p=0.2637, one-sided Fisher's Exact Test). Furthermore, 0/10 predicted SL pairs exhibit only concordant antagonism or additivity, compared to 3/5 predicted non-SL pairs (p=0.0220, one-sided Fisher's Exact Test).

Finally, we assess each drug pair for overall synergy and find that 3/10 predicted SL drug pairs exhibit overall synergy compared to 0/10 predicted non-SL pairs (non-significant; p=0.2637, one-sided Fisher's Exact Test). In contrast, 1/10 predicted SL pairs exhibit overall antagonism, compared to 3/5 predicted non-SL (p=0.0769, one-sided Fisher's Exact Test).

	Gene 1		Drug 1	Gene 2		Drug 2	SINaTRA	General Synergy				Specific Synergy				Concordance				Overall
								CAL148	HEP-3B217	Hs606T	MEG01	CAL148	HEP-3B217	Hs606T	MEG01	CAL148	HEP-3B217	Hs606T	MEG01	
A. Predicted SL	ERBB2	2064	Mubitrinib	PTK2B	2185	PF-431396	0.88	-1	1	1	1	-1	0	-1	1	-1	0	0	1	0
	KIT	3815	Amuvatinib	CSF1R	1436	BLZ945	0.79	0	0	0	-1	0	0	0	1	0	0	0	0	0
	PTK2B	2185	PF-431396	STAT5B	6777	CAS 285986-31-4	0.71	-1	0	1	0	-1	0	-1	1	-1	0	0	0	-1
	KIT	3815	Amuvatinib	ERBB2	2064	Mubitrinib	0.7	1	-1	0	0	1	-1	0	0	1	-1	0	0	0
	KIT	3815	Amuvatinib	PTK2B	2185	PF-431396	0.69	0	1	1	0	0	-1	1	0	0	0	1	0	1
	KIT	3815	Amuvatinib	STAT5B	6777	CAS 285986-31-4	0.68	0		0	0	0		0	0	0		0	0	0
	CSF1R	1436	BLZ945	PTK2B	2185	PF-431396	0.67	0	1	1	0	1	1	1	-1	0	1	1	0	1
	ERBB2	2064	Mubitrinib	STAT5B	6777	STAT5B	0.67	0	0	0	0	0	1	-1	1	0	0	0	0	0
	CSF1R	1436	BLZ945	ERBB2	2064	Mubitrinib	0.53	1	1	0	0	1	1	0	0	1	1	0	0	1
	CSF1R	1436	BLZ945	STAT5B	6777	CAS 285986-31-4	0.442	0	0	0	0	1	0	0	1	0	0	0	0	0
B. Pred. NSL	PDE10A	10846	PF-2545920	STAT5B	6777	CAS 285986-31-4	0.063	1	-1		0	-1	0		-1	0	0		0	0
	CSF1R	1436	BLZ945	PDE10A	10846	PF-2545920	0.058	1	-1		1	0	-1		-1	0	-1		0	-1
	ERBB2	2064	Mubitrinib	PDE10A	10846	PF-2545920	0.043	1	-1		-1	0	-1		0	0	-1		0	-1
	PTK2B	2185	PF-431396	PDE10A	10846	PF-2545920	0.04	0	-1		1	-1	-1		1	0	-1		1	0
	KIT	3815	Amuvatinib	PDE10A	10846	PF-2545920	0.038	0	-1		1	0	-1		0	0	-1		0	-1

Table 5.2: Results of statistical tests of synergy

In "General Synergy," cell-specific combinations are given +1 if they indicate overall synergy, -1 if they indicate overall antagonism, and 0 otherwise (Mann-Whitney U test). In "Specific Synergy," cell-specific combinations are given +1 if there are more synergistic concentration pairs than antagonistic ones; -1 if there are more antagonistic ones; and 0 otherwise (modified Z score). In "Concordance," cell-specific combinations are given +1 if it is both generally and specifically synergistic; -1 if it is both generally and specifically antagonistic; and 0 otherwise. A combination is given an overall score of +1 if all cell lines have a concordance score of +1 or 0; -1 if all cell lines have a concordance score of -1 or 0; and 0 otherwise.

DISCUSSION

The identification of drug combination therapy is important to the treatment of cancer because of its ability to prevent the development of drug resistance. Synthetic lethality (SL) has been suggested as a method of identifying DCT in humans; however, it is rare, occurring in only 1/1000 gene pairs [77]. Furthermore, the experimental elucidation of SL bears a high experimental and financial burden. Thus, in the previous chapter, we developed a computational model of SL by creating SINaTRA [66].

In this chapter, we assess the results of SINaTRA in fifteen drug pairs associated with either high-scoring putative SL pairs, or low-scoring predicted non-SL pairs. We test these in 3-4 cell lines using drugs specific for each gene (Table 5.1).

DAVISS: Data-driven Assessment of Variability In Synergy Scores

In order to assess the significance of each drug pair's interaction, we develop a novel statistical model called DAVISS, which is based on Bliss independence and integrates the calculated cell viability distribution of control wells. This allows us to calculate the statistical significance of EOB in drug-treated samples at both specific concentrations and across an entire combinatorial experiment in a simple, clear manner that allows us to quantitatively assess drug synergy and antagonism. We also show that we can use the concordance of general and specific synergy scores to assess the overall synergy of each drug combination across any number of cell lines. Furthermore, we show that our method is robust to experimental error (Figure 5.4D), in addition to suggesting further areas of inquiry if the original drug concentrations missed specific areas of potential synergy (Figure 5.4E).

In spite of these features, there are certain limitations to DAVISS. For example, the current iteration of this method necessitates the fitting of a dose-response curve. Although this is

beneficial in lowering the need to alter single-drug responses to avoid using negative values in the formula for EOB, this does mean that DAVISS requires
a larger number of concentrations than a less statistical approach to Bliss independence. Furthermore, although DAVISS is highly quantitative in its assessment of both combination-wide and specific synergy and antagonism, it still requires some qualitative interpretation in order to evaluate consistency and overall synergy.

Finally, it is worth noting that our current model only accounts for variability in control wells when denoting the significance of a particular combination's synergy. It may be possible to account for the complexity provided by the variation exhibited in the replicates of single-drug-dosed samples; however, we believed that the marginal benefit of such an analysis would be slim.

In future work, we also hope to update DAVISS for smaller numbers of concentrations and experiments.

SINaTRA as a guide for predicting drug combination therapy

Overall, we identify 3/10 predicted SL pairs associated with significant, consistent drug synergy over four cell lines (Amuvatinib/PF-431396, BLZ945/PF-431396, BLZ945/Mubritinib), compared to 0/5 predicted non-SL pairs in three cell lines. This suggests the utility of SL as a method of predicting cancer drug combinations (Table 5.2), as it significantly exceeds the expected hit rate of 0.1% ($p<0.0001$, one-tailed Fisher's Exact Test).

Although this is a promising first look at SINaTRA as a method for identifying novel drug combination therapies, we have only considered a small number of gene pairs and cell lines. Furthermore, this work has underscored the complications associated with identifying SL in multicellular organisms, as drug synergy is often inconsistent between cell lines.

In summary, we believe that SINaTRA is a viable tool for guiding the discovery of novel drug combination therapies for cancer; using DAVISS in conjunction with it allows for the rigorous assessment of synergy across combination experiments.

METHODS

Previous work suggests areas of possible drug synergy

In the previous chapter [66], we developed an interspecies model of synthetic lethality (SL) based on protein-protein interactions of two species. There, we predicted the SINaTRA score for over 100 million human gene pairs. We found that, when clustering 52 genes associated with cancer drugs (the "original gene set;" described in Chapter 4 of this thesis), areas of high SINaTRA scores were associated with high densities of known single and combination cancer drugs. We selected Box 2 from Figure 4.4 of the previous chapter, which contained 11 unique genes and was significantly enriched for cancer-drug associations ($p<2.2e-16$). From these, we selected our five genes of interest (*ERBB2, PTK2B, KIT, CSF1R, STAT5B*) because drugs inhibiting their activity had three or fewer targets according to SelleckChem.com.

Gene-Drug Database Provides Negative Controls

We selected all genes associated with cancer drugs from the following databases compiled in the Drug-Gene Interaction database (DGIdb [123,124]): Cancer Commons, CIViC, Clearity Foundation Clinical Trials, My Cancer Genome, TALC, and TTD. We labeled this gene list the "expanded gene set." We clustered these (Figure 5.A.1) and found that our five drugs of interest co-localize near each other. To identify a negative control gene, we first identified all genes in the expanded set that had a SINaTRA score of <0.2 (~49^{th} percentile) with all genes of interest. We next manually filtered these by selecting all genes associated with specific drugs (≤3 targets according to SelleckChem.com), then identified the gene with the lowest SINaTRA scores for all five of our original genes of interest.

Cell Growth, Drug Dosing, and Measurement

We select four cell lines for use in our experiments: CAL148, HEP-3B217, Hs606T, and MEG01. The density of each cell line was first optimized to ensure linear cell growth in the

tissue culture treated 384-well plates (Greiner Bio-One 781080) for the duration of the experiment. Starting with 16,000 cells per well, the cells were 2-fold serially diluted to test 10 different concentrations of growth in the microplate. Cell-Titer-Glo (Promega Corp.) was used to measure total ATP levels of the wells every 24 hours for 96 hours. Optimal cell density was chosen based on the linear growth of cells by graphing the total luminescence count versus time.

Each cell line was then plated employing the Cell::Explorer automation system (under HEPS filtered conditions) at the optimal density into white, sterile, tissue culture treated 384 well plates on a Perkin Elmer Janus Automated workstation. The Janus is equipped with a 96 tip Modular Dispense Technology (MDT) pipetting head and it was used with sterile tips (235µL, Perkin Elmer 69000045) for plating 50µL of the cell solution into the microplates. The plates were incubated in the Liconic STX-500 ICSA for 24 hours prior to drug addition.

To generate a concentration response curve of each compound in the combination, the HP D300 Digital Dispenser was used to dispense specific amounts of the drug for a titration curve. Each concentration of the drug was dispensed in triplicate by the Digital Dispenser using HP's inkjet technology.

After 48 hours of incubation with the drug, the Cell::Explorer removed the plates out of the incubator and placed them in the Liconic LPX 200 Hotel to let them equilibrate to room temperature. 25µL of Cell Titer Glo were added using the Perkin Elmer Flexdrop PLUS Precision Reagent Dispenser. After shaking at 600 rpm for 5 minutes, the plates were read by the Perkin Elmer Envision 2104 using an enhanced luminescence protocol. The viability of each well was then calculated utilizing the control wells in each plate.

Calculation of drug curves

Each drug combination was tested in triplicate over three plates. We quantile normalized each set of plates containing the same combination(s); therefore, if plates 1-3 contain

combinations of drugs A+B and drugs C+D, and plates 4-6 contain combinations of drugs A+E and drugs C+E, we quantile normalize plates 1-3 to each other, and plates 4-6 to each other; these two groups (1-3 and 4-6) are not normalized to each other.

Therefore, we computed individual drug curves for each plate set. We began each curve from a concentration of 0μM. In order to plot this with drug dosage on a log scale, we replaced [Drug]=0μ with $[\text{Drug}]_{\min}$*0.1μM for each drug in each cell line.

For the dose-response curve, we used a logistic curve following the equation:

$$y = \frac{c}{1 + e^{-k(x-x_o)}} + y_0$$

where x is the log(concentration) of drug and y is the calculated viability, referred to as "cell count" throughout the remainder of the methods for simplicity. We bootstrapped the data 100 times and used least-squares implemented with the SciPy package to fit the curve, beginning with seeds of x_0=median(drug concentration), y_0 = median(cell count), c=max(control cell counts)-min(control cell counts), and k=1.0. We then selected the curve with the lowest root mean squared error (RMSE) for the original data (Figure 5.A.2). In cases where the dose-response curve ends higher than the beginning, we fit a flat line (explained in next section). All curves are shown in Figure 5.A.3.

In order to measure the consistency between curve starting points, we calculated the median and percent difference from median for each starting value of the drug curve (where [Drug]=0μM). These are reported in Table 5.A.1.

Calculation of Drug Synergy using Bliss Independence

Drug effect is measured using percent inhibition. For each drug curve for each cell line, we calculate the effect of a single drug at each concentration XμM using:

$$\% \text{ Inhibition} = \frac{f(X\mu M)}{f(0\mu M)}$$

where *f*() is the function of the fitted curve, *f*(XμM) is cell count at a dose of XμM, and *f*(0μM) is the cell count at the highest point of the curve, where drug dose is 0μM.

For each drug combination in each cell line, we calculated the effect of both drugs using the same formula; however, as the denominator, we use the mean value for *f*(0μM) of both drugs.

To calculate drug synergy, we used the Bliss independence measure;[46] in particular, we used the excess over Bliss (EOB),

$$\text{EOB} = E_{AB} - (E_A + E_b - E_A E_B)$$

where E_A is the effect of drug A alone, E_B is the effect of drug B alone, and E_{AB} is the effect of both drugs in combination. Here, we measure effect as percent inhibition. An EOB > 0 implies synergy; EOB < 0 implies antagonism; EOB = 0 implies additivity.

We use percent inhibition for E_A and E_B. Importantly, neither E_A nor E_B can be negative; this is why flat lines must be used for certain curves.

The significance of combination-wide synergy can be calculated using null distributions

When [A] and [B] = 0μM, $E_A = E_B = 0$. Therefore, when [A]=[B]= 0μM, EOB = E_{AB}. Because we have 60 control wells per combination (20/plate), we are able to use them as a null distribution of EOB.

In order to account for the potential non-normality of the null or experimental distributions, we can use the Mann-Whitney U test to compare them. The expected value of U is described as:

$$E[U] = \frac{n_1 n_2}{2}$$

where n_1 is the number of data points in the null distribution, and n_2 is the number of data points in the combination. We describe the experimental dataset as synergistic if U>E[U] and $p \leq 0.05$; it is antagonistic if U<E[U] and $p<0.05$. Otherwise, it is additive.

We illustrate this principle using simulated data, where we sample 100 numbers each from normal distributions representing the null (μ=1.0, σ=0.0), additive (μ=0.01, σ=0.99), synergistic (μ=0.5, σ=1.0), and antagonistic (μ=-0.5, σ=1.0). An experimental distribution that meets the criteria of synergy is coloured red; those of antagonism, blue; and those of additivity, light gray.

The significance of synergy in at particular concentrations can be calculated using the null distribution

For each replicate of each drug combination, we calculated its significance using a modified Z score [126], such that:

$$\text{modZ} = \frac{c(x_i - \text{med}(x))}{\text{MAD}}$$

where x_i is the datum, x is the data, and c=0.6745 (E(MAD)=0.6745 σ) [126], and MAD (median absolute deviation) is defined as:

$$\text{MAD} = \text{med}(|x_i - \text{med}(x)|)$$

We define |modZ|≥3.5 as significant, where modZ ≥ 3.5 as significantly synergistic, and modZ≤3.5 as significantly antagonistic, for a given replicate. This level is chosen based on the suggestion of Iglewicz *et al*. [126].

We create a heat map of significance by adding one point for every dose combination replicate that has significant synergy, and subtracting one point for every one that has significant antagonism. Therefore, a dose combination with a value of +3 exhibits significant synergy at each replicate of the combination, while one with a value of -2 exhibits antagonism in only two replicates. We ensured that no single dose combination exhibits both significant synergy and

significant antagonism, which would lead to a nullification of significant replicates at that point (*i.e.* +1 for synergy and -1 for antagonism would lead to an overall indication of additivity).

The combination of EOB, concentration-specific significance, and synergistic trends illustrates a clear picture of drug synergy experiments

In order to assess overall synergy in each drug combination, we combine general and concentration-specific evaluations of synergy. In order to assess general synergy, all EOB scores of cell-specific combinations are tested against the null distribution using the Mann-Whitney U test. The combination is given +1 if we observe significant overall synergy, -1 if they indicate overall antagonism, and 0 otherwise.

In order to assess specific synergy, we identify the number of drug concentration combinations that are significantly synergistic and antagonistic according to modZ. We evaluate a cell line as having significant specific synergy (+1) if there are more synergistic concentration pairs than antagonistic ones; antagonistic (-1) if there are more antagonistic ones than synergistic ones; and additive (0) otherwise (modified Z score).

Concordance of a drug combination in a cell line is evaluated using the similarity between general and concentration-specific synergy. If both are +1, then the cell line/drug combination is given a concordance score of +1; if both are -1, then the concordance score is -1. Else, the concordance score is 0.

A drug combination is given an overall score of +1 if all cell lines have a concordance score of +1 or 0; -1 if all cell lines have a concordance score of -1 or 0; and 0 otherwise. Therefore, a drug combination with an overall score of -1 is evaluated as synergistic; one with an overall score of -1 is evaluated as antagonistic. Otherwise, it is considered additive.

APPENDIX

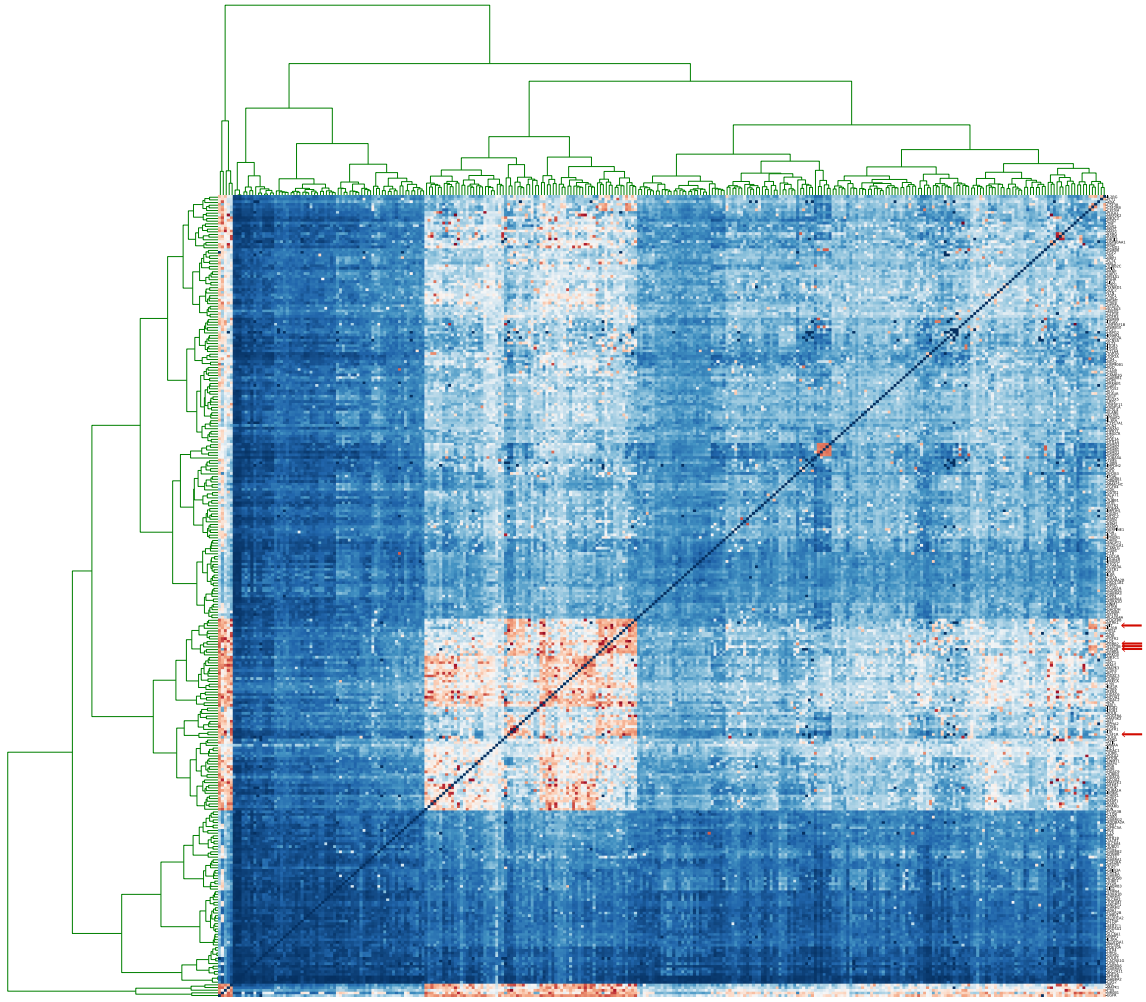


Figure 5.A.1: Cancer gene cluster
Cluster of cancer-associated genes from DGIdb. The distances of the five original genes of interest (red arrows) are significantly lower than the average distance (Mann-Whitney U, p=2.07e-7)

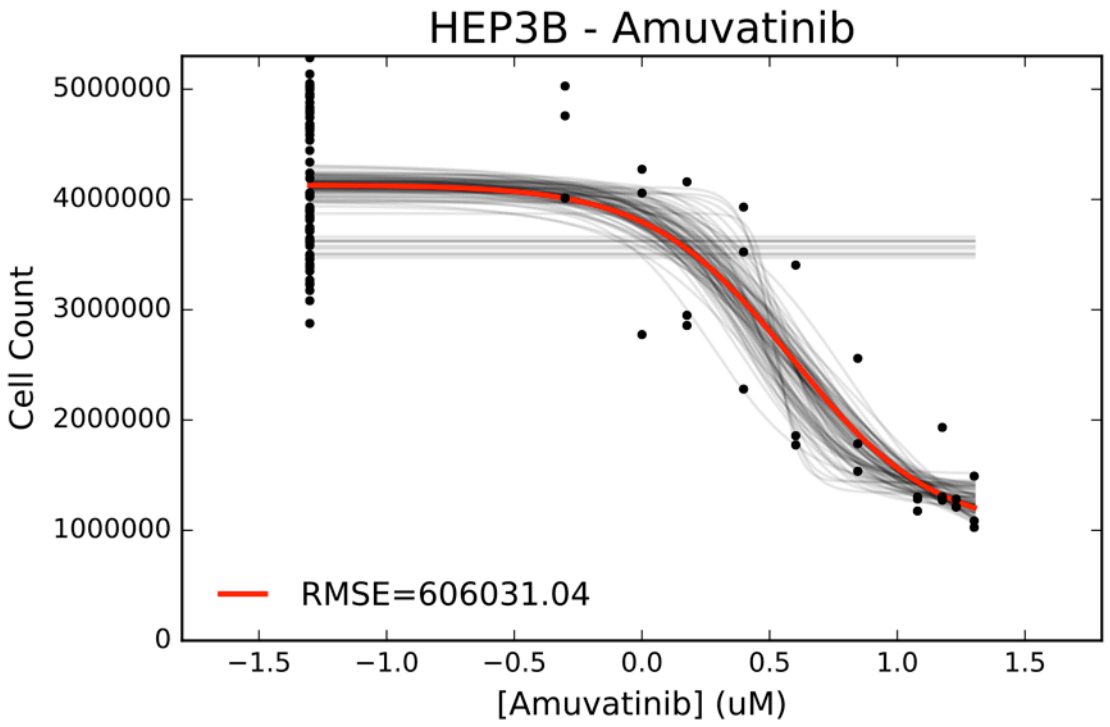


Figure 5.A.2: Curve-fitting example
In order to select the best curve fit, we ran 100 bootstraps of dose-response data (gray lines) and calculated the RMSE for the curve for the original data (black dots). We selected the curve with the lowest RMSE (red).

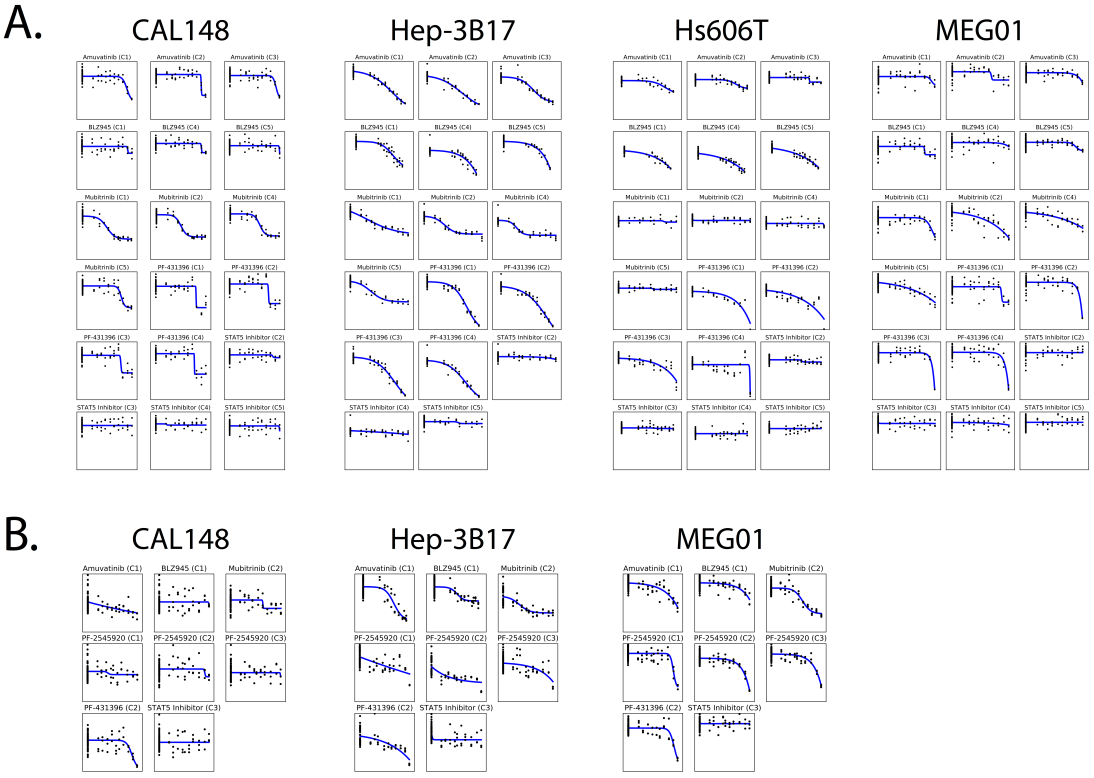


Figure 5.A.3: Dose-response curve fits
Curve fitting for A.) *Predicted SL combinations and* B.) *Predicted non-SL combinations.*

	Cell Line	Combination Set	Drug	Curve Start (Cell Ct)	Median	% Difference from Median
Predicted SL	CAL148	1	Amuvatinib	9.0E+06	9.0E+06	0.3%
			BLZ945	9.0E+06		0.4%
			Mubritinib	9.0E+06		0.4%
			PF-431396	9.0E+06		0.3%
		2	Amuvatinib	8.9E+06	8.9E+06	0.1%
			Mubritinib	8.9E+06		0.1%
			PF-431396	9.1E+06		1.6%
			STAT5 Inhibitor	8.9E+06		0.2%
		3	Amuvatinib	8.9E+06	9.0E+06	0.8%
			PF-431396	9.0E+06		0.5%
			STAT5 Inhibitor	9.0E+06		0.0%
		4	BLZ945	9.4E+06	9.4E+06	0.2%
			Mubritinib	9.3E+06		0.6%
			PF-431396	9.4E+06		0.2%
			STAT5 Inhibitor	9.4E+06		0.2%
		5	BLZ945	8.7E+06	8.7E+06	0.3%
			Mubritinib	8.7E+06		0.0%
			STAT5 Inhibitor	8.6E+06		0.2%
	HEP3B217	1	Amuvatinib	7.9E+06	7.9E+06	0.0%
			BLZ945	7.9E+06		0.0%
			Mubritinib	7.9E+06		0.1%
			PF-431396	7.9E+06		0.5%
		2	Amuvatinib	8.4E+06	8.4E+06	0.1%
			Mubritinib	8.5E+06		0.2%
			PF-431396	8.4E+06		0.1%
			STAT5 Inhibitor	8.4E+06		0.4%
		3	Amuvatinib	6.2E+06	6.2E+06	0.1%
			PF-431396	6.2E+06		0.1%
		4	BLZ945	6.2E+06	6.2E+06	0.0%
			Mubritinib	6.2E+06		0.1%
			PF-431396	6.2E+06		0.0%
			STAT5 Inhibitor	6.2E+06		0.1%
		5	BLZ945	6.2E+06	6.2E+06	0.1%
			Mubritinib	6.2E+06		0.0%
			STAT5 Inhibitor	6.2E+06		0.0%
	HS606T	1	Amuvatinib	1.5E+06	1.5E+06	0.5%
			BLZ945	1.5E+06		0.0%
			Mubritinib	1.5E+06		0.0%
			PF-431396	1.5E+06		1.1%
		2	Amuvatinib	1.6E+06	1.6E+06	0.1%
			Mubritinib	1.6E+06		0.1%
			PF-431396	1.6E+06		1.5%
			STAT5 Inhibitor	1.6E+06		0.1%
		3	Amuvatinib	2.2E+06	2.2E+06	0.1%
			PF-431396	2.2E+06		1.1%
			STAT5 Inhibitor	2.2E+06		0.0%
		4	BLZ945	1.9E+06	1.9E+06	0.1%
			Mubritinib	1.9E+06		0.1%
			PF-431396	1.8E+06		2.9%
			STAT5 Inhibitor	1.9E+06		0.1%
		5	BLZ945	2.1E+06	2.1E+06	0.0%
			Mubritinib	2.1E+06		0.2%
			STAT5 Inhibitor	2.1E+06		0.2%
	MEG01	1	Amuvatinib	8.9E+06	8.9E+06	0.0%
			BLZ945	8.9E+06		0.3%
			Mubritinib	8.7E+06		2.3%
			PF-431396	8.9E+06		0.0%
		2	Amuvatinib	1.2E+07	1.2E+07	0.8%
			Mubritinib	1.2E+07		0.0%
			PF-431396	1.2E+07		0.0%
			STAT5 Inhibitor	1.2E+07		0.6%
		3	Amuvatinib	1.2E+07	1.3E+07	0.2%
			PF-431396	1.3E+07		0.2%
			STAT5 Inhibitor	1.3E+07		0.0%
		4	BLZ945	1.1E+07	1.1E+07	0.6%
			Mubritinib	1.1E+07		0.0%
			PF-431396	1.1E+07		0.2%
			STAT5 Inhibitor	1.1E+07		0.0%
		5	BLZ945	1.1E+07	1.1E+07	0.0%
			Mubritinib	1.1E+07		0.1%
			STAT5 Inhibitor	1.1E+07		0.5%
Predicted non-SL	CAL148	1	Amuvatinib	2.1E+06	2.1E+06	0.0%
			BLZ945	2.1E+06		0.1%
			PF-2545920	2.1E+06		1.6%
		2	Mubritinib	2.0E+06	2.0E+06	0.0%
			PF-2545920	2.1E+06		2.1%
			PF-431396	2.0E+06		2.7%
		3	PF-2545920	1.2E+06	1.2E+06	0.0%
			STAT5 Inhibitor	1.2E+06		0.0%
	HEP3B	1	Amuvatinib	4.1E+06	4.1E+06	0.3%
			BLZ945	4.1E+06		0.0%
			PF-2545920	4.1E+06		0.9%
		2	Mubritinib	4.3E+06	4.3E+06	0.1%
			PF-2545920	4.3E+06		0.1%
			PF-431396	4.3E+06		0.0%
		3	PF-2545920	4.4E+06	4.4E+06	0.0%
			STAT5 Inhibitor	4.4E+06		0.0%
	MEG01	1	Amuvatinib	4.5E+06	4.5E+06	0.0%
			BLZ945	4.5E+06		0.3%
			PF-2545920	4.4E+06		1.5%
		2	Mubritinib	4.1E+06	4.0E+06	1.3%
			PF-2545920	4.0E+06		0.0%
			PF-431396	3.9E+06		4.2%
		3	PF-2545920	3.9E+06	3.9E+06	0.4%
			STAT5 Inhibitor	3.9E+06		0.4%

Table 5.A.1: Starting counts for drug curves

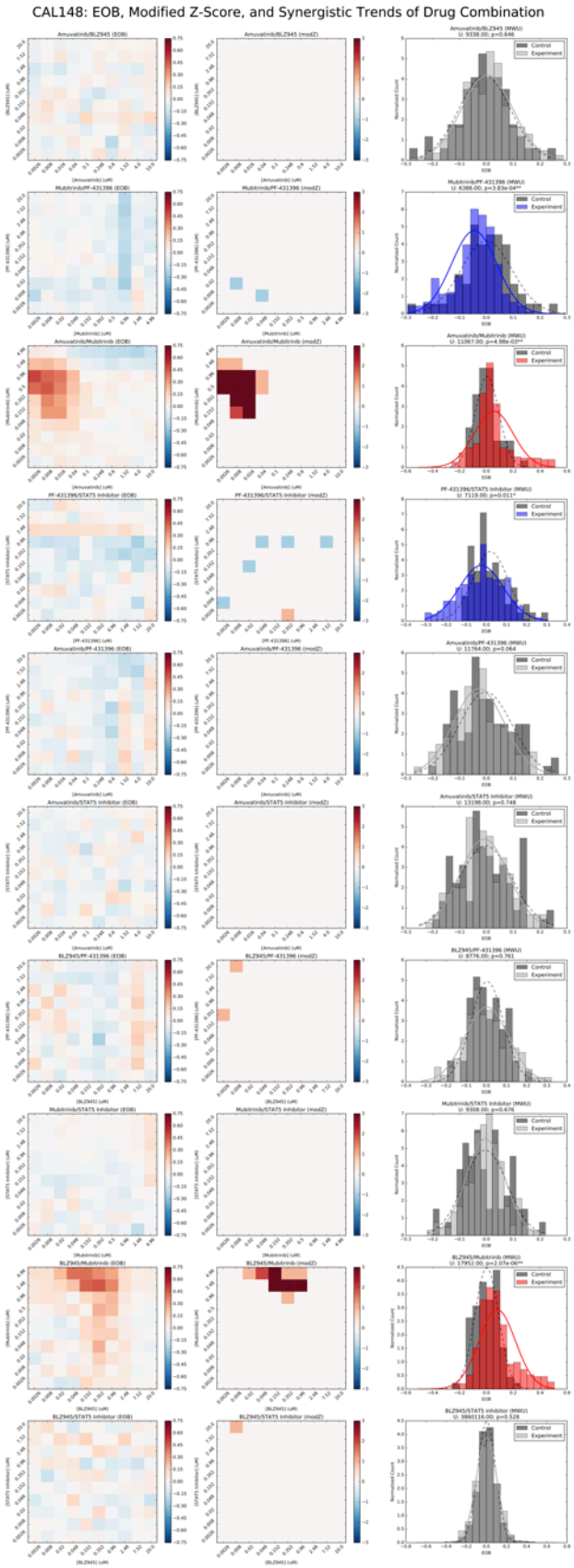


Figure 5.A.4: Putative SL pairs in CAL148
10 putative SL pairs in CAL148. We observe two consistently, significantly synergistic drug combinations: amuvatinib/mubritinib and BLZ945/mubritinib.

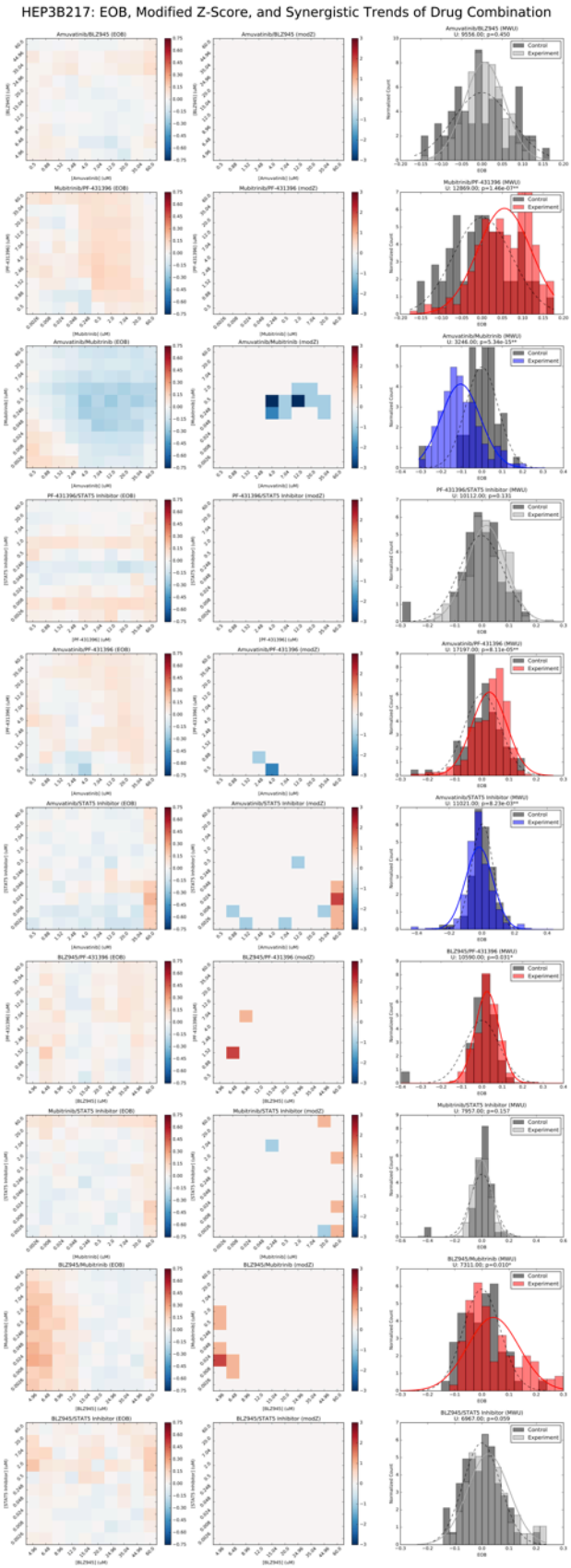


Figure 5.A.5: Putative SL pairs in Hep-3B217
10 putative SL pairs in Hep-3B217. We observe one consistently, significantly synergistic drug combination: mubritinib/PF-431396. We discard Amuvatinib/CAS 285986-31-4 (row 6) due to fault in experimental setup.

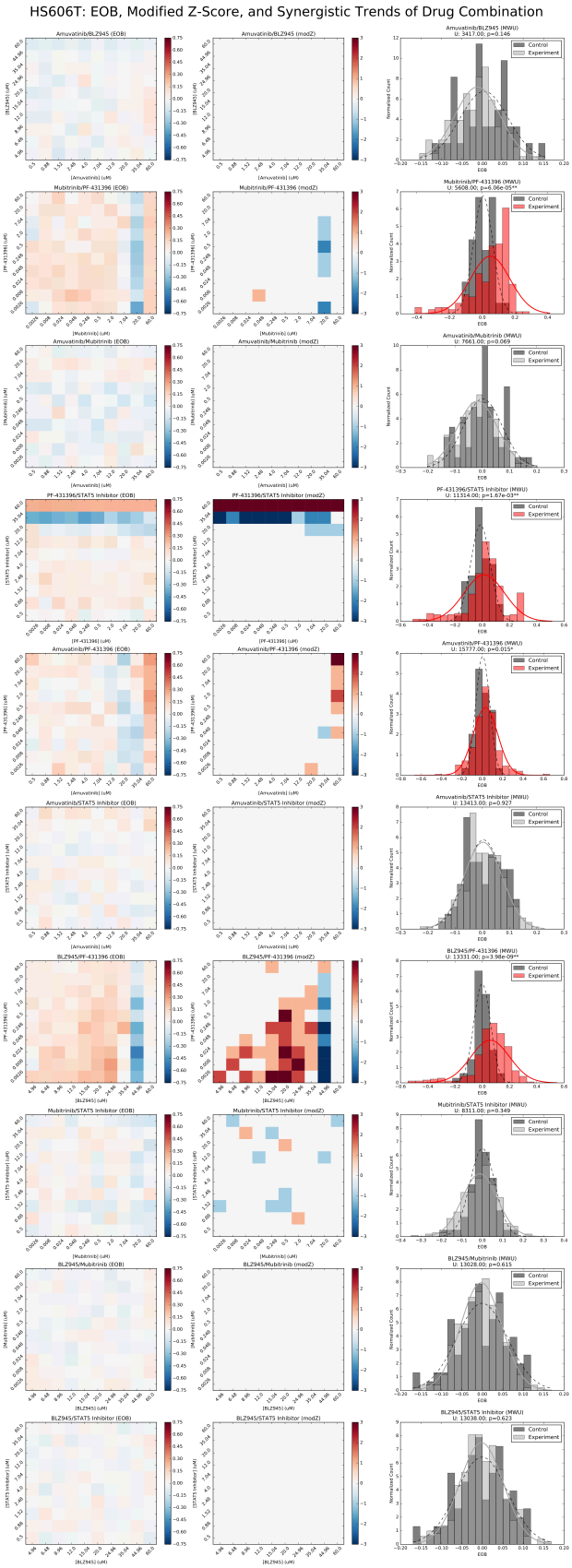


Figure 5.A.6: Putative SL pairs in Hs606T
10 putative SL pairs in Hs606T. We observe two consistently, significantly synergistic drug combinations: BLZ945/PF-431396 and BLZ945/mubritinib.

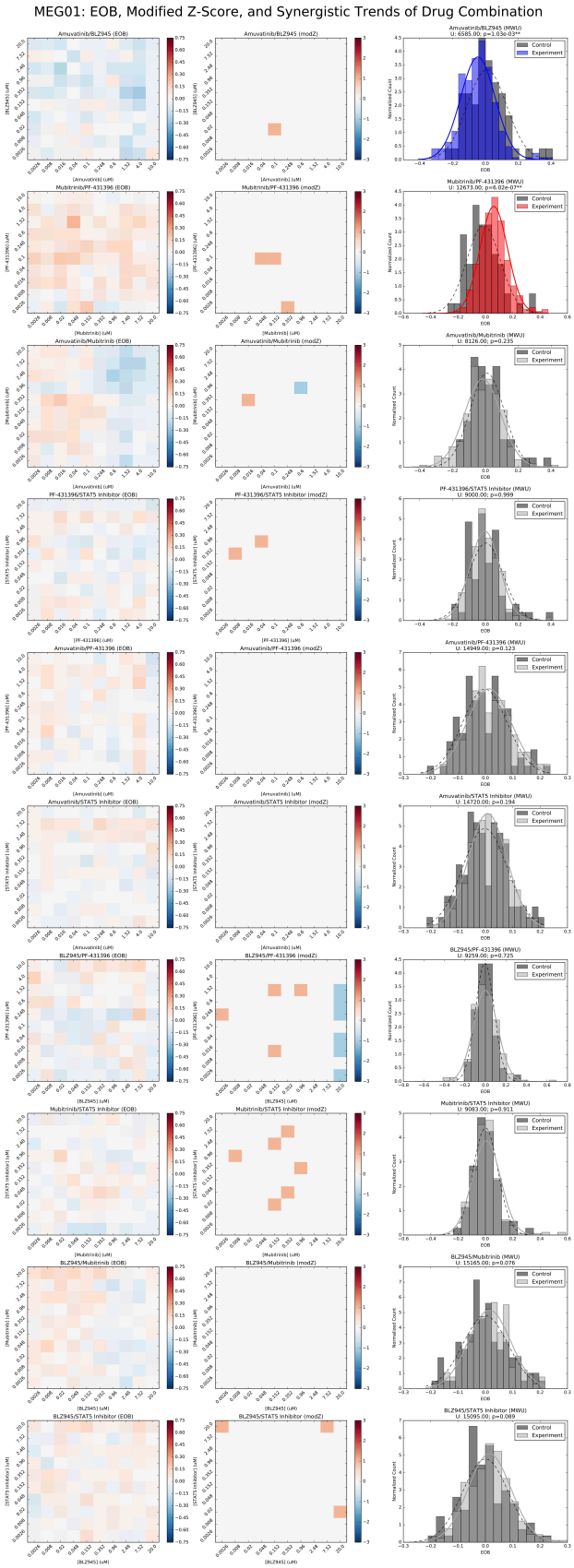


Figure 5.A.7: Putative SL pairs in MEG01
10 putative SL pairs in MEG01. We observe two consistently, significantly synergistic drug combination: amuvatinib/PF-431396 and BLZ945/PF-431396.

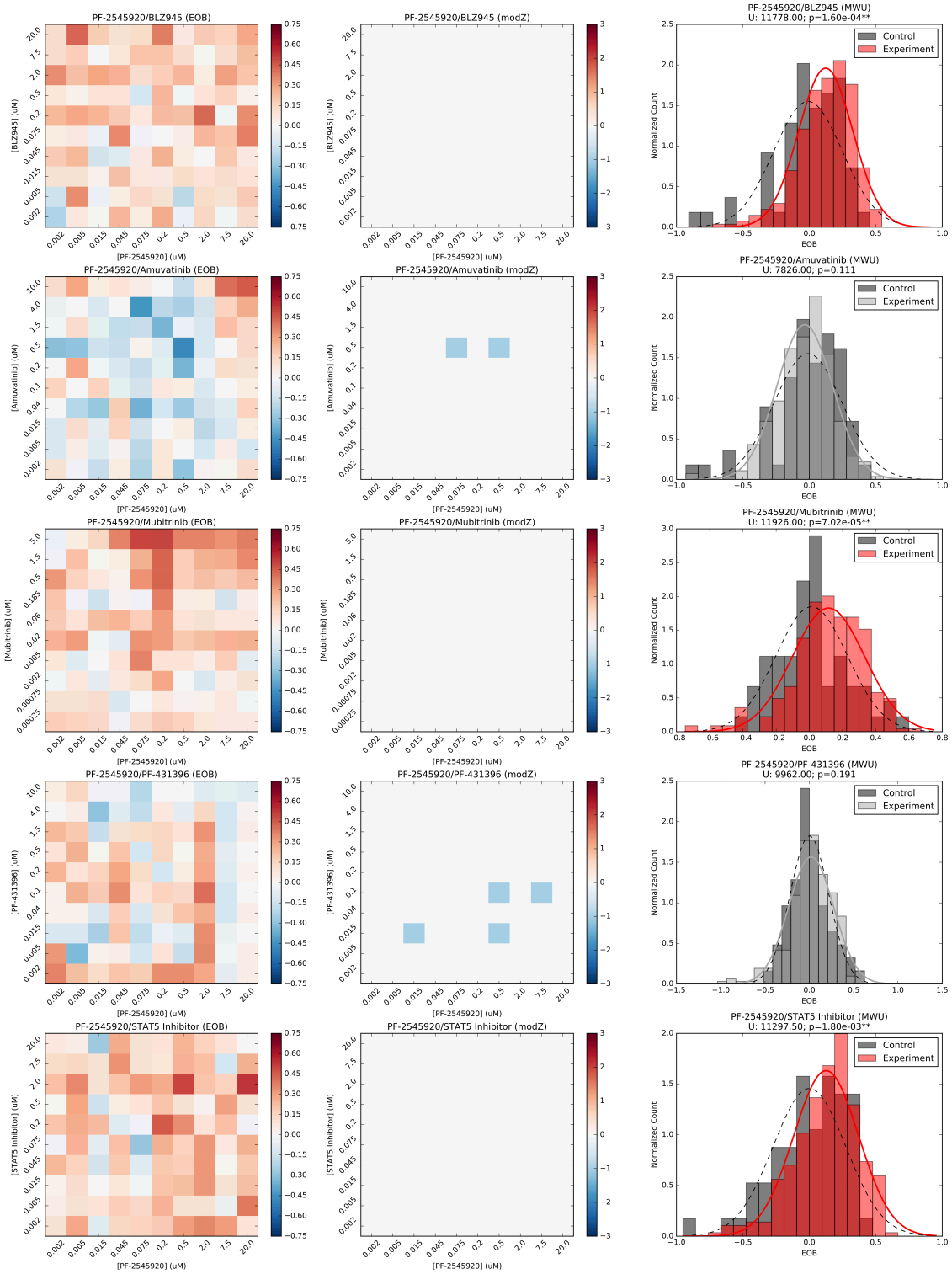


Figure 5.A.8: Putative non-SL pairs in CAL148

5 putative non-SL pairs in CAL148. We observe no consistently, significantly synergistic drug combinations.

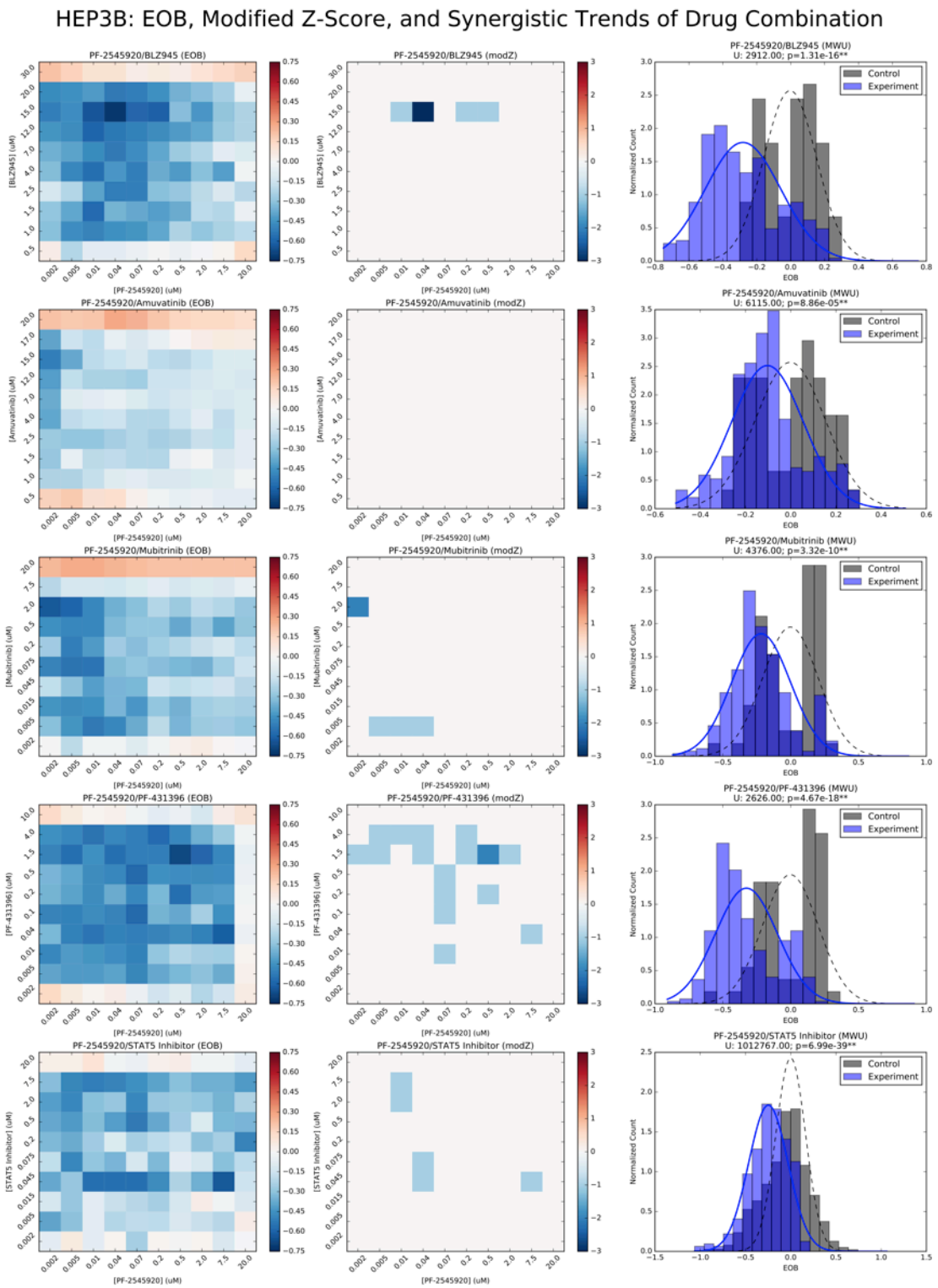


Figure 5.A.9: Putative non-SL pairs in Hep-3B217

5 putative non-SL pairs in Hep-3B217. We observe no consistently, significantly synergistic drug combinations.

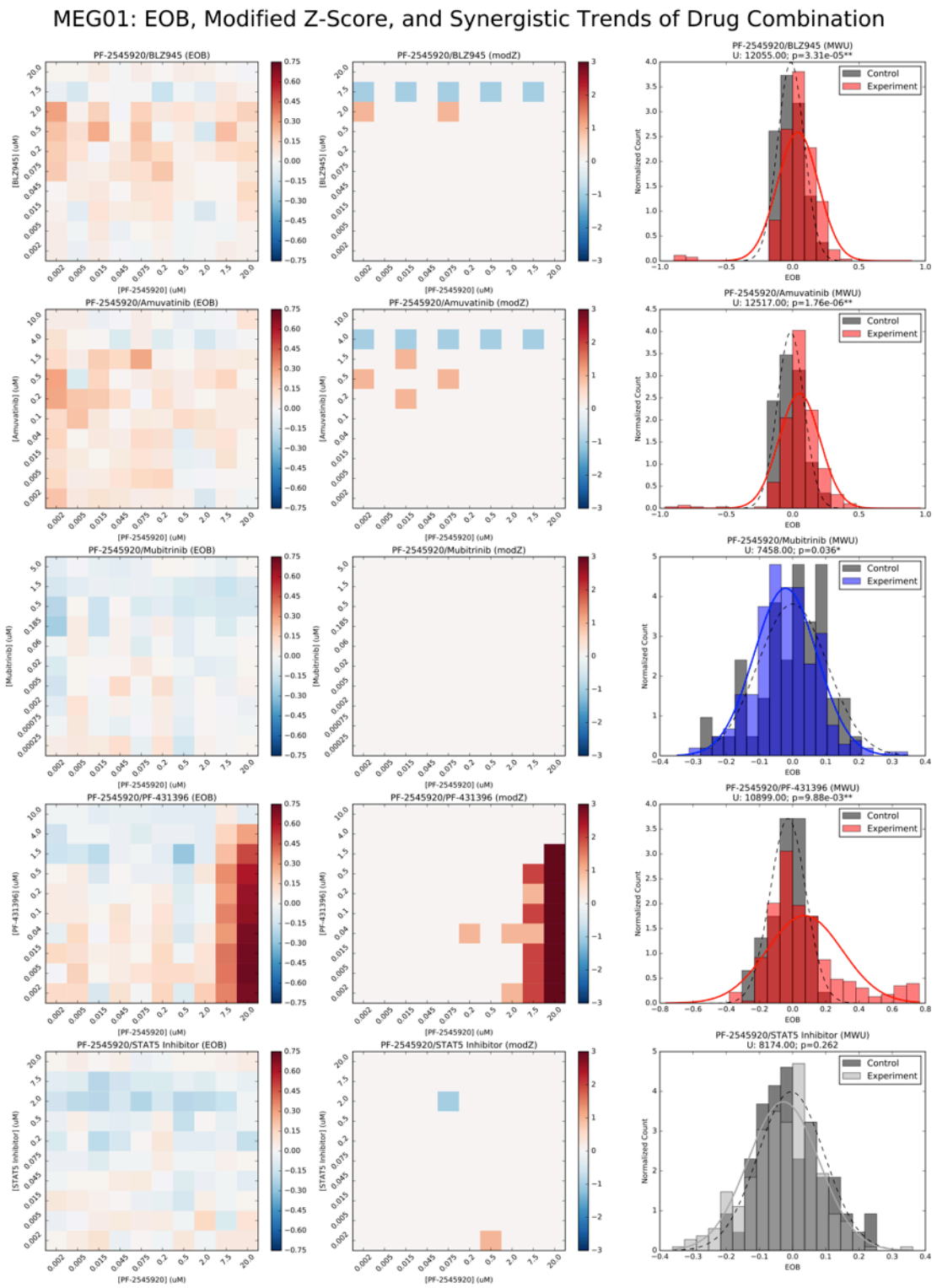


Figure 5.A.10: Putative non-SL pairs in MEG01

5 putative non-SL pairs in MEG01. We observe one consistently, significantly synergistic drug combinations: PF-2545920/PF-431396.

ACKNOWLEDGEMENTS

We thank the Columbia Genome Center (in particular, Ronald Realubit, Sergey Y Pampou, and Charles Karan) for conducting the experiments used in this chapter. Furthermore, we thank Hossein Khiabanian, Tal Lorberbaum, Rami Vanguri, and Kayla Quinnies for their careful reading of this work. Finally, as of the publication of this thesis (Summer 2017), the research in this chapter is under submission.

CHAPTER 6 – DISCUSSION AND CONCLUSIONS

Motivation

In this work, our overarching goal was to integrate work in systems biology, genetics, and pharmacology in order to explore the underlying properties of biological networks and to predict novel combination therapies to treat cancer in humans.

Summary

In Chapter 1, we introduced the concepts behind systems biology, and the use of networks in particular. We then explored the definition of synthetic lethality and its potential applications to the prediction of novel cancer combination therapies. We next outline the processes for elucidating drug synergy and the shortcomings of these methods. We synthesize these concepts to outline the questions addressed in this work: can we create interspecies models of synthetic lethality, and can the output of such

a model be used to successfully predict synergistic drug combinations in human cancer?

In Chapter 2, we outline the concept of connectivity homology, a novel measure of relatedness between genes based on protein-protein interaction networks that is independent of structure, function, or genetic homology. We first illustrate the concept using toy networks, then show that networks evolving using preferential attachment exhibit higher connectivity homology than random ones. Finally, we show that orthologous and non-orthologous genes of similar functions in S. cerevisiae, S. pombe, and human PPI networks exhibit significantly higher connectivity homology.

We use the concept of connectivity homology to demonstrate the viability of interspecies models of synthetic lethality in Chapter 3. We show that we can successfully predict synthetic

lethality from *S. cerevisiae* to *S. pombe* and M. musculus, and from S. pombe to S. cerevisiae using Species-INdependent TRAnslation (SINaTRA), a novel algorithm.

We applied SINaTRA to predict synthetic lethal gene pairs in humans using S. cerevisiae network data in Chapter 4. We found that, when cancer-therapy-associated genes were clustered together, several high-SINaTRA areas were enriched for known cancer combination therapies. This led us to hypothesize that SINaTRA may be a good way of finding novel cancer combination therapies that may exude their effect through a synthetic lethal mechanism.

We explored this hypothesis in Chapter 5, where we selected ten putative SL pairs and five putative non-SL pairs and tested them for synergy using specific drugs. We developed DAVISS (Data-driven Assessment of Variability In Synergy Scores), a new method of We found that 3/10 predicted SL pairs associated with significant, consistent drug synergy over four cell lines (Amuvatinib/PF-431396, BLZ945/PF-431396, BLZ945/Mubritinib), compared to 0/5 predicted non-SL pairs in three cell lines, which greatly exceeded the expected SL hit rate of 0.1% [77]. Furthermore, we found that putative SL pairs are enriched for synergy at specific concentrations compared to predicted non-SL pairs. Finally, we identified three novel, cell-specific drug combinations: Amuvatinib/Mubritinib and BLZ945/Mubritinib in CAL148, and BLZ945/PF-431396 in HS606T.

These results suggest that the underlying structures of biological networks can be leveraged to better understand human systems using model organisms. Furthermore, an interspecies, network-based model of synthetic lethality can help to identify novel synergistic drug pairs to treat human cancer.

Limitations

Although the results of each phase of our study are promising, they do have some limitations, and we will cover several of the key ones in this section.

First and foremost, our model of synthetic lethality is based on the protein-protein interaction network of *S. cerevisiae*, a monocellular organism. Although we addressed context-specific synthetic lethality in humans, we did not fully integrate expression data. This is in part because we have only a vague idea of how synthetic lethality changes between contexts, and no thorough study of the subject has yet occurred.

Next, our exploration of SINaTRA as a method to guide the discovery of novel synergistic drug pairs is rather small. We covered relatively few pairs in a small number of cell lines, and our analysis was limited to drugs alone. Although the drugs we used were fairly specific for our genes of interest, the possibility of off-target effects does exist. In an ideal, large-scale exploration of the subject, we would begin our analysis first with RNAi or another specific method of knocking down genes to show synthetic lethality, and then utilize drugs to show that drug synergy may be mediated through
a synthetic lethal mechanism.

Finally, although DAVISS has been shown to be useful and thorough in understanding drug synergy, it currently requires the creation of a full drug curve in order to assess synergy. Furthermore, it relies on the use of Bliss independence, which has its own limitations. In future work, expanding DAVISS to require fewer experiments, and developing it for other methods of testing drug synergy would make it a highly versatile and utile methodology.

Future directions

In our introduction, we mentioned the importance of a feedback loop in the development of systems biology. Therefore, the best method of further understanding and refining SINaTRA

would be to conduct as many experiments as possible to validate and update the model. This, in turn, would help us better understand the mechanisms and connectivity patterns associated with synthetic lethality.

AFTERWORD

The process of writing a thesis is a time for reflection, both on the work accomplished during the doctoral process, and on the relationships formed and nurtured during this time. It has made me take stock of everyone for whom I am grateful.

I thank my adviser, Dr. Nicholas Tatonetti, for the freedom and independence he gave me in my work, and his excitement towards my ideas (however small they seemed). I am also grateful to my thesis committee, including Dr. Raul Rabadan,
Dr. Saeed Tavazoie, Dr. Brent Stockwell, and Dr. Joel Dudley for their advice, patience, and generosity. I am indebted to my fellow TLab members for their comments on sections of this work and their help throughout my time at Columbia, and to the graduate students who have brightened my days (and especially my Fridays).

To my Weasels: my homes away from home. Thank you for the puns, the weirdness, and for being the worst SHG ever.

To Shelley and Kylie: my Bermuda Triangle and sisters from other misters. Thank you for fifteen years of friendship no distance could diminish.

To Hossein: my kahou and delbaram. Your laughter, love, and patience made this journey a significantly happier one ($p<2.2e-16$). Thank you for keeping me sane… and for teaching me to make tahdeeg.

And, above all, to my parents: for teaching me to aim high and put my nose to the grindstone. I owe every success I have to your love and support. I love you.

There are no endings, and never will be endings,
to the turning of the Wheel of Time.
But it was an *ending.*

REFERENCES

1. Lazebnik Y. Can a biologist fix a radio?—Or, what I learned while studying apoptosis. Cancer Cell. Elsevier; 2002;2: 179–182. doi:10.1016/S1535-6108(02)00133-2

2. Khetarpal P, Das S, Panigrahi I, Munshi A. Primordial dwarfism: overview of clinical and genetic aspects. Molecular Genetics and Genomics. Springer Berlin Heidelberg; 2015;291: 1–15. doi:10.1007/s00438-015-1110-y

3. Motulsky AG. Genetics of complex diseases. Journal of Zhejiang University SCIENCE B. Springer-Verlag; 2006;7: 167–168. doi:10.1631/jzus.2006.B0167

4. Sparrow DB, Chapman G, Smith AJ, Mattar MZ, Major JA, O'Reilly VC, et al. A Mechanism for Gene-Environment Interaction in the Etiology of Congenital Scoliosis. Cell. 2012;149: 295–306. doi:10.1016/j.cell.2012.02.054

5. Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. Science. American Association for the Advancement of Science; 2004;306: 640–643. doi:10.1126/science.1104635

6. Chen R, Snyder M. Systems biology: personalized medicine for the future? Current Opinion in Pharmacology. 2012;12: 623–628. doi:10.1016/j.coph.2012.07.011

7. Ciriello G, Cerami E, Koh SS, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Research. 2012;22: 398–406. doi:10.1101/gr.125567.111

8. Navratil V, de Chassey B, Rabourdin Combe C, Lotteau V. When the human viral infectome and diseasome networks collide: towards a systems biology platform for the aetiology of human diseases. BMC Systems Biology. BioMed Central; 2011;5: 13. doi:10.1186/1752-0509-5-13

9. Li X, Xia S, Bertisch HC, Branch CA, DeLisi LE. Unique topology of language processing brain network: A systems-level biomarker of schizophrenia. Schizophrenia Research. 2012;141: 128–136. doi:10.1016/j.schres.2012.07.026

10. Travers J, Milgram S. The small world problem. Phychology Today. 1967.

11. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems. 1998;30: 107–117. doi:10.1016/S0169-7552(98)00110-X

12. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature. Nature Publishing Group; 2000;407: 651–654. doi:10.1038/35036627

13. Minguez P, Dopazo J. Assessing the Biological Significance of Gene Expression

Signatures and Co-Expression Modules by Studying Their Network Properties. Falciani F, editor. PLoS ONE. Public Library of Science; 2011;6: e17474. doi:10.1371/journal.pone.0017474

14. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. Nature Publishing Group; 2011;12: 56–68. doi:doi:10.1038/nrg2918

15. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. 2013;: 1–6.

16. Lorberbaum T, Nasir M, Keiser MJ, Vilar S, Hripcsak G, Tatonetti NP. Systems Pharmacology Augments Drug Safety Surveillance. Clin Pharmacol Ther. 2014;97: 151–158. doi:10.1002/cpt.2

17. Bridges CB. The Origin of Variations in Sexual and Sex-Limited Characters: The American Naturalist: Vol 56, No 642. The American Naturalist. 1922. doi:10.2307/i323530;page:string:Article/Chapter

18. Dobzhansky T. Genetics of natural populations; recombination and variability in populations of Drosophila pseudoobscura. Genetics. Genetics Society of America; 1946;31: 269–290.

19. Ozier-Kalogeropoulos O, Adeline M-T, Yang W-L, Carman GM, Lacroute F. Use of synthetic lethal mutants to clone and characterize a novel CTP synthetase gene in Saccharomyces cerevisiae. Molec Gen Genet. Springer-Verlag; 1994;242: 431–439. doi:10.1007/BF00281793

20. Le Meur N, Gentleman R. Modeling synthetic lethality. Genome Biology. BioMed Central Ltd; 2008;9: R135. doi:10.1186/gb-2008-9-9-r135

21. Kaelin WG. The Concept of Synthetic Lethality in the Context of Anticancer Therapy. Nat Rev Cancer. 2005;5: 689–698. doi:10.1038/nrc1691

22. Nijman SMB. Synthetic lethality: General principles, utility and detection using genetic screens in human cells. FEBS Letters. 2011;585: 1–6. doi:10.1016/j.febslet.2010.11.024

23. Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N, et al. Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. Science. 2001;294: 2364–2368. doi:10.1126/science.1065810

24. Dixon SJ, Fedyshyn Y, Sonoda K, Prasad TSK, Chahwan C, Chua G, et al. Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. Proceedings of the National Academy of Sciences. 2008;105: 16653–16658. doi:10.1073/pnas.0806261105

25. Tarailo M, Tarailo S, Rose AM. Synthetic Lethal Interactions Identify Phenotypic “Interologs” of the Spindle Assembly Checkpoint Components. Genetics. Genetics

Society of America; 2007;177: 2525–2530. doi:10.1534/genetics.107.080408

26. Maia AF, Tanenbaum ME, Galli M, Lelieveld D, Egan DA, Gassmann R, et al. Genome-wide RNAi screen for synthetic lethal interactions with the C. elegans kinesin-5 homolog BMK-1. Sci Data. 2015;2: 150020. doi:10.1038/sdata.2015.20

27. Boettcher M, Lawson A, Ladenburger V, Fredebohm J, Wolf J, Hoheisel JD, et al. High throughput synthetic lethality screen reveals a tumorigenic role of adenylate cyclase in fumarate hydratase-deficient cancer cells. BMC Genomics. BioMed Central; 2014;15: 158. doi:10.1186/1471-2164-15-158

28. Drinnenberg IA, Weinberg DE, Xie KT, Mower JP, Wolfe KH, Fink GR, et al. RNAi in budding yeast. Science. American Association for the Advancement of Science; 2009;326: 544–550. doi:10.1126/science.1176945

29. Paladugu SR, Zhao S, Ray A, Raval A. Mining protein networks for synthetic genetic interactions. BMC Bioinformatics. BioMed Central Ltd; 2008;9: 426. doi:10.1186/1471-2105-9-426

30. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, et al. Global Mapping of the Yeast Genetic Interaction Network. Science. 2004;303: 808–813. doi:10.1126/science.1091317

31. Turner RM, Park BK, Pirmohamed M. Parsing interindividual drug variability: an emerging role for systems pharmacology. WIREs Syst Biol Med. 2015;7: 221–241. doi:10.1002/wsbm.1302

32. Jhamandas K, Sutak M. Morphine-naloxone interaction in the central cholinergic system: the influence of subcortical lesioning and electrical stimulation. British Journal of Pharmacology. Wiley-Blackwell; 1976;58: 101–107. doi:10.1111/(ISSN)1476-5381

33. Zhou L, Naraharisetti SB, Liu L, Wang H, Lin YS, Isoherranen N, et al. Contributions of Human Cytochrome P450 Enzymes to Glyburide Metabolism. Biopharmaceutics & drug disposition. NIH Public Access; 2010;31: 228–n/a. doi:10.1002/bdd.706

34. Roberge RJ, Kaplan R, Frank R, Fore C. Glyburide-ciprofloxacin interaction with resistant hypoglycemia. Annals of Emergency Medicine. 2000;36: 160–163. doi:10.1067/mem.2000.108617

35. Bleakley S. Identifying and reducing the risk of antipsychotic drug interactions. Progress in Neurology and Psychiatry. John Wiley & Sons, Ltd; 2012;16: 20–24. doi:10.1002/pnp.231

36. Michaels SH, Clark R, Kissinger P. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. N Engl J Med. Massachusetts Medical Society; 1998;339: 405–406. doi:10.1056/NEJM199808063390612

37. Bozic I, Reiter JG, Allen B, Antal T, Chatterjee K, Shah P, et al. Evolutionary dynamics

of cancer in response to targeted combination therapy. Elife. eLife Sciences Publications Limited; 2013;2: e00747. doi:10.7554/eLife.00747

38. Hunter P. The paradox of model organisms. The use of model organisms in research will continue despite their shortcomings. EMBO Rep. Nature Publishing Group; 2008;9: 717–720. doi:10.1038/embor.2008.142

39. Foucquier J, Guedj M. Analysis of drug combinations: current methodological landscape. Pharmacol Res Perspect. 2015;3: e00149. doi:10.1002/prp2.149

40. Tallarida RJ. An Overview of Drug Combination Analysis with Isobolograms. Journal of Pharmacology and Experimental Therapeutics. 2006;319: 1–7. doi:10.1124/jpet.106.104117

41. LOEWE S. The Problem of Synergism and Antagonism of Combined Drugs. Arzneimittelforschung. 1953;3: 285–290.

42. Nieuwenhuis S, Forstmann BU, Wagenmakers E-J. Erroneous analyses of interactions in neuroscience: a problem of significance. Nature Neuroscience. 2011;14: 1105–1107. doi:10.1038/nn.2886

43. Lehár J, Zimmermann GR, Krueger AS, Molnar RA, Ledell JT, Heilbut AM, et al. Chemical combination effects predict connectivity in biological systems. Mol Syst Biol. 2007;3: 80. doi:10.1038/msb4100116

44. Berenbaum MC. What is synergy? Pharmacological Reviews. 1989;41: 93–141. Available: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=2692037&retmode=ref&cmd=prlinks

45. Slinker BK. The statistics of synergism. J Mol Cell Cardiol. 1998;30: 723–731. doi:10.1006/jmcc.1998.0655

46. Bliss CI. The toxicity of poisons applied jointly. Annals of applied biology. 1939;26: 585–615. doi:10.1111/j.1744-7348.1939.tb06990.x

47. Jacunski A, Tatonetti NP. Connecting the dots: applications of network medicine in pharmacology and disease. Clin Pharmacol Ther. 2013;94: 659–669. doi:10.1038/clpt.2013.168

48. Boland MR, Jacunski A, Lorberbaum T, Romano JD, Moskovitch R, Tatonetti NP. Systems biology approaches for identifying adverse drug reactions and elucidating their underlying biological mechanisms. WIREs Syst Biol Med. 2016;8: 104–122. doi:10.1002/wsbm.1323

49. Barabási A-L, Albert R. Emergence of Scaling in Random Networks. Science. 1999;286: 509–512. doi:10.1126/science.286.5439.509

50. van Noort V, Snel B, Huynen MA. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep. European Molecular Biology Organization; 2004;5: 280–284. doi:10.1038/sj.embor.7400090

51. Mering von C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, et al. Comparative assessment of large-scale data sets of protein|[ndash]|protein interactions. Nature. Nature Publishing Group; 2002;417: 399–403. doi:10.1038/nature750

52. Deshpande R, Asiedu M, Klebig M, Sutor S, Kuzmin E, Nelson J, et al. A comparative genomic approach for identifying synthetic lethal interactions in human cancer. Cancer Research. 2013;73: 6128–6136. doi:10.1158/0008-5472.CAN-12-3956

53. Goh KI, Choi IG. Exploring the human diseasome: the human disease network. Briefings in Functional Genomics. 2012;11: 533–542. doi:10.1093/bfgp/els032

54. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. 2006;34: D535–D539. doi:10.1093/nar/gkj109

55. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25: 25–29. doi:10.1038/75556

56. Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Research. 2010;38: D331–5. doi:10.1093/nar/gkp1018

57. Ispolatov I, Krapivsky PL, Yuryev A. Duplication-divergence model of protein interaction network. Phys Rev E Stat Nonlin Soft Matter Phys. American Physical Society; 2005;71: 061911. doi:10.1103/PhysRevE.71.061911

58. Pastor-Satorras R, Smith E, Solé RV. Evolving protein interaction networks through gene duplication. J Theor Biol. 2003;222: 199–210.

59. Hughes AL, Friedman R. Parallel evolution by gene duplication in the genomes of two unicellular fungi. Genome Research. 2003;13: 794–799. doi:10.1101/gr.714603

60. Amoutzias GD, Robertson DL, Oliver SG, Bornberg-Bauer E. Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. EMBO Rep. EMBO Press; 2004;5: 274–279. doi:10.1038/sj.embor.7400096

61. Shannon P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003;13: 2498–2504. doi:10.1101/gr.1239303

62. Batagelj V, Brandes U. Efficient generation of large random networks. Phys Rev E Stat Nonlin Soft Matter Phys. American Physical Society; 2005;71: 036113. doi:10.1103/PhysRevE.71.036113

63. Holme P, Kim BJ. Growing scale-free networks with tunable clustering. Phys Rev E Stat

Nonlin Soft Matter Phys. American Physical Society; 2002;65: 026107. doi:10.1103/PhysRevE.65.026107

64. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Statist. Institute of Mathematical Statistics; 1947;18: 50–60. doi:10.1214/aoms/1177730491

65. Winter C, Kristiansen G, Kersting S, Roy J, Aust D, Knösel T, et al. Google Goes Cancer: Improving Outcome Prediction for Cancer Patients by Network-Based Ranking of Marker Genes. Slonim DK, editor. PLoS Comput Biol. Public Library of Science; 2012;8: e1002511. doi:10.1371/journal.pcbi.1002511.s010

66. Jacunski A, Dixon SJ, Tatonetti NP. Connectivity Homology Enables Inter-Species Network Models of Synthetic Lethality. Iakoucheva LM, editor. PLoS Comput Biol. Public Library of Science; 2015;11: e1004506. doi:10.1371/journal.pcbi.1004506

67. Whitehurst AW, Bodemann BO, Cardenas J, Ferguson D, Girard L, Peyton M, et al. Synthetic lethal screen identification of chemosensitizer loci in cancer cells. Nature. 2007;446: 815–819. doi:10.1038/nature05697

68. Conde-Pueyo N, Munteanu A, Solé RV, Rodríguez-Caso C. Human synthetic lethal inference as potential anti-cancer target gene detection. BMC Systems Biology. 2009;3: 116. doi:10.1186/1752-0509-3-116

69. Chan N, Pires IM, Bencokova Z, Coackley C, Luoto KR, Bhogal N, et al. Contextual Synthetic Lethality of Cancer Cell Kill Based on the Tumor Microenvironment. Cancer Research. 2010;70: 8045–8054. doi:10.1158/0008-5472.CAN-10-2352

70. Jerby-Arnon L, Pfetzer N, Waldman YY, McGarry L, James D, Shanks E, et al. Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell. 2014;158: 1199–1209. doi:10.1016/j.cell.2014.07.027

71. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2012;41: D8–D20. doi:10.1093/nar/gks1189

72. Sipiczki M. Where does fission yeast sit on the tree of life? Genome Biology. BioMed Central; 2000;1: REVIEWS1011. doi:10.1186/gb-2000-1-2-reviews1011

73. Wu M, Li X, Zhang F, Li X, Kwoh C-K, Zheng J. Meta-analysis of Genomic and Proteomic Features to Predict Synthetic Lethality of Yeast and Human Cancer. New York, New York, USA: ACM Press; 2013. pp. 384–391. doi:10.1145/2506583.2506653

74. Kranthi T, Rao SB, Manimaran P. Identification of synthetic lethal pairs in biological systems through network information centrality. Mol BioSyst. 2013;9: 2163. doi:10.1039/c3mb25589a

75. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein

domain families based on seed alignments. Proteins. 1997;28: 405–420.

76. Finn RD, Bateman A, Clements J, Coggill P. Pfam: the protein families database. Nucleic acids …. 2013.

77. Phillips PC, Johnson NA. The Population Genetics of Synthetic Lethals. Genetics Society of America. 1998;: 449–458.

78. Hart GT, Ramani AK, Marcotte EM. How complete are current yeast and human protein-interaction networks? Genome Biology. BioMed Central Ltd; 2006;7: 120. doi:10.1186/gb-2006-7-11-120

79. Conte LL, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. SCOP: a Structural Classification of Proteins database. Nucleic Acids Research. 2000;28: 257–259. doi:10.1093/nar/28.1.257

80. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Research. 2013;42: D304–D309. doi:10.1093/nar/gkt1240

81. Maglott D. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research. 2004;33: D54–D58. doi:10.1093/nar/gki031

82. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research. 2009;37: 1–13. doi:10.1093/nar/gkn923

83. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4: 44–57. doi:10.1038/nprot.2008.211

84. Amaratunga D, Cabrera J. Analysis of Data From Viral DNA Microchips. Journal of the American Statistical Association. 2001;96: 1161–1170. doi:10.1198/016214501753381814

85. Hagberg A, Schult D, Swart P. Exploring Network Structure, Dynamics, and Function using NetworkX. Varoquaux G, Vaught T, Millman J, editors. In: Proceedings of the Python in Science Conference (SciPy) [Internet]. 8 Nov 2008 [cited 11 Nov 2013] pp. 11–16. Available: http://conference.scipy.org/proceedings/SciPy2008/paper_2/

86. Breiman L. Random forests. Machine learning. Springer; 2001;45: 5–32.

87. Hill C. SciPy. Learning Scientific Programming with Python. Cambridge: Cambridge University Press; 2016. pp. 333–401. doi:10.1017/CBO9781139871754.008

88. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12: 77. doi:10.1186/1471-2105-12-77

89. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9: 90–95. doi:10.1109/MCSE.2007.55

90. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467: 1061–1073. doi:10.1038/nature09534

91. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491: 56–65. doi:10.1038/nature11632

92. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012;44: 623–630. doi:10.1038/ng.2303

93. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet. 2012;44: 631–635. doi:10.1038/ng.2283

94. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012;91: 597–607. doi:10.1016/j.ajhg.2012.08.005

95. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Research. 2004;32: D277–80. doi:10.1093/nar/gkh063

96. Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, et al. CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Research. 2007;36: D646–D650. doi:10.1093/nar/gkm936

97. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Research. 2009;38: D497–D501. doi:10.1093/nar/gkp914

98. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, et al. Remodeling of yeast genome expression in response to environmental changes. Molecular Biology of the Cell. American Society for Cell Biology; 2001;12: 323–337.

99. Pontén F, Jirström K, Uhlen M. The Human Protein Atlas--a tool for pathology. J Pathol. 2008;216: 387–393. doi:10.1002/path.2440

100. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28: 1248–1250. doi:10.1038/nbt1210-1248

101. Li X-J, Mishra SK, Wu M, Zhang F, Zheng J. Syn-lethality: an integrative knowledge base of synthetic lethality towards discovery of selective anticancer therapies. Biomed

Res Int. 2014;2014: 196034–7. doi:10.1155/2014/196034

102. Ryan CJ, Lord CJ, Ashworth A. DAISY: Picking Synthetic Lethals from Cancer Genomes. Cancer Cell. 2014.

103. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Research. 2014;42: D472–7. doi:10.1093/nar/gkt1102

104. Rains JL, Jain SK. Oxidative stress, insulin signaling, and diabetes. Free Radical Biology and Medicine. 2011;50: 567–575. doi:10.1016/j.freeradbiomed.2010.12.006

105. Liu Y, Hu B, Fu C, Chen X. DCDB: Drug combination database. Bioinformatics. 2010;26: 587–588. doi:10.1093/bioinformatics/btp697

106. Al-Lazikani B, Banerji U, Workman P. Combinatorial drug therapy for cancer in the post-genomic era. Nat Biotechnol. 2012;30: 679–692. doi:10.1038/nbt.2284

107. Maenza J, Flexner C. Combination antiretroviral therapy for HIV infection. Am Fam Physician. 1998;57: 2789–2798.

108. Frank J. Managing hypertension using combination therapy. Am Fam Physician. 2008;77: 1279–1286.

109. Garraway LA, Jänne PA. Circumventing cancer drug resistance in the era of personalized medicine. Cancer Discov. 2012;2: 214–226. doi:10.1158/2159-8290.CD-12-0012

110. Housman G, Byler S, Heerboth S, Lapinska K, Longacre M, Snyder N, et al. Drug resistance in cancer: an overview. Cancers. Multidisciplinary Digital Publishing Institute; 2014;6: 1769–1792. doi:10.3390/cancers6031769

111. Paul SM, Mytelka DS, Dunwiddie CT. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug …. 2010.

112. Brough R, Frankum JR, Costa-Cabral S, Lord CJ, Ashworth A. Searching for synthetic lethality in cancer. Current Opinion in Genetics & Development. 2011;21: 34–41. doi:10.1016/j.gde.2010.10.009

113. Yu D, Hung MC. Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. Oncogene. Nature Publishing Group; 2000;19: 6115–6121. doi:10.1038/sj.onc.1203972

114. Pittoni P, Piconese S, Tripodo C, Colombo MP. Tumor-intrinsic and -extrinsic roles of c-Kit: mast cells as the primary off-target of tyrosine kinase inhibitors. Oncogene. 2011;30: 757–769. doi:10.1038/onc.2010.494

115. Zhu Y, Knolhoff BL, Meyer MA, Nywening TM, West BL, Luo J, et al. CSF1/CSF1R Blockade Reprograms Tumor-Infiltrating Macrophages and Improves Response to T-cell Checkpoint Immunotherapy in Pancreatic Cancer Models. Cancer Research. American

Association for Cancer Research; 2014;74: 5057–5069. doi:10.1158/0008-5472.CAN-13-3723

116. Patsialou A, Wang Y, Pignatelli J, Chen X, Entenberg D, Oktay M, et al. Autocrine CSF1R signaling mediates switching between invasion and proliferation downstream of TGFβ in claudin-low breast tumor cells. Oncogene. Nature Publishing Group; 2015;34: 2721–2731. doi:10.1038/onc.2014.226

117. Wu X, Zahari MS, Renuse S, Nirujogi RS, Kim M-S, Manda SS, et al. Phosphoproteomic Analysis Identifies Focal Adhesion Kinase 2 (FAK2) as a Potential Therapeutic Target for Tamoxifen Resistance in Breast Cancer. Molecular & Cellular Proteomics. 2015;14: 2887–2900. doi:10.1074/mcp.M115.050484

118. Rajala HLM, Eldfors S, Kuusanmäki H, van Adrichem AJ, Olson T, Lagström S, et al. Discovery of somatic STAT5b mutations in large granular lymphocytic leukemia. Blood. American Society of Hematology; 2013;121: 4541–4550. doi:10.1182/blood-2012-12-474577

119. Müller J, Sperl B, Reindl W, Kiessling A, Berg T. Discovery of chromone-based inhibitors of the transcription factor STAT5. Chembiochem. WILEY- VCH Verlag; 2008;9: 723–727. doi:10.1002/cbic.200700701

120. Nagasawa J, Mizokami A, Koshida K, Yoshida S, Naito K, Namiki M. Novel HER2 selective tyrosine kinase inhibitor, TAK-165, inhibits bladder, kidney and androgen-independent prostate cancer in vitro and in vivo. Int J Urol. Blackwell Publishing Asia; 2006;13: 587–592. doi:10.1111/j.1442-2042.2006.01342.x

121. Pyonteck SM, Akkari L, Schuhmacher AJ, Bowman RL, Sevenich L, Quail DF, et al. CSF-1R inhibition alters macrophage polarization and blocks glioma progression. Nat Med. 2013;19: 1264–1272. doi:10.1038/nm.3337

122. Han S, Mistry A, Chang JS, Cunningham D, Griffor M, Bonnette PC, et al. Structural characterization of proline-rich tyrosine kinase 2 (PYK2) reveals a unique (DFG-out) conformation and enables inhibitor design. Journal of Biological Chemistry. American Society for Biochemistry and Molecular Biology; 2009;284: 13193–13201. doi:10.1074/jbc.M809038200

123. Griffith M, Griffith OL, Coffman AC, Weible JV, McMichael JF, Spies NC, et al. DGIdb: mining the druggable genome. Nat Meth. Nature Research; 2013;10: 1209–1210. doi:10.1038/nmeth.2689

124. Wagner AH, Coffman AC, Ainscough BJ, Spies NC, Skidmore ZL, Campbell KM, et al. DGIdb 2.0: mining clinically relevant drug–gene interactions. Nucleic Acids Research. Oxford University Press; 2016;44: D1036–D1044. doi:10.1093/nar/gkv1165

125. Verhoest PR, Chapin DS, Corman M, Fonseca K, Harms JF, Hou X, et al. Discovery of a Novel Class of Phosphodiesterase 10A Inhibitors and Identification of Clinical Candidate 2-[4-(1-Methyl-4-pyridin-4-yl-1 H-pyrazol-3-yl)-phenoxymethyl]-quinoline

(PF-2545920) for the Treatment of Schizophrenia †† Coordinates of the PDE10A crystal structures have been deposited in the Protein Data Bank for compound 1 (3HQW), 2 (3HQY), 3 (3HQW) and 9 (3HR1). J Med Chem. 2009;52: 5188–5196. doi:10.1021/jm900521k

126. Iglewicz B, Hoaglin DC. How to detect and handle outliers. ASQC basic references in quality control. Milwaukee; 1993.