Evaluating the correspondence between genetic interactions and gene order of action in *S. cerevisiae*

by

Yingzhou Wu

A thesis submitted in conformity with the requirements for the degree of Master of Science

Molecular Genetics

University of Toronto

© Copyright by Yingzhou Wu 2015
Evaluating the correspondence between genetic interactions and gene order of action in *S. cerevisiae*

Yingzhou Wu  
Master of Science  
Molecular Genetics  
2015

**Abstract**

Although a molecular function has been revealed for over 80% of *Saccharomyces cerevisiae* genes, the specific position of genes within ordered biological pathways remains elusive. Much of the information we have about connectivity and directionality of genes within pathways comes from genetic interaction analysis. Positive interactions, defined by double mutant phenotypes that are less extreme than expected, suggest that genes work in the same pathway or in series. Masking and suppression relationships are important subtypes of positive genetic interactions that can sometimes indicate gene order of action\textsuperscript{1,2,3,4,5,6,7,8}. Here I describe studies of the relationship between genetic interactions and functional gene order by extracting genetic interaction information of gene pairs with known order from large-scale Synthetic Genetic Array (SGA) data\textsuperscript{7} and small-scale DNA repair related genetic interaction data\textsuperscript{8}. This represents the first systematic evaluation of the applicability of quantitative epistasis analysis to reveal gene order.
# Table of Contents

Abstract ................................................................. ii

List of Figures ......................................................... v

List of Tables .......................................................... vi

1 Introduction ............................................................ 1
  1.1 Classical methods for gene order determination................................. 1
  1.2 Epistasis analysis ........................................................................ 2
  1.3 Quantitative epistasis analysis and genetic interaction.......................... 4
  1.4 Recent studies of gene functional order prediction.............................. 5
  1.5 Principle and overview of this study.................................................. 8

2 Methods and results ....................................................... 10
  2.1 Six different training data sets (ordered gene pairs collection)............... 10
  2.2 Six different training data sets (raw features and sample labeling)........... 12
  2.3 New feature: subtype of positive genetic interaction............................ 15
  2.4 New feature: Scaled Fitness............................................................. 16
  2.5 Mathematical model relating gene order to observed phenotypes............ 18
  2.6 Fixed “straw man” prediction models................................................. 23
  2.7 A data-driven “scaled fitness” model to predict gene order.................... 25
  2.8 Training data constraints.................................................................... 26
  2.9 Performance of the different models.................................................... 29
  2.10 Genetic interaction and the specificity of phenotypes............................ 37
  2.11 Expansion of DR_SENS dataset – DNA repair related genes collection......... 44
  2.12 Expansion of DR_SENS dataset – Gene pair selection............................ 46
3 Conclusion and discussion

3.1 Model-wise comparison: EUD is the overall best model ........................................50

3.2 Dataset constraint comparison: keep all gene pairs with positive GI in the training set. ..........55

3.3 Datasets comparison: epistatic gene is upstream is a general trend on substrate dependent pathways........................................................................................................................................57

3.4 Best combination: EUD + PG + DR_SENS ........................................................................59

3.5 Potentially confounding factors in gene order prediction: phenotype specificity, pleiotropy and circumvention of the upstream gene. .............................................................................................................59

3.6 Impact of dataset quality on gene order predictions ..............................................................60

4 Future direction

4.1 Computational future direction .............................................................................................61

4.2 Experimental future direction ................................................................................................62

5 Reference List

5.1 Reference List ........................................................................................................................62
List of Figures

Figure 1: The reciprocal shift method for inferring order in a substrate-dependent pathway..2

Figure 2: The missing circumvention of the theoretical model in Phenix et al (2010). ..........8

Figure 3: The subtypes of the positive genetic interaction..................................................16

Figure 4: The definition of the “scaled fitness” feature..........................................................18

Figure 5: The mathematical model..........................................................................................22

Figure 6: Default definition of an epistatic gene. .................................................................24

Figure 7: Alternative definition of an epistatic gene...............................................................24

Figure 8: The “Scaled Fitness” plot .......................................................................................26

Figure 9: Alternative definitions of precision and recall for gene order predictions..........30

Figure 10: Model-wise prediction performance comparison...............................................32

Figure 11: Dataset-wise prediction performance comparison............................................34

Figure 12: Constraint-wise prediction performance comparison.......................................36

Figure 13: Prediction performance on datasets DR NOMMS, DR MMS and DR SENS....40

Figure 14: Epsilon changes on three different datasets.......................................................41

Figure 15: Protein-protein interaction map in the space of the genes of interest.................48

Figure 16: Evaluating gene pair prioritization criteria using DR SENS dataset.................49

Figure 17: Model-wise prediction performance comparison for all gene pairs with genetic interaction information (AG constraint). .................................................................51
Figure 18: Model-wise prediction performance comparison for gene pairs with positive genetic interaction (PG constraint). .................................................................52

Figure 19: Model-wise prediction performance comparison for gene pairs with partial and complete epistatic to relationship (ET constraint). .......................................................53

Figure 20: Model-wise prediction performance comparison for gene pairs with complete epistatic to relationship (CET constraint). ...........................................................................54

Figure 21: Constraint-wise prediction performance comparison. .......................................56

Figure 22: Dataset-wise prediction performance comparison. ..............................................58

**List of Tables**

Table 1: Description of the six different training data sets.............................................12

Table 2: Four different fixed models..............................................................................25

Table 3: Number of training samples with different constraints ..................................28

Table 4: Number of training samples for different types of genetic interaction ............28

Table 5: Gene pairs predicted wrong in DR_NOMMS but correct in DR_MMS ..........42

Table 6: Gene pairs predicted wrong in DR_MMS but correct in DR_SENS .................42

Table 7: Gene pairs predicted wrong in DR_SENS dataset ............................................43

Table 8: List of DNA damaging drugs and their mechanism of action ......................45
1 Introduction

1.1 Classical methods for gene order determination

The characterization of the functional order of genes in biological pathways has been an important task as well as a major challenge in molecular genetics. Classically two methods have been used by geneticists: the “reciprocal shift” method and the double-mutant method known as “epistasis analysis” both of which have been used by Hereford and Hartwell (1974) to study the sequential gene function in the initiation of DNA synthesis of budding yeast. Temperature-sensitive mutants (cdc28, cdc4 and cdc7) that inhibited the cell cycle at different morphologically distinguishable stages were used to order their execution points relative to the block produced by the cell cycle inhibitor alpha factor using reciprocal shift method. The double mutant method was used to order the genes relative to each other. The principle of the reciprocal shift method is to infer gene order by observing the unique phenotypic pattern exhibited by a pair of gene (gene X and gene Y) with different order relationship when the experimental condition is reciprocallly shifted from one condition (permits gene X but restrict gene Y) to the other (restricts gene X but permits gene Y, Figure 1). The reciprocal shift method was first used by Jarvik & Botstein (1973) to determine the ordered events in phage assembly. In this study, temperature sensitive and cold sensitive mutants that block phage P22 morphogenesis were isolated, the phenotype of mutants with one low and one high temperature sensitive gene were observed before and after shifting the temperature from permissive to restrictive. This work successfully elucidated that the head and tail assembly were both dependent on phage DNA replication but were independent of each other. However the reciprocal shift method is not generally applicable, because it is not always an easy task to create both low and high temperature sensitive mutants that are
required for each gene. The more predominant epistasis analysis method requires only single and double mutants and an observable phenotype.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X is upstream of Y</td>
<td><img src="image" alt="Diagram" /></td>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Y is upstream of X</td>
<td><img src="image" alt="Diagram" /></td>
<td>+</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>X and Y are independent</td>
<td><img src="image" alt="Diagram" /></td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>X and Y are interdependent</td>
<td><img src="image" alt="Diagram" /></td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1:** The reciprocal shift method for inferring order in a substrate-dependent pathway. Four different possible relationships between gene X and gene Y are listed in column 1. Column 2 gives a graphical presentation for each relationship. The boxed P symbol represents the observed phenotype. Column 3 & 4 show different phenotypic patterns after reciprocal shift. The patterns shown, which assume a substrate-dependent pathway and complete loss of function of each allele in the non-permissive condition, are different and unique for each possible relationship.

**1.2 Epistasis analysis**

Epistasis analysis has been used to determine whether or not two genes are in the same biological pathway, and if so to infer their functional order. A gene X is defined as “epistatic to” the other gene Y when: 1) X and Y have a positive genetic interaction (the
phenotype is less severe than expected under a model of independent action), 2) the phenotypes of two single mutants differ from each other and 3) the double mutant produces the phenotype of the mutant gene X. Whether the epistatic gene is upstream or downstream depends on the type of biological pathway in which the gene pair is working. For epistasis analysis there are two types of biological pathway: a) substrate-dependent and b) switch regulatory. A substrate-dependent pathway is often biosynthetic and consists of a series of reactions that are each dependent on a source of substrate and a functional gene product for producing a final outcome. Straightforward rules have been described for inferring gene order for substrate-dependent pathways: 1) All mutants involved must be null alleles. 2) Upstream mutants are epistatic to downstream mutants (because the downstream reaction depends on the substrate produced by the upstream reaction). The study of eye color pigmentation in *Drosophila melanogaster* where the *white* gene is epistatic to and upstream of the *brown* and *scarlet* genes is a typical example of epistasis analysis in substrate-dependent pathway. The transportation of precursors of the brown and scarlet pigment depends on the product of the *white* gene, an ATP-binding cassette (ABC) transporter protein. A switch regulatory pathway consists of a series of genes or gene products that have either an “on” or “off” state. The activity of a switch regulatory pathway is usually governed by an upstream signal that stimulates the pathway and produces the downstream response. Avery and Wasserman (1992) summarized rules of epistasis analysis for switch regulatory pathways: (1) each single mutant strain must impact the trait when the signal is either ON or OFF, but not both. (2) If two single mutants impact the trait in opposite signal states and one is epistatic to the other, then the epistatic gene is downstream and is repressed by the upstream gene. (3) If two single mutants impact the trait in the same signal state and one is
epistatic to the other, then the epistatic gene is upstream and activates the downstream gene. Order determination for the genes (her-1, tra-1) in the sex determination pathway of Caenorhabditis elegans is an example of epistasis analysis where the single mutants have a phenotype in opposite signal states\(^4\). The signal is the X chromosome dosage (either XX or XO, which triggers hermaphrodite or male development respectively). The her-1 mutant showed a hermaphrodite phenotype under the signal state XO, which was expected to trigger male development, while the tra-1 mutant showed a male phenotype under a signal state XX that was expected to trigger hermaphrodite development. The double mutant showed exactly the same phenotype as the tra-1 single mutant, which implied that tra-1 is the epistatic gene and is downstream to and inhibited by the her-1 gene based on the rules summarized by Avery and Wasserman (1992). The study of the cell death cascade in C. elegans gives a counterexample where the upstream gene is epistatic to the downstream gene when the single mutants have a phenotype in the same signal state\(^4\). In this study ced-3 and ced-1 single mutants both had a phenotype in the same signal state (the cell death signal is ON), and the double mutant (ced-3, cde-1) produced the exactly same phenotype as the ced-3 single mutant. Therefore ced-3 is the epistatic gene and is inferred to be upstream of ced-1.

1.3 Quantitative epistasis analysis and genetic interaction

Classical epistasis analysis has typically been applied to categorical phenotypes. The color of the fly eyes in the eye color pigmentation study in Drosophila melanogaster, the sex type in the sex determination pathway study in Caenorhabditis elegans and the bud morphology in the sequential gene function study in the initiation of Saccharomyces cerevisiae DNA synthesis are all examples of such phenotypes. However, observation of quantitative phenotypes is also possible, and thanks to technological developments in
robotics and optical density readers, large-scale collections of quantitative phenotypes such as growth rate of yeast mutants have become possible, offering opportunity for further development of quantitative epistasis analysis. The term “genetic interaction” has been used interchangeably with “epistasis”, especially among population geneticists\textsuperscript{10,11}. In this study, however, quantitative epistasis is not identical to genetic interaction and the word “epistatic” is used in William Bateson’s now-century-old classical definition\textsuperscript{1}. Bateson coined the word “epistasis” and “hypostasis” to express the masking or suppression relationships between alleles of different genes and to distinguish from the terms “dominant” and “recessive” used to describe a similar relationship between alleles of the same gene. “Epistasis” under the classical definition describes interactions in which one mutant phenotype is masked or suppressed in the presence of another mutation in another locus. This classical definition corresponds to a subset of positive genetic interaction. By contrast, among population geneticists, the “epistasis” encompasses all positive and negative genetic interactions. Occurrences of term “epistasis” or “epistatic” for the rest of this document correspond to the classical definition.

1.4 Recent studies of gene functional order prediction

Epistasis analysis on categorical phenotypes has been successfully used to infer gene functional order in many biological pathways\textsuperscript{3,4}. Given the availability of an increasing amount of quantitative phenotypic data, the applicability of quantitative epistasis analysis to infer gene order is of great interest. Phenix et al (2011) developed a theoretical model to infer gene order within pathways responding to an external signal\textsuperscript{6}. The model used boolean variables to represent the ON/OFF state of the signal, gene deletion/no-deletion and the activation/inhibition relationship between genes, also used real number variables to present
the phenotypic contributions provided by each gene and in combination as well as directly from the upstream signal itself. Phenix et al (2011) claimed that the epistatic gene is upstream when the observed phenotype is entirely signal-dependent regardless the sign of the regulatory relationship (inhibition/activation) between genes. When the model was applied to infer the order of genes involved in the galactose utilization pathway in *S. cerevisiae*, it recaptured 80% of the known relationships without any false positives. Battle et al (2010) took a Bayesian network approach to infer gene order considering not only the pairwise epistasis but also the genetic interaction network of the target genes as a whole. For every gene pair, a score was assigned for each possible order relationship, with higher scores for more probable ordered relationship based on the rules summarized by Avery and Wasserman (1992). Then an overall score was computed for many candidate networks, such that networks corresponding better to the most probable ordering were scored more highly. This method was applied to a genetic interaction dataset for endoplasmic reticulum-related genes using the unfolded protein response as a quantitative phenotype. The result showed the Bayesian network model was able to reconstruct known pathways such as ER-associated protein degradation and N-linked glycosylation.

These studies were successful but not without limitations. First, it is not known if the above-described approaches are generally applicable, as each method was only applied to a single dataset. Specifically, Phenix et al (2011) applied the theoretical model only to glucose utilization-related genes and the Bayesian model in Battle et al (2010) was only applied to the ER unfolded protein response pathway. Second, the Phenix et al (2011) model is only suitable for gene pairs with complete epistasis: not only do the single mutant phenotypes need to be different and positively interacting, but the double mutant needs to exhibit a
phenotype that is essentially identical to one of the single mutants. A gene pair is said to have a partial epistasis relationship when there is a positive interaction such that the single mutants differ, but the double mutant phenotype is not identical to either single mutant. The complexity of real biological networks, replete with pleiotropic effects and genetic redundancy, and the limitations of real observations subject to measurement error, makes complete epistasis more the exception than the rule. The model of Phenix et al (2011) took the pleiotropic effect of the upstream gene into consideration but ignored the possible circumvention that a signal could regulate the downstream gene via an alternative pathway (in addition to regulation via the upstream gene), as there might be another gene that could also regulate the downstream gene when the signal is ON (Figure 2). This model is, therefore, unable to infer order from partial epistasis gene pairs observed frequently in real biological networks. Third, in Battle et al (2010), the rationale for interpreting partial epistatic relationships was unclear, in that they interpreted complete and partial masking (where the double mutant phenotype indicates a positive interaction, but yields a phenotype that is still more severe than the most severe single mutant) to have opposite pathway order. Finally, the pairwise score of Battle et al (2010) was assigned to each possible order relationship of a pair of genes based on the assumption that the epistatic gene is more likely to be downstream. If the epistatic gene is more likely to be upstream (as in substrate-dependent and positive regulatory pathways), all arrows will be reversed in the inferred pathway. Thus, in a real network which may contain substrate-dependent relationships as well as positive and negative regulatory relationships, the inference of gene order from epistatic-to relationships is dependent on context not considered by Battle et al (2010) or Phenix et al (2010).
Figure 2: The missing circumvention of the theoretical model in Phenix et al (2010). Figure 2A is the theoretical model from Phenix et al (2011). Figure 2B included the possible contribution of signal dependent regulation not via the upstream gene.

1.5 Principle and overview of this study

In order to evaluate the extent to which quantitative epistasis analysis (complete and partial epistasis) can be applied to gene order inference, and to study the potential of different models to generalize to different biological pathways, I measured the prediction performance of five different prediction models on six different types of dataset. Specifically, I studied four fixed models, where the epistatic gene is assumed to be upstream/downstream with varying definitions of epistasis. I also studied one data-driven model involving supervised learning. The datasets included three sets of varying types of biological processes, including transcription regulation, phosphorylation and metabolic, and three describing different quantitative phenotypes for gene pairs involved in the DNA homologous recombination repair pathway: a) fitness without methyl methanesulfonate (MMS), b) fitness with MMS, and c) sensitivity to MMS. Moreover four levels of training data constraint were applied to each dataset: a) All gene pairs with genetic interaction information available (AG), b) positive genetic interacting gene pairs only (PG), c) ‘epistatic to’ gene pairs (ET), including
those with both partial and complete epistasis, and d) ‘complete epistatic to’ gene pairs (CET). Therefore in total there were 120 prediction results, each corresponding to a specific training data set and model type (5 model types, each fit to 4 constraints for each of 6 input datasets). A prediction performance comparison was made at three levels: model-wise, constraint-wise and dataset-wise. The model-wise comparison showed that the epistatic gene is generally upstream in substrate dependent pathways, such as metabolic pathways and the DNA homologous recombination repair pathway, but not in switch regulatory pathways, like transcription regulation and kinase-substrate phosphorylation pathways. The constraint-wise comparison showed that, in general, except for gene pairs with no or negative genetic interaction, a relatively accurate prediction could be made on gene pairs with positive genetic interactions when the right prediction model is used. The dataset-wise comparison showed that more accurate prediction of the functional order of a gene pair could be achieved when a more specific phenotype was used to define genetic interaction between two genes. The best performance (precision: 81.8%, recall: 75%) was achieved by using “epistatic gene is upstream” model to predict the functional order of the gene pairs in the DNA homologous recombination repair pathway while drug sensitivity was being used as the quantitative phenotype. The high performance motivated me to further collect 371 DNA-repair-related gene pairs corresponding to known protein-protein interactions. A future plan is to measure genetic interaction of these gene pairs with respect to growth and drug sensitivity phenotypes, for further model evaluation and novel gene order prediction in DNA repair pathways.
2 Methods and results

2.1 Six different training data sets (ordered gene pairs collection)

The collection of known ordered gene pairs relies on the availability of genetic interaction data as only the gene pairs with quantitative genetic interaction information are useful to this study. Two genetic interaction studies have been used for the generation of these datasets. The first is a large-scale S. cerevisiae synthetic genetic array dataset (SGA dataset) composed of genetic interactions of 1711 query single mutant strains crossed to 3885 array single mutant strains. The second dataset is a small-scale study by St Onge et al 2007 (SO dataset) in which the genetic interactions between all possible pair-wise combinations of 26 genes that are sensitive to MMS were measured with and without MMS.

Six different datasets, each consisting of a number of gene pairs with known functional order and quantitative genetic interaction, were collected. These can be divided into two categories, each presenting a unique research perspective in this study: biological process-centric and quantitative phenotype-centric. The biological process-centric datasets are used for studying and comparing the correspondence between gene functional order and genetic interaction of gene pairs involved in different biological processes. Three datasets are collected in the biological process-centric category: transcription factors and binding targets gene pairs (TFT), kinase and substrate gene pairs (KS) and metabolic enzyme pairs (MP). Specifically for the TFT dataset 6142 transcription factor – binding target gene pairs with p-value <0.001 were extracted from the study of Harbison et al, of which 1472 gene pairs were finally chosen due to their availability in the SGA dataset. Similarly the gene pairs in the KS dataset were collected from Sharifpoor et al in which 884 kinase and substrate gene pairs
with p-value <0.05 were originally selected, and 194 of which were found in the SGA dataset\textsuperscript{13}. Finally, for the MP metabolic enzyme gene pairs, I wrote a script to retrieve the ordered pairs from the yeast pathways database downloaded from Saccharomyces Genome Database (SGD) since no comprehensive collection of ordered metabolic enzyme pairs exists in the literature. The script has collected 798 gene pairs and 233 of them were found in SGA dataset.

To better understand how correspondence between gene functional order and genetic interaction changes with different quantitative phenotypes, we also examined quantitative phenotype-centric datasets. The three datasets in this category were: gene pairs related to DNA repair with genetic interaction calculated with respect either to growth fitness with (DR\textsubscript{MMS}) or without (DR\textsubscript{NOMMS}) the DNA damaging agent methyl methanesulfonate (MMS), or with respect to sensitivity to MMS (DR\textsubscript{SENS}). The gene pairs for the DR\textsubscript{NOMMS}, DR\textsubscript{MMS} and DR\textsubscript{SENS} datasets were collected from the aforementioned SO dataset. The genetic interaction of 325 gene pairs in SO dataset were measured with and without MMS in the original study. Through literature mining I confirmed 32 gene pairs have curated functional order. Most of these pairs are from the DNA homologous recombination repair pathway. The detailed information about the number of collected known ordered gene pairs in each dataset and the source of the collection are summarized in Table 1.
Table 1: Description of the six different training data sets

<table>
<thead>
<tr>
<th>Datasets</th>
<th># of ordered gene pairs</th>
<th># of ordered gene pairs with GI</th>
<th>GI information source</th>
<th>Ordered gene pairs source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biological process centric</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFT</td>
<td>6142</td>
<td>1427</td>
<td>Costanzo et al 2010</td>
<td>Harbison et al 2004</td>
</tr>
<tr>
<td>KS</td>
<td>884</td>
<td>194</td>
<td>Costanzo et al 2010</td>
<td>Sharipoor et al 2011</td>
</tr>
<tr>
<td>MP</td>
<td>798</td>
<td>233</td>
<td>Costanzo et al 2010</td>
<td>SGD database</td>
</tr>
<tr>
<td><strong>Quantitative phenotype centric</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR_NOMMS</td>
<td>32</td>
<td>32</td>
<td>St Onge et al 2007</td>
<td>St Onge et al 2007</td>
</tr>
<tr>
<td>DR_MMS</td>
<td>32</td>
<td>32</td>
<td>St Onge et al 2007</td>
<td>St Onge et al 2007</td>
</tr>
<tr>
<td>DR_SENS</td>
<td>32</td>
<td>32</td>
<td>St Onge et al 2007</td>
<td>St Onge et al 2007</td>
</tr>
</tbody>
</table>

2.2 Six different training data sets (raw features and sample labeling)

To study the correspondence between gene functional order and genetic interaction, I decided to treat it as a classification problem that classifies each pair of gene to be “upstream” or “downstream” based on the model learned from genetic interaction information of the known ordered gene pairs. The first crucial step of a classification problem is to collect the raw features of the training samples and appropriately label the order of these gene pairs. The raw features of a known ordered gene pair (gene X and gene Y) include an estimated quantitative phenotype of both single mutants (Wx and Wy) and uncertainty in these estimates in terms of standard error of the mean or “s.e.m.” (Wx_sem, Wy_sem). Features also include the double mutant the quantitative phenotype (Wxy) and uncertainty (Wxy_sem), as well as the genetic interaction score and its significance (p-value). Wherever the source data didn’t report uncertainty in estimates or reported a different measure of uncertainty such as standard deviation, I calculated the standard error of the mean. The raw features for TFT, KS and MP are collected from SGA dataset. Each single mutant fitness and its uncertainty
from the SGA dataset were reported as bootstrapped sample mean and bootstrapped standard error of the mean from eighty independent screens, therefore I collected them directly. The double mutant fitness was also reported as the sample mean, however the uncertainty of the double mutant fitness was reported as the standard deviation (Wxy_sd) of four replicates instead of the standard error of mean (Wxy_sem), therefore the standard error of mean was calculated as the reported standard deviation divided by the square root of the sample size which was two in this case (Wxy_sem = Wxy_sd / 2). The reported genetic interaction score Epsilon and its p-value from the SGA dataset were also collected directly as raw features. For the datasets DR_NOMMS and DR_MMS both single and double mutant phenotype and uncertainties were calculated from the original raw data obtained from Dr. St Onge, an author of the SO dataset. For each single or double mutant, the phenotypes (the growth rate with or without MMS) of the technical replicates were averaged and treated as a single independent sample, the mean of the phenotype of all independent samples was recorded as the final phenotype of the mutant and its standard error of mean was calculated by the standard derivation of all independent samples divided by the square root of the number of independent samples. Finally, for the DR_SENS dataset, the mutant phenotypes (sensitivity to MMS) of both single and double mutants were calculated as the mutant growth rate with MMS divided by the growth rate of the same mutant without MMS. The uncertainty of the mutant phenotype was calculated using the delta method\textsuperscript{14}, which derives uncertainty of any arithmetic function of random variables by propagating the error of the input random variables. In this case of calculating drug sensitivity, the input random variables were the sample mean of the mutant growth rate with MMS and sample mean of the mutant growth rate without MMS, and the function was simply division. The genetic interaction score
epsilon and its uncertainty for each gene pair in all three DNA repair related datasets was also calculated by the delta method: the input three random variables are the sample mean of two single (Wx and Wy) and one double mutant (Wxy), and the formula for Epsilon is (Wxy-Wx×Wy) based on the multiplicative model of genetic interaction\textsuperscript{15}. The p-value of Epsilon was calculated by a one-location z-test to show the significance of a pair of gene having a positive or negative genetic interaction. In terms of labeling the training samples, I first defined gene X as the gene whose single mutant has a stronger phenotype to wild type and gene Y as the gene whose single mutant has a weaker phenotype. In this case the stronger/weaker phenotype is the lower/higher growth fitness of a yeast mutant strain. Each gene pair in all six training datasets is assigned to four possible categories: 1) X→Y, gene X is upstream of gene Y. 2) Y→X, gene X is downstream of gene Y 3) X-Y, gene X and gene Y are in the same complex or work as a cohesive unit. 4) X?Y, the order between gene X and gene Y is unknown. There are several ways to label the order of a pair of gene and each exhibits a different prior probability for the ordered labels (X→Y or Y→X). For example I could label a pair of genes as X→Y where gene X precedes gene Y alphabetically with respect to the gene name, the prior probability of X→Y by this definition is clearly different from the previous definition of X→Y where gene X mutant is a slower grower than gene Y. Different ways of training sample labeling will not affect the prediction model evaluation if the chosen prediction performance estimator is insensitive to the prior. Nevertheless, labeling gene order based on definitions of a gene mutant as a slower or faster grower is particularly useful for designing new predictive features, which I will discuss in later sections.
2.3 New feature: subtype of positive genetic interaction

One useful categorical feature for gene order prediction that can be easily derived from the raw features of each gene pair is the subtype of each positive interaction. One of the prerequisites of gene order prediction is that two genes are truly working together in the same pathway, which can be supported by various measures of functional similarity. The subtype of the positive interaction sometimes can provide information about the gene order of action\textsuperscript{4}. In classical epistasis analysis, the observed phenotype is often categorical rather than quantitative, such that it is straightforward to determine which gene is the epistatic gene. The gene whose mutant phenotype exactly resembles the double mutant but different from another single mutant is defined as the epistatic gene. However, in the case of quantitative phenotype, it is almost impossible to observe exactly the same phenotype between mutants. In this case, statistical tools such as the two-location test can help compare the similarity of two phenotypes and its significance. Eight subtypes of positive genetic interaction were defined based on the relative position of the double mutant fitness relative to the fitness of each single mutant (Figure 3). Whether the fitness of one mutant is significantly larger or smaller than that of another is determined by the P-value (cutoff: 0.05), calculated by applying the two-locations Z test using the mean of the fitness of mutant replicates and the standard error of the mean derived previously. The fitness of the two mutants is considered to be the same when they are not significantly different from each other (P > 0.05).
Figure 3: The subtypes of the positive genetic interaction.
The blue arrows represent the fitness of two single mutants and the red arrow represents the fitness of the observed double mutant. The eight different subtypes of positive genetic interaction are based on the relative numerical positions among the three fitness values.

2.4 New feature: Scaled Fitness

In order to graphically observe the correspondence between gene functional order and genetic interaction, I designed a new feature termed “Scaled Fitness” \((W'_{xy})\). Scaled Fitness places the phenotype of a double mutant on a common scale with respect to its relative distance to the phenotype of each single mutant \((\text{Figure 4})\). \(W'_{xy} = (W_{xy} - W_x) / (W_y - W_x)\) where \(W_x, W_y\) and \(W_{xy}\) represent the fitness of the slower single mutant \(X\), the fitness of the faster single mutant \(Y\) and the fitness of the double mutant \(XY\) respectively. In comparison to
the subtype of positive interaction feature, which classifies each gene pair with positive interaction into eight distinct categories, the Scaled Fitness presents positive interaction in a more continuous fashion. Scaled fitness features can be used to determine subtypes of positive interaction as follows:

1) \((Wx*Wy - Wx)/(Wy – Wx) < W’xy < 0 \rightarrow \) partial masking

2) \(W’xy = 0 \rightarrow \) complete masking

3) \(0 < W’xy < 1 \rightarrow \) partial suppression

4) \(W’xy = 1 \rightarrow \) complete suppression

5) \(W’xy > 1 \rightarrow \) super suppression

Scaled Fitness can be used not only to determine subtype of positive interaction, but also to estimate the strength of suppression and masking because of its quantitative and continuous nature.
Figure 4: The definition of the “scaled fitness” feature.

Wx and Wy are the fitness of two single mutants, Wxy is the fitness of the observed double mutant, and WxWy is the expected double mutant fitness under multiplicative model. W’ (xy) is the defined “scaled fitness”. W’ (xy) is equal to zero in cases of complete masking and equal to one in cases of complete suppression.

2.5 Mathematical model relating gene order to observed phenotypes

The mathematical model proposed by Phenix et al (2011) provided a theoretical solution for gene functional order determination. They claimed the epistatic gene is always upstream irrespective of the pathway architecture, when only the signal-dependent effect is considered. However their model cannot handle partial epistasis. Moreover, this model ignored the phenotype contribution from the upstream signal directly to the downstream gene independent of the upstream gene. In order to make gene functional order predictions on the
gene pairs with partial epistasis observed frequently in real biological pathways, I redesigned and expanded the mathematical model of Phenix et al (2011) taking possible circumvention into account (Figure 5). The new model used both real-valued and Boolean variables to form a list of linear equations. The real number variables are the quantitative phenotype contributions by the upstream signal and each gene alone and in combination, and the boolean variables are: variable j which represents the downstream or upstream relationship for a pair of genes, and variable k which represents whether the regulatory relationship between genes is negative or positive. Essentially the purpose of the model is to infer gene order and the sign of regulatory influence (by inferring j and k, respectively).

**Figure 5** illustrates the building blocks of the mathematical model including the real number and Boolean variables. The signal and two target genes each has two different states. In total we have eight linear equations to model the phenotype by summation of phenotypic contribution from different combinations of signal and gene states. For example P (s+a+b+) presents the phenotype of wild type with signal and P (s-a+b+) represents the phenotype of wild type without signal. The eight linear equations are as follows:

1. \[ P_{(s+a+b+)} = \delta_s + \delta_{sx} + \delta_{sy} + k*\delta_{sxy} + \delta_{ns(a+b+)} \]

2. \[ P_{(s-a+b+)} = (1-k)*\delta_{sxy} + \delta_{ns(a+b+)} \]

3. \[ P_{(s+a-b+)} = \delta_s + (1-j)*\delta_{sx} + j*\delta_{sy} + j*(1-k)*\delta_{sxy} + \delta_{ns(a-b+)} \]

4. \[ P_{(s-a-b+)} = j*(1-k)*\delta_{sxy} + \delta_{ns(a-b+)} \]

5. \[ P_{(s+a+b-)} = \delta_s + j*\delta_{sx} + (1-j)*\delta_{sy} + (1-j)*(1-k)*\delta_{sxy} + \delta_{ns(a+b-)} \]
6. \[ P_{(s-a+b-)} = (1-j)(1-k)\delta_{sxy} + \delta_{ns(a+b-)} \]

7. \[ P_{(s+a-b-)} = \delta_s + \delta_{ns(a-b-)} \]

8. \[ P_{(s-a-b-)} = \delta_{ns(a-b-)} \]

Then I subtracted the phenotype with signal (e.g. MMS) from the phenotype without signal (e.g. no MMS) to get the sensitivity of each single mutant (\( S_a \) and \( S_b \)), wild type (\( S_w \)) and double mutant (\( S_{ab} \)) to the signal. There are four equations left when the signal-independent contributions are removed:

1. \[ S_w = \delta_s + \delta_{sx} + \delta_{sy} + (2k-1)\delta_{sxy} \]
2. \[ S_a = \delta_s + (1-j)\delta_{sx} + j\delta_{sy} \]
3. \[ S_b = \delta_s + j\delta_{sx} + (1-j)\delta_{sy} \]
4. \[ S_{ab} = \delta_s \]

In an alternative definition of epistasis for quantitative phenotypes, we can define a gene to be “epistatic to” another when the phenotype of the double mutant is more similar to that gene’s single mutant than to the other gene’s single mutant. In this case, gene ‘a’ is epistatic to gene ‘b’ requires \( S_{ab} \) to be more similar to \( S_a \) than to \( S_b \), and this requirement can be stated as \[ | S_a - S_{ab} | < | S_b - S_{ab} | \] which can be expanded and is equivalent to the following formula:

\[ (2j-1)(\delta_{sx}^2 - \delta_{sy}^2) > 0 \]

This formula indicates that \( j \) equals to 1 (gene ‘a’ is upstream of gene ‘b’) when \[ |\delta_{sx}| > |\delta_{sy}| \]. Through this derivation, we summarize the theoretical rules to infer gene order when sensitivity to an upstream signal is used as phenotype in switch regulatory pathways as following:
1) All mutants involved must result in complete loss of function.

2) The sensitivity to the upstream signal of each single mutant needs to be different from each other and different from the wild type.

3) A gene is defined to be epistatic to another when sensitivity of the double mutant is more similar to the phenotype of that gene’s single mutant than the phenotype of the other gene’s single mutant.

4) The epistatic gene is upstream if the signal-dependent effect of the upstream gene ($\delta_{sx}$, independent of the downstream gene) is stronger than signal-dependent effect of the downstream gene ($\delta_{sy}$, independent of the upstream gene).
Figure 5: The mathematical model.

**Real number variables:** 1) $\delta_s$: the phenotype contribution from signal that is independent of genes X and Y. 2) $\delta_{sx}$: the phenotype contribution from signal through both gene X and Y. 3) $\delta_{sy}$: the phenotype contribution from signal through downstream gene but independent of the upstream gene. 4) $\delta_{ns}$: the phenotype contribution from signal through upstream gene but independent of the downstream gene. 5) $\delta_{ns}$: Signal-independent contribution the phenotype.

**Boolean variables:** 1) $j$: equals one when gene ‘a’ is upstream of gene ‘b’ and zero when gene ‘a’ is downstream of gene ‘b’. 2) $k$: equals one when positive regulatory between genes and equals zero when negative regulatory between genes.
2.6 Fixed “straw man” prediction models

Before delving into the different machine-learning models to make gene order predictions, I first evaluated several simple fixed models such as “the epistatic gene is always upstream” and “the epistatic gene is always downstream” using the aforementioned training datasets. There is no supervised learning involved and the training datasets are used only to assess prediction performance of these “straw man” models. The essential missing element of these models is a definition of which gene is the epistatic gene. It is straightforward to determine which gene is epistatic when the observed phenotype is categorical, however alternative choices exist when the phenotype is quantitative. In the case where the double mutant phenotype is in between of the phenotype of two single mutants (partial suppression), the mutant of an epistatic gene produces less severe phenotype under default definition than the mutant of the other gene (Figure 6), whereas under alternative definition the mutant of the epistatic gene produces the phenotype closer to the double mutant than the other gene as I have assumed in the mathematical model in section 4.5 (Figure 7). Four fixed models each present a unique combination of epistatic gene definition and gene order direction are summarized in Table 2.
Figure 6: Default definition of an epistatic gene.
The blue and green arrows are the fitness of two single mutants, and the green arrow is the fitness of the single mutant with the epistatic gene. The red arrow is the fitness of the double mutant.

Figure 7: Alternative definition of an epistatic gene.
The alternative definition of epistatic gene is only different in partial suppression cases. The fitness of the single mutant with the epistatic gene (green arrow) is closer to the fitness of double mutant (red arrow).
Table 2: Four different fixed models

<table>
<thead>
<tr>
<th>Model abbreviation</th>
<th>Description</th>
<th>Definition of epistatic gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUD</td>
<td>Epistatic gene is upstream</td>
<td>Default definition (<a href="#">Figure 6</a>)</td>
</tr>
<tr>
<td></td>
<td>Default definition of epistatic gene</td>
<td></td>
</tr>
<tr>
<td>EDD</td>
<td>Epistatic gene is downstream</td>
<td>Default definition (<a href="#">Figure 6</a>)</td>
</tr>
<tr>
<td></td>
<td>Default definition of epistatic gene</td>
<td></td>
</tr>
<tr>
<td>EUA</td>
<td>Epistatic gene is upstream</td>
<td>Alternative definition (<a href="#">Figure 7</a>)</td>
</tr>
<tr>
<td></td>
<td>Alternative definition of epistatic gene</td>
<td></td>
</tr>
<tr>
<td>EDA</td>
<td>Epistatic gene is downstream</td>
<td>Alternative definition (<a href="#">Figure 7</a>)</td>
</tr>
<tr>
<td></td>
<td>Alternative definition of epistatic gene</td>
<td></td>
</tr>
</tbody>
</table>

2.7 A data-driven “scaled fitness” model to predict gene order

I used the aforementioned “Scaled Fitness” feature to build a “Scaled Fitness” data driven model (SF model) to predict gene order. I estimated the fraction of gene pairs that are the \( X \rightarrow Y \) category in each moving window as a function of the “Scaled Fitness” value (Figure 8). Each training dataset provides a unique SF model of the correspondence between the gene functional order and genetic interaction in that dataset. Each SF model can provide probabilistic predictions of gene order a given type of dataset. For a new pair of gene \( X \) and gene \( Y \), the gene order can be predicted by seeking the corresponding probability of the scaled fitness of the new gene pair. The simplest use of the model is to infer \( X \rightarrow Y \) if the probability of \( X \rightarrow Y \) is bigger than 0.5, otherwise the order is \( Y \rightarrow X \). Other thresholds are possible if greater confidence is required, or if it is known that the prior probability of ordering differs from that of the training set.
Figure 8: The “Scaled Fitness” plot.

Scaled fitness is a rescaled representation of double mutant fitness defined to be zero at the least fit single mutant and one at the most-fit single mutant. **X axis:** Scaled fitness is defined as \( W'_{xy} = \frac{W_{xy} - W_x}{W_y - W_x} \), where \( W, W' \) and \( W_{xy} \) are fitness of single and double mutants and \( x \) and \( y \) are chosen such that \( W_y > W_x \). **Y axis:** Probability of \( X \rightarrow Y \) is defined by fraction of ordered gene pairs \( (X \rightarrow Y) \) where \( X \) is upstream of \( Y \). **Blue lines:** Represents gene pair has complete masking \( (W'_{xy}=0) \) or complete suppression \( (W'_{xy}=1) \). **Red Point:** Example of a data point on the plot that has a scaled fitness of 0.25 and 60% probability \( (X \rightarrow Y) \). The 0.25 scaled fitness of the red point comes from the median of scaled fitness of the points in a moving windows, in this case the window size is 10. Sixty percent probability \( (X \rightarrow Y) \) of the red point is calculated as the fraction of the samples in \( X \rightarrow Y \) category in the moving window, in this case there are 6 points in the window labeled as \( X \rightarrow Y \).

2.8 Training data constraints

The training datasets I assembled are gene pairs with known order and available genetic interaction information, but not all of them are useful training samples for order prediction, particularly when gene pairs have negative or no genetic interaction. It is reasonable to
believe that we could learn some information of gene order from “epistatic to” relationships like the classical epistasis analysis has successfully done, however not all the ordered genes show “epistatic to” relationships especially when the observed phenotype is irrelevant to the pathway in which the ordered gene pair work together. A pair of genes could show different genetic interactions when genetic interaction is calculated with respect to different phenotypes. Therefore to evaluate which gene pairs are useful in training, four different constraints on the training data set have been used in this study: 1) All gene pairs with available genetic interaction information (AG), 2) Gene pairs with positive genetic interaction only (PG), 3) Gene pairs with either partial or complete ‘epistatic to’ relationships (ET), and 4) Gene pairs only with complete ‘epistatic to’ relationships (CET). The number of training samples for each dataset with different constraints are listed in Table 3 and the number of training samples for each different type of genetic interaction and subtype of positive interaction are listed in Table 4. The constraint column in Table 4 indicates which constraint filters allow this type of genetic interaction to be present.
### Table 3: Number of training samples with different constraints

<table>
<thead>
<tr>
<th>Constraint levels</th>
<th>TFT</th>
<th>KS</th>
<th>MP</th>
<th>DR_NOMMS</th>
<th>DR_MMS</th>
<th>DR_SENS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) All genetic interactions (AG)</td>
<td>1427</td>
<td>117</td>
<td>233</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>2) Positive interactions only (PG)</td>
<td>80</td>
<td>15</td>
<td>33</td>
<td>12</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>3) Epistatic to only (ET)</td>
<td>72</td>
<td>14</td>
<td>22</td>
<td>6</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>4) Complete epistatic to only (CET)</td>
<td>33</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

### Table 4: Number of training samples for different types of genetic interaction

<table>
<thead>
<tr>
<th>Type of genetic interaction</th>
<th>Constraints (Table 3) allowing this type</th>
<th>TFT</th>
<th>KS</th>
<th>MP</th>
<th>DR_NOMMS</th>
<th>DR_MMS</th>
<th>DR_SENS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Interaction</td>
<td></td>
<td>1202</td>
<td>90</td>
<td>181</td>
<td>16</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Negative interaction</td>
<td></td>
<td>145</td>
<td>12</td>
<td>19</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Positive interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudo Coequal</td>
<td>1, 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Partial Coequal</td>
<td>1, 2</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Coequal</td>
<td>1, 2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Partial Masking</td>
<td>1, 2, 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Complete masking</td>
<td>1, 2, 3, 4</td>
<td>21</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Partial suppression</td>
<td>1, 2, 3</td>
<td>18</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Complete suppression</td>
<td>1, 2, 3, 4</td>
<td>12</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Super suppression</td>
<td>1, 2, 3</td>
<td>20</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Positive interaction Total</td>
<td></td>
<td>80</td>
<td>15</td>
<td>33</td>
<td>12</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1427</td>
<td>117</td>
<td>233</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
</tbody>
</table>
2.9 Performance of the different models

To evaluate how well the fixed and data driven models predict gene order I calculated different measures of precision and recall performance. There are in total three possible classes (X→Y, Y→X, X–Y) in the training datasets and four possible classes the prediction model could generate ((X→Y, Y→X, X–Y, X?Y). The precision here is defined as the ratio of how many correct predictions on the space of all predictions were made as X→Y, Y→X and X–Y, the recall is defined as the ratio of how many correct predictions made on the space of all training samples includes the predictions made as X?Y. Figure 9 illustrates the definition of precision and recall using a confusion matrix. As mentioned in previous sections, each of the six different datasets filtered by four different levels of constraint have been collected and five fixed and data driven models have been proposed. Therefore, in total there are 120 (5 models, 6 datasets and 4 levels of constraint) prediction performance results. These results can be viewed from different angles: dataset-wise, model-wise and constraint level wise. For the purpose of seeking which model performs the best, model-wise performance comparison across all possible combinations of datasets and levels of constraint are plotted in Figure 10. To illustrate how the prediction performance changes among different datasets, especially among those datasets with different quantitative phenotype, the dataset-wise performance comparison across all possible combinations of models and levels of constraint can be found in Figure 11. Finally the constraint-wise comparison across all possible combinations of datasets and models plotted in Figure 12 are used to evaluate whether different constraint levels may change the prediction performance.
**Figure 9:** Alternative definitions of precision and recall for gene order predictions.

**Precision:** the fraction of the correct predicted pairs in the space of gene pairs predicted as X → Y, Y → X and X − Y. **Recall:** the fraction of the correct predicted pairs in the space of all predicted gene pairs.
Figure 10: Model-wise prediction performance comparison. The prediction performance comparison was made on five different models for all possible combinations of six datasets and four levels of constraint. A) DR_NOMMS, DR_MMS, DRSENS datasets, and B) TFT, KS, MP datasets.
A)

dataset ▲ DR_NOMMS ● DR_MMS ■ DR_SENS ★ TFT + KS ☒ MP

Model: EDD
Constraint: CET

Model: EDA
Constraint: CET

Model: EDD
Constraint: ET

Model: EDA
Constraint: ET

Model: EDD
Constraint: PG

Model: EDA
Constraint: PG

Model: EDD
Constraint: AG

Model: EDA
Constraint: AG
Figure 11: Dataset-wise prediction performance comparison. The prediction performance (precision/recall) comparison made on six different datasets for all possible combinations of five models and four levels of constraint. A) EDD and EDA model, and B) EUD, EUA and SF models.
A) constraint ● AG + PG ■ ET ▲ CET

Dataset: DR NOMMS
Model: SF

Dataset: DR MMS
Model: SF

Dataset: DR SENS
Model: SF

Dataset: DR NOMMS
Model: EUA

Dataset: DR MMS
Model: EUA

Dataset: DR SENS
Model: EUA

Dataset: DR NOMMS
Model: EUD

Dataset: DR MMS
Model: EUD

Dataset: DR SENS
Model: EUD

Dataset: DR NOMMS
Model: EDA

Dataset: DR MMS
Model: EDA

Dataset: DR SENS
Model: EDA

Dataset: DR NOMMS
Model: EDD

Dataset: DR MMS
Model: EDD

Dataset: DR SENS
Model: EDD
Figure 12: Constraint-wise prediction performance comparison. The prediction performance (precision/recall) comparison made on four different levels of constraint for all possible combinations of five models and six datasets. A) DR_NOMMS, DR_MMS and DR_SENS datasets, and B) TFT, KS, MP datasets.
2.10 Genetic interaction and the specificity of phenotypes

The EUD model (epistatic gene is upstream with default definition of epistatic gene) was applied to the PG set (all positive interactions) of DNA repair genes (Figure 13). Genetic interaction was measured with respect to different phenotypes (DR_NOMMS, DR_MMS and DR_SENS). The prediction performance using MMS sensitivity (DR_SENS; precision: 81.8%, recall: 75%) was better than that using growth with MMS (DR_MMS; precision: 63.2%, recall: 53.3%) or without MMS (DR_NOMMS; precision: 57.1%, recall: 33.3%). Since the three datasets have the same set of gene pairs, the change of prediction performance is attributable to the change of the quantitative phenotype. Notably, these phenotypes differ in specificity, in that impairment of many gene functions could inhibit growth (with or without MMS), but impairment of fewer functions will yield growth changes that are specific to the MMS condition. The NOMMS phenotype (growth fitness with no MMS) is less specific than the MMS phenotype (growth fitness with MMS) because without applying MMS some mutants may not show phenotype even they are sensitive to MMS, for example the null mutants created by knocking out the genes in SHU complex (shu1, shu2, csm2, and psy3) show no phenotype without MMS but show a clear phenotype with MMS. The SENS phenotype (sensitivity to MMS) is even more specific than the MMS phenotype as the calculation of the sensitivity (MMS phenotype divided by NOMSS phenotype) eliminated the phenotypic effect unrelated to the MMS perturbation. The magnitude and even sign of genetic interaction for most gene pairs are changing towards more positive from DR_NOMMS to DR_MMS as well as from DR_MMS to DR_SENS, which probably explains the increasing prediction performance with more specific phenotypes. Figure 14 illustrates how the genetic interaction score epsilon changes for every gene pair with respect
to the change of the quantitative phenotype from DR_NOMMS to DR_MMS and from DR_MMS to DR_SENS.

In total there were eight gene pairs that were correctly predicted by DR_MMS dataset but incorrectly predicted by DR_NOMMS dataset (Figure 14, Table 5). Of these, four pairs involve genes in SHU complex that show no phenotype without MMS but clear phenotype with MMS yielding observed positive genetic interactions (co-equal) under MMS and the correct predictions. The remaining four out of eight gene pairs show either partial or complete masking relationship with MMS but show no genetic interaction without MMS. Although the genes in these four pairs have a phenotype without MMS (RAD51, RAD52, RAD55, RAD57, SGS1, MMS4), the phenotypes are not strong and in some cases (e.g., SGS1) may be due to functions other than DNA repair. The strong and DNA-repair-related phenotype of these genes under MMS made the correct predictions possible. Another seven gene pairs have been predicted correctly in the DR_SENS dataset but not in the DR_MMS or DR_NOMMS datasets (Table 6). Using MMS sensitivity as the quantitative phenotype ‘normalized out’ phenotypic effects not induced by MMS, which reduces the possible confounding pleiotropic effect for gene order inference.

Even using the more specific drug sensitivity phenotype fourteen gene pairs are still predicted incorrectly: most of these fourteen involve only three genes (MMS4, MUS81, and SGS1) working in the late stage of homologous recombination (HR) DNA repair pathway resolving the double Holliday junctions (Table 7). There are eight gene pairs among these fourteen that are combinations of one of MUS81 or MMS4 with one of the genes RAD51, RAD52, RAD54, RAD55 (involved in
the early stage of HR). *MUS81* and *MMS4* only work in double Holliday junction pathway (dHJ), one of three possible sub-pathways in HR, but not in the other two sub-pathways are synthesis-dependent strand annealing (SDSA) and break induced replication (BIR)\(^\text{17}\). The role of *RAD51, RAD52, RAD54*, and *RAD55* in multiple HR sub pathways can be thought of as pleiotropy that can obscure epistasis with the dHJ subpathway, especially if the SDSA and BIR subpathways are the dominant mode of DNA repair\(^\text{18}\).

There are another four gene pairs among the fourteen not predicted correctly that are each a combination of *SGS1* with one of the genes of *RAD51, RAD52, RAD54* and *RAD57*, these four pairs all show positive interactions but the predictions of order disagree with our ‘gold standard’ set. In fact, with hindsight these pairs should have been excluded from the gold standard set given the multiple functions of *SGS1* in both early and late stage of the HR pathway. The Sgs1-Top3 complex has the function of helping extend resection at DNA double strand breaks in the early stage of the HR pathway, as well as the function of double Holliday junction dissolution in the late stage of the HR pathway\(^\text{17}\). For the gene pair *MUS81-MMS4*, the phenotype of *mms4* is statistically indistinguishable from *mus81* and the double mutant, but the phenotype of *mus81* is statistically different from the double mutant *mus81-mms4* due to the different standard error between the phenotype of *mus81* and *mms4*. The order would have been predicted correctly if the p-value cutoff of determining whether two mutants phenotype are significantly different were set to be more stringent, for example a P-value threshold of 0.01 rather than 0.05. For the gene pair *RAD51* and *RAD54*, the prediction was co-equality, suggesting that a lack of *RAD51* filament or *RAD54* motor protein to aid strand invasion and D loop formation represented equal impairment of the HR pathway. This example illustrates that epistasis analysis can indicate where two genes work
together (through a prediction of co-equality), but fail to determine upstream/downstream order where it exists.

Figure 13: Prediction performance on datasets DR_NOMMS, DR_MMS and DR_SENS. The precision and recall are calculated when EUD model (epistatic gene is upstream with default definition of an epistatic gene) and PG constraint (gene pairs with positive genetic interaction) were applied on DR_NOMMS, DR_MMS and DR_SENS dataset.
Figure 14: Epsilon changes on three different datasets. 
**Green color:** the epsilon score for each pair in DR_NOMMS dataset. **Red color:** the epsilon score for each pair in DR_MMS dataset. **Blue color:** the epsilon score for each pair in DR_SENS dataset. The triangle indicates the wrong prediction and circle indicates the correct prediction using the genetic interaction information in the corresponding dataset.
<table>
<thead>
<tr>
<th>Gene X</th>
<th>Gene Y</th>
<th>True Order</th>
<th>DR_NOMMS</th>
<th>DR_MMS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>GI types</td>
<td>Predicted order</td>
</tr>
<tr>
<td>RAD57</td>
<td>MUS81</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD51</td>
<td>RAD55</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD52</td>
<td>RAD55</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD52</td>
<td>SGS1</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>SHU2</td>
<td>PSY3</td>
<td>In complex</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>SHU1</td>
<td>CSM2</td>
<td>In complex</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>SHU1</td>
<td>PSY3</td>
<td>In complex</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>CSM2</td>
<td>PSY3</td>
<td>In complex</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
</tbody>
</table>

Table 6: Gene pairs predicted wrong in DR_MMS but correct in DR_SENS

<table>
<thead>
<tr>
<th>Gene X</th>
<th>Gene Y</th>
<th>True Order</th>
<th>DR_MMS</th>
<th>DR_SENS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>GI types</td>
<td>Predicted order</td>
</tr>
<tr>
<td>RAD57</td>
<td>MMS4</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD57</td>
<td>RAD55</td>
<td>In complex</td>
<td>Partial Co-equal</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD54</td>
<td>RAD57</td>
<td>Downstream</td>
<td>Co-equal</td>
<td>In complex</td>
</tr>
<tr>
<td>RAD54</td>
<td>RAD55</td>
<td>Downstream</td>
<td>Co-equal</td>
<td>In complex</td>
</tr>
<tr>
<td>SGS1</td>
<td>RAD55</td>
<td>Downstream</td>
<td>Co-equal</td>
<td>In complex</td>
</tr>
<tr>
<td>SHU2</td>
<td>SHU1</td>
<td>In complex</td>
<td>Partial Co-equal</td>
<td>Unknown order</td>
</tr>
<tr>
<td>SHU2</td>
<td>CSM2</td>
<td>In complex</td>
<td>Partial Co-equal</td>
<td>Unknown order</td>
</tr>
<tr>
<td>Gene X</td>
<td>Gene Y</td>
<td>True Order</td>
<td>GI types</td>
<td>Predicted order</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>------------</td>
<td>---------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>RAD55</td>
<td>MMS4</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>MUS81</td>
<td>MMS4</td>
<td>Upstream</td>
<td>Partial Co-equal</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD51</td>
<td>MMS4</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD54</td>
<td>MMS4</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD52</td>
<td>MMS4</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>SGS1</td>
<td>RAD57</td>
<td>Downstream</td>
<td>Complete masking</td>
<td>Upstream</td>
</tr>
<tr>
<td>RAD55</td>
<td>MUS81</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD51</td>
<td>MUS81</td>
<td>Upstream</td>
<td>Negative Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD54</td>
<td>MUS81</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD52</td>
<td>MUS81</td>
<td>Upstream</td>
<td>No Genetic interaction</td>
<td>Unknown order</td>
</tr>
<tr>
<td>RAD54</td>
<td>RAD51</td>
<td>Downstream</td>
<td>Co-equal</td>
<td>In complex</td>
</tr>
<tr>
<td>SGS1</td>
<td>RAD51</td>
<td>Downstream</td>
<td>Partial masking</td>
<td>Upstream</td>
</tr>
<tr>
<td>SGS1</td>
<td>RAD54</td>
<td>Downstream</td>
<td>Partial masking</td>
<td>Upstream</td>
</tr>
<tr>
<td>RAD52</td>
<td>SGS1</td>
<td>Upstream</td>
<td>Partial Co-equal</td>
<td>Unknown order</td>
</tr>
</tbody>
</table>
2.11 Expansion of DR_SENS dataset –DNA repair related genes collection

The high performance of order prediction with the DR_SENS dataset motivated me to measure the genetic interaction of a larger number of gene pairs involved in DNA repair under different DNA damaging drugs. This would not only allow prediction models to be further refined and verified but also permit novel discoveries about gene order amongst DNA repair genes. The DNA repair-related genes of interest were collected first by searching for additional candidate genes. Because high GI profile correlated genes are more likely to have similar function and to act in the same pathway, we identified candidates as those having a high GI profile correlation with any gene in a list of 278 manually curated DNA repair related genes. The correlation cutoff used here is 0.3, which was derived from the 99.99 percentile of the GI profile correlation distribution generated from the collection of unpublished new SGA data from the Boone/Andrews Labs (retrieved JAN 2013). A total of 26 additional candidate genes were identified, expanding our genes set from 278 to 304. This set of 304 was subsequently filtered down to 186 by whether the non-essential genes in the list are sensitive to drugs that cause DNA damage. The drug sensitivity data was collected from a recent HIP-HOP study that involves 1800 biological active compounds\(^9\). Eleven drugs that cause DNA damage (Table 8) were used to filter out the genes that are not sensitive to any of them, so that selected genes of interest are those sensitive to at least one of the 11 drugs.
Table 8: List of DNA damaging drugs and their mechanism of action

<table>
<thead>
<tr>
<th>Drug</th>
<th>Mechanism of Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ravidomycin</td>
<td>Inhibits DNA synthesis followed by inhibition of RNA synthesis, proposed to affect Topoisomerase II function</td>
</tr>
<tr>
<td>Camptothecin</td>
<td>Generates toxic topoisomerase I complex, inhibits DNA synthesis</td>
</tr>
<tr>
<td>5-Fluorouracil</td>
<td>Pyrimidine analogue, incorporated into DNA and RNA</td>
</tr>
<tr>
<td>Aclacinomycin A, Aclarubicin</td>
<td>Intercalates into DNA and interacts with topoisomerases I and II</td>
</tr>
<tr>
<td>Doxorubicin</td>
<td>DNA intercalator, inhibits DNA synthesis by blocking the progression of topoisomerase II</td>
</tr>
<tr>
<td>Cryptolepine</td>
<td>DNA intercalator</td>
</tr>
<tr>
<td>Mechlorethamine</td>
<td>Damages DNA by forming crosslinks within DNA, and between DNA and protein</td>
</tr>
<tr>
<td>Methyl methanesulfonate (MMS)</td>
<td>Stalls replication forks during DNA replication and repair</td>
</tr>
<tr>
<td>Flucytosine</td>
<td>Fluorinated pyrimidine analogue that inhibits RNA and DNA synthesis</td>
</tr>
<tr>
<td>Bleomycin A2</td>
<td>Binds DNA and induces double-stranded breaks</td>
</tr>
<tr>
<td>Zeocin, Zeomycin</td>
<td>Leads to double-stranded DNA breaks</td>
</tr>
</tbody>
</table>
2.12 Expansion of DR_SENS dataset – Gene pair selection

One of the prerequisites of gene order prediction is that two genes are located in the same pathway. Although the genes in the same pathway do not necessarily encode physically interacting proteins, those that do are more likely to act in the same pathway. Moreover, gene pairs with high GI profile correlation are also more likely to have the same function and to act in the same pathway. Therefore protein-protein interacting and GI profile-correlated gene pairs in the space of the previously selected DNA repair related genes are the desired set of gene pairs to measure genetic interaction and to predict gene order (Figure 15). The PPI relationship between genes are retrieved from the HINT database including both binary and co-complex interaction. The GI profile-correlated gene pairs were selected similarly to the DNA repair related genes selected by GI profile correlation above: in the space of 186 DNA repair related genes, any gene pair that has a high GI profile correlation (cutoff: 0.3) is included in the experiment. Once the gene order of direct PPI gene pairs is determined, the order of indirect PPI gene pairs can be inferred without GI measurements, for example gene A could be inferred as upstream of gene C if it is known that gene A is upstream of gene B and gene B is upstream of gene C. Nevertheless GI measurements for the indirect PPI gene pairs could be very useful to confirm the inferred gene order and make prediction for the gene pairs that cannot be inferred, but with the cost of measuring more gene pairs. Two genes separated by distance n in a PPI network are defined as n-hop pairs. After balancing the number of experiments and the prediction performance on the DR_SENS dataset (Figure 16), up to 2-hop PPI gene pairs (direct 1-hop gene pairs and 2-hop gene pairs) are prioritized for testing of genetic interaction. The total number of the 1-hop and 2-hop PPI gene pairs plus the high GI profile correlating pairs is 545 in the space of 186 previous selected DNA
repair related genes. Essential-essential gene pairs represent a special case for two reasons: 1) Both alleles are hypomorphic in our dataset, whereas the ‘rules of epistasis’ typically assume null mutations, and 2) most of the essential-essential gene pairs are in the same complex, where we do not expect to find many ordered relationships. Therefore, we removed 174 essential-essential gene pairs, making the final number of pairs to test 371 involving 112 genes (90 non-essential, 22 essential). The gene pair selection criteria used in this proposal significantly reduces the number of gene pairs (371 pairs) that need to be measured, a factor of 98% below the 17,298 gene pairs that would represent all pairings of target genes.
**Figure 15:** Protein-protein interaction map in the space of the genes of interest. Protein-protein interacting gene pairs involving 112 DNA repair related genes. **Black edge:** Binary protein-protein interacting pairs. **Green edge:** Co-complex protein-protein interacting pairs. **Red edge:** High SGA profile correlated pairs. In the case that one edge has multiple PPI relationships, only the relationship with the highest priority was shown. The priority of the three types of relationship is black, green and red from high to low.
Figure 16: Evaluating gene pair prioritization criteria using DR_SENS dataset.

A) Gene order prediction performance for various selection criteria based on network distance (hops) in the PPI graph, measured in the space of DR_SENS gene list. "Full matrix" represents performance using all gene pairs within the DR_SENS set. B) Number of gene pairs tested (experiments) when up to n hops gene pairs are selected (n = 1 to 10). Red Dot: Indicates a good balance in terms of a relatively small number of gene pairs measured and strong prediction performance.


3 Conclusion and discussion

3.1 Model-wise comparison: EUD is the overall best model

To seek which model is overall the best, I used precision vs. recall scatter plot and precision/recall violin-box plots representing the overall prediction performance of each model on all datasets filtered by different constraint (Figure 17, Figure 18, Figure 19, Figure 20 each represents a different constraint: AG, PG, ET, CET respectively). These plots showed that the EUD and EUA models (epistatic gene is upstream models) are overall better models than EDD and EDA models (epsistatic gene is downstream models) with all dataset constraints. The performance of SF model is as good as the EUD and EUA models in all cases except when PG constraint was applied. This may be due to the fact that the SF model cannot handle co-equal predictions (the scaled fitness value is not appropriate for gene pairs with co-equal relationship, given that the two single mutants are indistinguishable). There is almost no performance difference between EUD and EUA model when the overall prediction performance was compared across all datasets, however EUD performed slightly better than or equivalent to EUA on every individual dataset except TFT dataset (Figure 17A, Figure 18A, Figure 19A, Figure 20A). The EUA outperforms EUD on TFT dataset perhaps because TFT has a larger number of gene pairs with partial suppression relationships (Table 4). Therefore I picked EUD as the overall best model in this study. Nevertheless, it will be interesting to see how EUA, EUD and SF perform when larger datasets become available. Especially, EUA might outperform EUD when more partial suppression cases available in the dataset, given the EUA model is supported by the aforementioned mathematical model of phenotypes expected given known gene order.
Figure 17: Model-wise prediction performance comparison for all gene pairs with genetic interaction information (AG constraint).

A) Precision vs. recall scatter plot for different combinations of models and datasets
B) Precision violin plot (X-Axis: different models Y-axis: precision)
C) Recall violin plot (X-Axis: different models Y-axis: recall)
Figure 18: Model-wise prediction performance comparison for gene pairs with positive genetic interaction (PG constraint).
A) Precision vs. recall scatter plot for different combinations of models and datasets
B) Precision violin plot (X-Axis: different models Y-axis: precision)
C) Recall violin plot (X-Axis: different models Y-axis: recall)
Figure 19: Model-wise prediction performance comparison for gene pairs with partial and complete epistatic to relationship (ET constraint).
A) Precision vs. recall scatter plot for different combinations of models and datasets
B) Precision violin plot (X-Axis: different models Y-axis: precision)
C) Recall violin plot (X-Axis: different models Y-axis: recall)
Figure 20: Model-wise prediction performance comparison for gene pairs with complete epistatic to relationship (CET constraint).
A) Precision vs. recall scatter plot for different combinations of models and datasets
B) Precision violin plot (X-Axis: different models Y-axis: precision)
C) Recall violin plot (X-Axis: different models Y-axis: recall)
3.2 Dataset constraint comparison: keep all gene pairs with positive GI in the training set.

No significant overall precision performance difference was observed when the EUD model was applied to gene sets defined by different constraints (Figure 21). Moreover, as expected, the recall became higher when the more stringent constraint was applied to the dataset because fewer unknown order predictions were made (for non- or negatively-interacting gene pairs) with the AG (all genetic interactions) constraint than with the PG (positive interactions only) constraint. Similarly, fewer unknown order predictions on pseudo and partial co-equal cases were made when the dataset was filter by PG constraint than ET (epistatic pairs only) constraint. Moreover, the result implied that the co-equal cases existing in PG but not in ET didn’t affect the prediction performance negatively. Similarly, the partial epistasis cases existing with the ET but not the “complete epistatic to” (CET) constraint did not negatively affect performance. Moreover, the sample size shrinks when more stringent cutoff is applied, therefore considering the overall balance of precision, recall, and the sample size, I suggest keeping all gene pairs with positive genetic interaction in the training dataset when making gene functional order inference.
Figure 21: Constraint-wise prediction performance comparison. The comparison was made on all six datasets using the EUD (epistatic gene is upstream with default definition of an epistasis gene) model.
A) Precision VS recall scatter plot for different combinations of datasets and constraints
B) Precision violin plot (X-Axis: different levels of constraint Y-axis: precision)
C) Recall violin plot (X-Axis: different levels of constraint Y-axis: recall)
3.3 Datasets comparison: epistatic gene is upstream is a general trend on
substrate dependent pathways

The prediction performance in terms of precision and recall of each dataset when
EUD model and PG constraint were applied can be found in Figure 22. Although it is a more
general trend that the epistatic gene is upstream than downstream based on the comparison
on different models, the significance of the trend varies among different datasets as each
contains different types of biological pathway (TFT, KS, MP) or different quantitative
phenotypes (DR_NOMMS, DR_MMS, DR_SENS). The trend almost does not exist on TFT
and KS datasets, both of which are switch-regulatory pathways. The trend becomes relatively
stronger with the MP, DR_NOMMS and DR_MMS datasets and peaks with the DR_SENS
dataset. Interestingly, the datasets that show stronger trends all represent substrate dependent
pathways. The MP dataset represents metabolic pathways that are mainly biosynthetic, and
DR_NOMMS, DR_MMS and DR_SENS represent DNA homologous recombination repair
pathway in which each step depends on the previous intermediate product such as RAD51
filament and D-loop intermediate. Therefore I conclude “epistatic gene is upstream” is a
useful general model for predicting order for substrate dependent pathways. The missing or
weak trend on switch- regulatory pathways may be explained by a lack of information about
the sign of the regulation, for example a transcription factor can activate or inhibit the
expression of a target gene, and a kinase can activate or inhibit the activity of the substrate by
phosphorylation.
Figure 22: Dataset-wise prediction performance comparison.
The prediction performance (precision and recall) on each of six datasets was measured when EUD model and PG constraint were applied.
3.4 Best combination: EUD + PG + DR_SENS

Taking all of the results on gene order prediction performance above together, across different models, datasets, phenotypes, and constraints defining different gene pair subsets, the best performance was the EUD model applied to positive genetic interacting gene pairs related to DNA repair (a substrate dependent pathway) using drug sensitivity as a quantitative phenotype. This conclusion motivated me to collect more DNA repair related gene pairs corresponding to PPIs, as these should be enriched for working in the same pathway. The genetic interaction of these gene pairs with/without different drugs can be measured to provide ‘gold standard’ training samples for gene order inference.

3.5 Potentially confounding factors in gene order prediction: phenotype specificity, pleiotropy and circumvention of the upstream gene.

The selection of a quantitative phenotype is crucial for an epistasis analysis to correctly infer the gene functional order in pathways. The more specific the phenotype is to the gene’s function, the more accurately order inference can be achieved. For example, in the St Onge et al 2007 study, the target genes were selected for involvement in DNA repair, so that it makes sense that the precision and recall increased significantly when predicting gene functional order on DR_SENS dataset compare to DR_NOMMS dataset since the phenotype used in DR_SENS dataset is more specific and more close to the function of DNA repair pathways than a simple growth phenotype (DR_NOMMS). When a non-specific phenotype is selected to define the genetic interaction, there are several possible reasons that might cause the false gene order prediction: First, the mutant is more likely to show no phenotype if the function of the pathway is not required for growth under a non-specific condition; consequently the gene pairs will tend to show no genetic interaction leading to their
exclusion from the epistasis analysis. Second, even if the mutant’s phenotype differs from wild type, the phenotype may be the aggregate result of the gene’s multiple functions if the gene is pleiotropic, as a result the observed genetic interaction might not accurately reflect the epistasis in the pathway that relates solely to the shared function of the gene pair.

Pleiotropy represents an opportunity for order prediction, as well as a challenge. For example, success of order prediction in substrate-dependent pathways using the “epistatic gene is upstream” rule is likely to often depend on pleiotropy, where the upstream gene has additional functions that give it a stronger single-mutant phenotype, and the phenotypic difference between two single mutants is a prerequisite for gene order inference.

For a pair of genes working together in a pathway, the upstream gene can be completely or partially circumvented (i.e., if the signal to or the substrate for the downstream gene can be obtained from another source). This can have an impact on order prediction, and can also lead to partial epistasis that is not accommodated by conventional epistasis analysis. In summary, the specificity of phenotypes, pleiotropic effects and circumvention of upstream genes are all confounding factors in gene order inference.

### 3.6 Impact of dataset quality on gene order predictions

Another possible impediment to accurately infer gene functional order is the low quality of the dataset. The subtle differences between the phenotype of different mutants are crucial in this study as a small change in the phenotype may change the subtype of positive genetic interaction (for example from complete masking to partial suppression), which changes the direction of the predicted order. Therefore, not only the accuracy of the phenotype measurement itself but also the uncertainty of each measurement becomes important. Lack of
sufficient biological replicates is a problem common to both large-scale small-scale genetic interaction datasets. Uncertainty inferred from technical replicates often underestimates the true uncertainty, and uncertainty inferred from a small number of biological replicates (2-4) is often overestimated. The lack of independent samples (biological replicates) reduces the power of more robust paired or independent two location t-tests because fewer degrees of freedom can cause two different phenotypes to be statistically indistinguishable.

4 Future direction

4.1 Computational future direction

More research on computational analysis methods has the potential to yield improved data-driven models. First, the sign of the regulation (activation/inhibition) can be collected for TFT and KS dataset by deeper literature mining, and used as an important feature when making order prediction on switch regulatory pathways. Second, uncertainty of phenotypic measurements needs to be incorporated into models appropriately in order to make better predictions: the SF model is essentially a k-nearest neighbor classification model based on the scaled fitness feature. Better ways of calculating uncertainty of the scaled fitness feature and integrating this into neighbor distance calculation will likely to improve predictions. Third, more network-topology features can be collected, such as the in-degree and out-degree of target genes in other biological networks, which may usefully predict pleiotropy. Network consistent with circumvention of the upstream gene might be identified. Finally, new features designed based on combinations of other features could also lead to better performance.
4.2 Experimental future direction

As mentioned in the previous results section, motivated by the high prediction performance on the combination of EUD + PG + DR_SENS, 371 PPI gene pairs among 122 DNA repair-related genes have been collected. The genetic interaction of these pairs with and without different drugs could be measured, and the drug sensitivity to each gene calculated and treated as the quantitative phenotype in genetic interaction estimates. Power analysis could be done to determine how many biological replicates to use in order to achieve a better statistically inference. With this new dataset and larger number of gene pairs, the prediction models could be further evaluated, and we could make novel predictions of gene order in DNA related pathways. In the future, we can envision expanding prioritized order-inference screens to genome scale across multiple phenotypes.

5 Reference List


