Derivation and use of gene network models to make quantitative predictions of genetic interaction data

Hilary Phenix

This thesis is submitted as partial fulfillment
of the Ph.D. program in
Cellular and Molecular Medicine

University of Ottawa
Ottawa, Canada

Abstract

This thesis investigates how pairwise combinatorial gene and stimulus perturbation experiments are conducted and interpreted. In particular, I investigate gene perturbation in the form of knockout, which can be achieved in a pairwise manner by SGA or CRISPR/Cas9 methods. In the present literature, I distinguish two approaches to interpretation: the calculation of stimulus and gene interactions, and the identification of equality among phenotypes measured for distinct perturbation conditions. I describe how each approach has been applied to derive hypotheses about gene regulatory networks. I identify conflicts and uncertainties in the assumptions allowing these derivations, and explore theoretically and experimentally approaches to improve the interpretation of genetic interaction data. I apply the approaches to a well-studied gene regulatory branch of the DNA damage checkpoint (DDC) pathway of *Saccharomyces cerevisiae*, and confirm the known order of genes within this pathway. I also describe observations that seem inconsistent with this pathway structure. I explore this inconsistency experimentally and discover that high concentrations of the DNA alkylating drug methyl methanesulfonate cause a cell division arrest program distinct from a G1 or G2/M checkpoint or from DNA damage adaptation, that resembles an endocycle.

Acknowledgements

The author is infinitely indebted to her advisor, Mads Kærn, for the opportunities in learning and research experience he has enabled. This thesis would not exist without his guidance and training, nor the guidance of Theodore Perkins. The experiments presented would not exist without the guidance and training of Mila Tepliakova. Contributions of and discussions with Kristin Baetz, Vida Abedi, Cory Batenchuk and Daniel Jedrysiak shaped the directions taken in methodology and research. The University of Ottawa Admission Scholarship, Ontario Graduate Scholarship, Alexander Graham Bell Canada Graduate Scholarship, and Genome Canada Computational Regulomics Training Program Award have financially supported this work.

To my family

Contents

List of Figures

List of Tables

Overview

The aim of this thesis is to increase knowledge about hypotheses that may be derived by analysis of pair-wise combinatorial gene and stimulus perturbation experiments. I focus on hypotheses having the form of topologies with directed edges between gene nodes. In the Introduction (Chapter 1), I identify conflicts or ambiguities in how such hypotheses are derived among previous studies, as well as limits in applicability.

To address these problems, in Chapter 2, the assumptions of classical epistasis analysis are formalized to examine if and how these assumptions can be adapted to quantitative phenotypes. Classical assumptions nonetheless limit hypotheses to be in the form of hierarchical topologies. In Chapter 3 the assumptions of epistasis analysis are generalized to allow any acyclic topology to be hypothesized. In both chapters, models of gene topologies are used to derive inference methods, which are then tested on simulated or experimental datasets.

By applying the inference methods to data obtained for genes in the DNA damage checkpoint (DCC) pathway, I identify a response to high doses of a DNA alkylating drug (methyl methanesulfonate/MMS) that is inconsistent with the known biology of the pathway. In Chapter 4, I conduct experiments to examine this inconsistency. These experiments suggest that high drug dose causes cells to override a G2/M checkpoint, which may explain how the frequency of cells in G1 increases as function of MMS dose.

This thesis is written as a monograph, with deviation in structure owing to the unique background information required to articulate the gap in knowledge addressed in Chapter 4. Each of Chapters 2 through 4 has sections for summary, author contributions, background, objectives, results, discussion and methodology. Chapter 5 is the Conclusion.

1

Introduction

Summary

In this chapter, I introduce how pair-wise combinatorial gene and stimulus perturbation experiments are conducted and interpreted. I distinguish two approaches to interpretation: the calculation of stimulus and gene interactions, and the identification of equality among phenotypes measured for distinct perturbation conditions. I describe how each approach has been applied to derive hypotheses about how products of genes regulate one another and the phenotype. I identify conflicts and ambiguities in the assumptions allowing these derivations, which set the motivation for this thesis.

Epistasis is a pattern of phenotypic equality predicted by gene topology models wherein genes function in a hierarchical manner to regulate a phenotype. Such models are the basis for several methods developed to use the identification of epistasis to infer or hypothesize how genes function in hierarchical pathways. These models and methods are the focus of this thesis.

Contents

1.1 Gene Interactions

Genetic interaction data are the phenotypic measurements of organisms having both, one or none of two genes perturbed. Depending on the organism, genetic interaction data may be acquired on the scale of millions of gene pairs. The data can be analyzed in many ways, but ultimately the aim of all analyzes is to advance knowledge about how the effect of a genetic perturbation depends on perturbations to other genes in the genome. This dependency is known as a gene interaction.

Knowledge of gene interactions has importance for allowing a basic understanding of how to predict phenotype from genotype [1], of genome evolution [2] and of how to build synthetic genomes. The last is exemplified by the requirement of both essential and non-essential genes in the first synthesized minimal genome [3], since deleting particular combinations of non-essential genes can be lethal.

Knowledge of gene interactions allows discovery of previously uncharacterized gene functions [4], and may be useful to design gene therapeutic treatments or improve prediction of heritability of disease. In the case of polygenic disease wherein heritability is not Mendelian, lack of knowledge of gene interactions is hypothesized to be a cause of unpredictable heritability [5, 6]. In the case of a disease having a known single gene cause (for e.g. childhood Mendelian disorders), discovery of suppressing gene interactions provide knowledge about secondary mutations which may be targeted because they suppress the effect of the disease causing-mutation [7]. In the case of disease caused by acquired mutation(s), synthetic lethal gene interactions provide knowledge about secondary mutations which may be targeted because they cause death selectively in cells having the disease-causing mutation [8, 9].

Suppressing [10] and synthetic lethal [11–13] interactions represent the opposite extremes of the phenomenon of gene interaction. When quantitative phenotypes

are the basis for genetic interaction data, gene interactions can be calculated within these extremes. This calculation is described in §1.3. By this calculation, suppression corresponds to an extreme positive gene interaction [14], whereas synthetic lethality to an extreme negative gene interaction (Eq 1.10). An alternative sub-type of gene interaction additionally has a pattern of equality among phenotypes. This pattern, wherein a double gene perturbation has an effect that mimics the effect of perturbing only one of the two corresponding genes, is known as epistasis. Suppressing gene interactions, for example, can take the form of epistasis. The identification of epistasis is described in §1.5.

1.2 Conducting combinatorial gene perturbation experiments

I begin by describing how genetic interaction experiments are conducted. Several advancements in array-based gene perturbation allow these experiments to be conducted in high-throughput, in model organisms from *Escherichia coli* to human.

While genetic interaction data can be generated for any organism, they are generated on the greatest scale for *Saccharomyces cerevisiae*. *S. cerevisiae* is unicellular with a short division time ($\sim$ 90 to 180 min), can be used to generate gene knockout strains by a single PCR reaction and quick transformation protocol for gene replacement [15] ($\sim$ 5h over 3 to 5 d) with high ($\sim$ 90%) efficacy [16], and arrays of strains from a collection of all non-essential gene knockout strains [17] can be mated to one another to obtain combinatorial gene knockout strains in high-throughput by synthetic genetic array (SGA) technology [18] ($<$ 1 mo). For example, [19] describe a genetic interaction dataset generated by SGA consisting of 23 million combinatorial gene interaction experiments.

While methods analogous to SGA have been designed for other unicellular organisms (*Schizosaccharomyces pombe* [20, 21], *E. coli* [22]), alternative gene perturbation strategies are required for model multicellular organisms. Combinatorial RNA interference (RNAi) has been used to generate combinatorial gene perturbation experiments in *Caenorhabditis elegans* [23], *Drosophila melanogaster* [24], *Mus musculus* [25] and human [26] cells. RNAi, however, does not knockdown GFP expression [27] or essential gene function [28] as effectively nor reproducibly as nuclease-induced frameshift mutations within gene exon regions.

While numerous nuclease-based methods for gene knockout have been applied to multicellular organisms, CRISPR/Cas9 is most readily adaptable to target specific

DNA sequences. The site-specificity of this RNA-nuclease complex can be modified by alteration of an RNA-encoding sequence, guided by rules of DNA-RNA base pairing, without requirement to alter the Cas9p nuclease-encoding sequence. In contrast, alternative nuclease systems (meganucleases, zinc finger nucleases, transcription activator-like effector nucleases) consist of multi-domain proteins and require alteration of the DNA binding domain-encoding sequence to modify site-specificity (see [29] for a review).

CRISPR/Cas9 has been applied to generate large ($\sim$ 20,000 to 142,000 gene pairs) genetic interaction datasets for human cells [30, 31]. The CRISPR/Cas9 system requires Cas9p protein and an RNA molecule (*guide RNA* or *gRNA*) encoding both a Cas9p-binding region and a DNA-recognition region. This experimental design allows multiple guide RNAs directed to different DNA sequences to be co-transformed and co-expressed in individual cells. The CRISPR/Cas9 system has documented off-target recognition [32, 33], and can result in variable genotypes. To reduce false positive effects of perturbation owing to off-target effects, generally two replicates of a perturbation condition are used, achieved by alternating the guide RNA sequence to target unique sites within exon regions of the same gene. Notably, some studies have reduced off-target effects by mutating Cas9p residues hypothesized to recognize DNA non-specifically [34]. Nuclease-dead CRISPR/Cas9 can also be used to combinatorially knockdown the expression of two genes [35], termed CRISPR interference (CRISPRi).

1.3 Calculation of stimulus and gene interactions

In the previous section, I summarized how gene perturbations could be conducted in high-throughput for various model organisms. Gene perturbation experiments may be followed by high-throughput phenotype measurements, which can take many forms. For example, array-based fluorescence measurements by imaging, microscopy or flow cytometry [36], allow quantification of estimates of transcriptional promoter activity [37], or protein expression and localization [38], depending on the nature of the fluorescent reporter and the measurement. The most common phenotype described in the literature, particularly for *S. cerevisiae*, is fitness, estimated from imaged colony size or measurements of turbidity of cell cultures over time. Gene interactions are contextual to the phenotype from which they are identified, and several recent studies overcome this problem by combining multiple distinct phenotypes to identify gene interactions [39, 40]. In the present section, I formalize the definition of gene interaction and its calculation from phenotypes.

Gene interaction can be defined as a calculated deviation from a null expectation. For example, a multiplicative null expectation [41] is,

$$\frac{T_{wt}}{T_{\Delta x}} = \frac{T_{\Delta y}}{T_{\Delta x \Delta y}} \tag{1.1}$$

where T_{wt} is a phenotypic trait measurement of a wildtype haploid strain, and $T_{\Delta x}$, $T_{\Delta y}$ and $T_{\Delta x \Delta y}$ are the measurements when the open reading frame (ORF) of gene x, gene y or both genes are deleted, respectively. An alternative null expectation is additive,

$$T_{\Delta x} - T_{wt} = T_{\Delta x \Delta y} - T_{\Delta y}. \tag{1.2}$$

Mani et al. [42] describe how these alternative null expectations can modify the identification of gene interactions from the same datasets. For both definitions, the null expectation can be paraphrased as a first gene deletion having an effect on the phenotype that is the same regardless of a second gene deletion [41].

In the case of a multiplicative null expectation, a gene interaction Γ_{xy} is calculated as

$$\Gamma_{xy} = \frac{T_{\Delta x \Delta y} T_{wt}}{T_{\Delta x} T_{\Delta y}} - 1 \tag{1.3}$$

In the case of an additive null expectation, and when multiple replicate experiments are available, linear regression can be applied to estimate a gene interaction β_I as follows:

$$T(X,Y) = \beta_0 + \beta_X(1-X) + \beta_Y(1-Y) + \beta_I(1-X)(1-Y) + \epsilon, \tag{1.4}$$

where $X, Y \in \{0,1\}$ indicates if the genes are deleted (zero) or not (one), and ϵ is an error term. In this equation (Eq 1.4), β coefficients are estimates of the effects of gene deletions on a measured phenotypic trait T. Coefficients can be estimated by linear least squares regression of replicate T measurements, for all perturbation conditions.

The two definitions of gene interaction coincide, i.e. $\Gamma_{xy} = \beta_I$, when Γ_{xy} is based on the log-transformation of both sides of Eq 1.1 and β_I is calculated from log-transformed $T(X,Y)$ (for e.g. in [24]) wherein one replicate of each phenotype is considered.

I refer to the set of four phenotypes considered in the calculation of a gene interaction as genetic interaction data, as well as the replicates of these sets or of multiple sets

for different combinations of genes.

A stimulus-gene interaction can be calculated analogously [41], wherein the multiplicative null expectation is

$$\frac{T^0_{\Delta x}}{T^0_{wt}} = \frac{T^1_{\Delta x}}{T^1_{wt}}, \tag{1.5}$$

where T^0 and T^1 denote the phenotypes in absence (zero) or presence (one) of a stimulus. I note that the ratio of a phenotype measured with and without the stimulus is analogous to the calculation of sensitivity S [43], i.e. $S_x = T^0_x/T^1_x$.

In the case of a multiplicative null expectation, a stimulus-gene interaction Γs_x is calculated as follows,

$$\Gamma s_x = \frac{T^1_{\Delta x} T^0_{wt}}{T^1_{wt} T^0_{\Delta x}} - 1. \tag{1.6}$$

In terms of S, Γs_x is alternatively calculated as $\Gamma s_x = S_{wt}/S_{\Delta x} - 1$.

By simultaneously considering genetic interaction data acquired with and without a stimulus, one can also calculate a stimulus-dependent gene interaction, from the following multiplicative null expectation,

$$\frac{S_{wt}}{S_{\Delta x}} = \frac{S_{\Delta y}}{S_{\Delta x \Delta y}}. \tag{1.7}$$

This null expectation can be paraphrased as the expectation that a stimulus-gene interaction should be the same regardless of whether a second gene is deleted. Based on this expectation, a stimulus-dependent gene interaction Γs_{xy} is calculated as follows [41],

$$\Gamma s_{xy} = \frac{S_{\Delta x \Delta y} S_{wt}}{S_{\Delta x} S_{\Delta y}} - 1. \tag{1.8}$$

Note that the same calculation is derived from the alternative null expectation, paraphrased as the expectation that a gene interaction should be the same regardless of whether it is calculated from phenotypes measured when the stimulus is present or absent. Alternative calculations and interpretations of stimulus-dependent gene interactions are provided in [43] and [44].

In the case of an additive null expectation, and when multiple replicate experiments are available, linear regression can be applied to estimate a stimulus-dependent gene interaction δ_I as follows:

$$D(X,Y) = D_0 + \delta_X(1-X) + \delta_Y(1-Y) + \delta_I(1-X)(1-Y) + \epsilon \tag{1.9}$$

where $D(X,Y) = T(S=0,X,Y) - T(S=1,X,Y)$, and where $S, X, Y \in 0, 1$ and ϵ is an error term. Notably, the definitions of δ_I and Γs_{xy} coincide when δ_I is calculated from $log(T)$ and Γs_{xy} is based on the null hypothesis corresponding to log-transformation of each side of Eq 1.7.

1.4 Interpretation of gene interactions

In the previous section, I provided the calculation of gene interaction, stimulus-gene interaction, and stimulus-dependent gene interaction. Regardless of whether each of these calculations is based on additive or multiplicative null expectations, a calculated interaction equal to zero exemplifies theoretically non-interacting genes. In the present section, I describe how the calculation is interpreted to assign biological meaning based on the magnitude, uncertainty and sign.

In the first applications of SGA technology [45] with *S. cerevisiae*, synthetic *lethal/sick* interactions could be identified in absence of a calculation of gene interaction. This sub-type of interaction was identified for genes x and y from colony size phenotypes T when

$$T_{\Delta x \Delta y} \approx 0, \{T_{\Delta x}, T_{\Delta y}\} \approx T_{wt}. \tag{1.10}$$

In this application, deleting gene x or y had no visible effect on colony size, but deleting both genes resulted in a *lethal/sick* colony where growth was not detectable. Gene interaction is not calculated in this application since the phenotypes are qualitative, but is assumed to be infinitely negative.

Two genes having a synthetic lethal interaction have been interpreted to each have a function allowing buffering of variation in the other gene's function. This may be explained by the genes having redundant (parallel) contributions to an essential cellular process, or wherein genes contribute to a negative feedback regulatory loop that allows homeostasis [46]. Interpretations of synthetic lethality may be extended to quantified gene interactions deemed significantly negative.

Interpretations of quantitative gene interactions are often determined empirically and statistically. Such interpretations can vary between studies, according to the genes selected for analysis, biases in gene ontology annotations [47], and experimental design. In the present section I aim to provide an idea of how the data can be interpreted, by providing examples with a focus on interpretations allowed by the most comprehensive and recent analysis of gene interactions conducted in *S. cerevisiae*.

I consider two steps to the interpretation of these data. The first step is the determination of an interaction value that is significantly different from zero. Interactions may be calculated by Eq 1.1, Eq 1.4 or alternatives (for e.g. S-score in [48]). These calculations are generally corrected for errors characteristic to the experimental design and design of the perturbation strategy. For example, in the case where phenotypes correspond to arrayed colony size and SGA is the perturbation strategy, errors may include positions of colonies in an array, uncertainty in colony size estimates, variations in time colonies are grown, and limitations when gene pairs are nearby on a chromosome which do not follow Mendel's law of segregation [49]. Alternatively, when RNAi or RNA-guided nuclease is employed as the perturbation strategy, gene pairs may be discounted due to off-target RNA-DNA annealing inferred by variance in replicate gene interaction calculations.

The assignment of genetic interaction value as significantly deviating from zero typically applies the assumption that most gene pairs do not interact. In high-throughput studies, this assumption takes the form of significant values identified as outliers on each tail of the distribution that consists of all calculated values (this assumption may alternatively be incorporated earlier in the calculation of the gene interaction [48]). Thresholds of significance vary, for example in [19] outliers are identified to have a p-value ≤ 0.05, and may additionally incorporate thresholds in the calculated values.

Because the calculation of gene interaction alone does not incorporate assumptions about how two gene functions interact, the interpretation of a significantly negative or positive gene interaction has been allowed by combining the analysis of gene interaction data and alternative empirical measures of gene product functions. Such empirical data may include documented annotations of whether pairs of genes are co-expressed, have a physical (protein-protein or protein-DNA) interaction, or have shared gene ontology (GO) categorizations, such as in a biological process, complex or cellular compartment [19]. Two methods are commonly used to determine if gene interactions are reliable indicators of an interaction based on empirical data: calculation of enrichment (hypergeometric test [19]) or predictive value (precision calculated from true and false positives versus recall calculated from true positives and false negatives [50]).

There are primarily two approaches to analyzing the relation between gene interactions and empirical data. First is a similarity measure of two gene interaction profiles [50]. Similarity among profiles has the highest predictive value for pairs of genes having annotated protein-protein interactions or the same biological process. In [19], for example at the lowest recall value, precision is > 0.9 (maximum is one) for predicting a protein-protein interaction or shared biological process from similarity between genes profiles. A gene i's interaction profile is the vector of gene interaction values between i and each other gene in the dataset. Similarity between profiles is often measured by the Pearson correlation [19, 50]. High predictive value of profile similarity to identify complexes or pathways has been explained by the assumption that two genes in the same complex or pathway will are statistically more likely to have the similar patterns of gene interactions with all other genes in the genome, compared to a randomly selected pair of genes [51]. Thus large scale genetic interaction datasets are required to allow this interpretation.

This similarity measure is also used as a distance measure to visualize and analyze gene networks based on hierarchical clustering, wherein clusters may represent groups of genes having a similarity measure above the a set threshold [19]. Thus gene interaction profiles allow derivation of the connectivity structure of undirected gene networks. The connectivity patterns in these networks can be interpreted by comparison to hierarchical versus scale-free networks, for example which are derived from particular assumptions (reviewed in [2]).

A second approach is to analyze the relation between gene interaction sign and empirical data. A generalized interpretation of gene interaction sign has been that positive or interactions allow identification of genes functioning in the same pathway, whereas negative interactions identify two genes functioning in parallel or redundant pathways [52]. While this is true for a fraction of gene pairs, gene interaction sign alone does not definitively allow one to derive hypotheses about how genes function in complexes or pathways (reviewed in [53]). This conclusion agrees logically with the nature of gene interaction calculation, which does not make any assumptions specific to how genes function.

In [19], the relation between gene interaction sign and empirical measures of shared gene function are reported to be contextual to the biological process or complex considered, and whether the genes are essential (perturbed by temperature sensitive mutations) or non-essential (perturbed by gene replacement). This dataset corresponds to 23 million gene interaction experiments, corresponding to $> 90\%$ (5416) of genes in *S. cerevisiae* genome, which allowed identification of close to 1 million significant gene interactions ($\sim$ 550K negative, $\sim$ 350K positive). For essential genes, a documented protein-protein interaction had high predictive value (50% of physical interactions) for having a negative gene interaction. Negative interactions among essential genes also had high predictive value of being in the same complex if the complex was essential (63% of gene pairs in an essential complex also have a

negative interaction). These trends were absent or weak for negative interactions among non-essential genes or positive genetic interactions for any gene type. Positive genetic interactions are statistically more likely to predict genes in distinct cellular compartments (75-78% of genes with positive interactions). Nonetheless, if genes in annotated complexes are specifically analyzed in terms of genetic interaction sign, the most common trend is a given complex will be enriched for having predominantly positive or predominantly negative interactions (43% of complexes). Whether a complex is predominantly negative or positive can in part be predicted by whether the genes are essential (predominantly negative) or correspond to a specific biological process.

1.5 Identification of phenotypic equality

In the previous section I summarized approaches to the interpretations derived from the calculation of gene interaction. In the present section, I summarize an alternate approach to analyze genetic interaction data, which is the identification of epistasis. In the literature, *epistasis* is sometimes used interchangeably with *gene interaction* [54]. In the present study I use the term epistasis to reflect a sub-type of gene interaction which additionally has a pattern of phenotypic equality.

By this classical definition of epistasis [54–56], if one considers two genes X and Y, gene X is considered epistatic to gene Y when

$$T_{\Delta x} = T_{\Delta xy} \neq T_{\Delta y} \text{ and } T_{wt} \neq \{T_{\Delta y}, T_{\Delta x}, T_{\Delta xy}\}, \tag{1.11}$$

where T_{wt} is a phenotypic measurement of a wildtype strain, and $T_{\Delta x}$, $T_{\Delta y}$ and $T_{\Delta x \Delta y}$ are the measurements when when gene X, gene Y or both genes are deleted, respectively. Epistasis therefore marks a pattern of phenotypic equality between a double deletion strain and a single deletion strain.

The identification of phenotype equality requires either a qualitative and unambiguously categorizable phenotype or a method to determine if two quantitative phenotypes are statistically equivalent. Qualitative categorized phenotypes are exemplified by study [55], which makes the distinctions between sets of phenotypic outcomes such as {male, female, or hermaphrodite} or {alive, dead and engulfed by neighbouring cells, or dead with persistent corpse}. In contrast, studies [43] and [57] determine equality based on uncertainty of each phenotypic trait as estimated from replicate measurements of quantitative phenotypes such as growth rate or colony size,

respectively. St Onge et al [43] considered phenotypic traits $T_{\Delta x}$ and $T_{\Delta x \Delta y}$ equal when

$$\frac{\mu_{T_{\Delta x}} - \mu_{T_{\Delta x \Delta y}}}{\sqrt{se^2_{T_{\Delta x}} + se^2_{T_{\Delta x \Delta y}}}} < 0.75, \tag{1.12}$$

where μ_{T_x} and se_{T_x} are the mean and standard error of replicate measurements of trait T_x, respectively. Notably, this measure is similar to the calculation of a z-score, i.e. the number of standard deviations z by which a value x deviates from a population with mean μ and standard deviation σ ($z = (x - \mu)/\sigma$), or a t-statistic. By an alternative calculation, Drees et al. [57] considered two phenotypes equal if the confidence intervals of their medians overlapped. I apply the methods of Drees et al. and St Onge et al. in Chapter 3.

Alternative approaches to the identification of epistasis arise when combinatorial gene perturbation effects are estimated by linear regression (Eq 1.4). If I consider two genes X and Y, gene X is considered epistatic to gene Y when

$$\beta_I = -\beta_Y \neq -\beta_X, \{\beta_X, \beta_Y, \beta_I\} \neq 0 \tag{1.13}$$

where β parameters are estimated and defined as in Eq 1.4 for two genes X and Y. This equality can be paraphrased as the phenotypic effect of deleting gene Y (β_Y) is negated when gene X is also deleted.

Equality between β parameters in Eq 1.13 has been determined indirectly by analyzing the effect on model fitting when terms β_I and β_Y and cancelled out from the full linear model [39, 58] (Eq 1.4).

While the above examples present alternative statistical approaches to the identification of equality from a phenotype T measured multiple times under different gene perturbation conditions, equality has also been identified from multiple distinct phenotypes. For example, studies [58] and [59] consider the transcript levels obtained by microarray (> 1500 genes) to identify epistasis, whereas [39] considers features extracted from analysis of fluorescence microscopy images of antibody-stained cells (> 20 features including cellular area, nucleus area, cell count, and mitotic cell count). Notably, distinct phenotypes may be simultaneously considered to estimate gene deletion effects with a linear regression model [58]. St Onge et al considered a ratio of two phenotypes, measured in presence and absence of a stimulus, as the basis for identifying epistasis. This is analogous to the identification of epistasis from sensitivity (defined in §1.3).

1.6 Interpretation of phenotypic equality

In the previous section I briefly summarized alternative approaches to the identification of epistasis. In contrast to the calculation of gene interaction, epistasis can be predicted based on specific assumptions about how genes function. The present section will summarize these assumptions and how the identification of epistasis has allowed inference of gene order and regulation in hierarchical pathways.

In a review in 1992 [55], Avery and Wasserman defined a generalized set of assumptions about how genes function in pathways. Avery and Wasserman considered the assumptions that a signal and two genes have Boolean (ON/OFF) activity, and through activating or repressing regulation, the signal determines the activity of the upstream gene, which in turn determines the activity of the downstream gene. They assumed that the activities of the signal and the two genes determine the phenotype. I will describe the assumptions in further detail in §2.1.

From their assumptions, Avery and Wasserman derived a set of rules to apply to combinatorial gene and stimulus perturbation data to hypothesize which gene is upstream, and whether the activity of the upstream gene represses or activates the activity of downstream gene. I summarize the rules as follows:

- If deleting each gene affects the phenotype with or without the stimulus, but not both (*Rule 1*), then
  - If there is masking and if individual gene deletions affect the phenotype in the same stimulus condition, then the masking gene is an upstream activator (*Rule 2*),
  - If there is masking and if individual gene deletions affect the phenotype in the opposite stimulus conditions, then the masked gene is an upstream

repressor (*Rule 3*).

where I have modified the rules to read *gene deletion* rather than *full loss of function mutation*, masking is equivalent to epistasis as defined in Eq 1.11, and the stimulus or signal is assumed to be controlled by the experimenter independent of the other genes. A signal may be a third genetic perturbation or a change in environmental condition such as addition of a drug or nutrient.

The rules reflect several conclusions. First, the interpretation of epistasis allows distinct hypotheses when genetic interaction data are obtained with and without a stimulus. If no stimulus is tested, only a subset of hypotheses can be determined, i.e. the upstream gene is an activator. Second, the rule to interpret which gene is upstream is different depending on how the upstream gene regulates the downstream gene. Third, perturbation of the signal and the genes must cause a change in the phenotype in comparison to an unperturbed organism, and perturbation of each gene may cause a change in one stimulus condition only. Avery and Wasserman emphasized that the assumptions and rules needed to be met by a dataset to make their distinct hypotheses.

Aylor and Zeng extended the research of Avery and Wasserman to be applicable to quantitative phenotypes. Avery and Wasserman described their rules in the context of analyzing qualitative discrete phenotypes, for example *cell engulfed* versus *cell not engulfed*, which allowed identification of epistasis in absence of statistical criteria (see §1.5). Also, Avery and Wasserman had not formalized, in their assumptions, how the signal and the genes regulate the phenotype. Aylor and Zeng addressed these limitations by formalizing how the signal and genes function in a topology diagram, requiring the following additional assumption: the quantity of a phenotypic measurement is regulated by the activity of the downstream gene. Aylor and

Zeng modelled continuous phenotypes representing combinatorial stimulus and gene perturbation experiments by a linear regression model (Eq 1.4).

These extensions allowed Aylor and Zeng to propose a method for hypothesizing whether genes function in one of sixteen topologies (permutations of Topologies 1 or 2 in Figure 1.1). The method consists of finding best-fit linear models for eight phenotypes (permutations of the conditions wherein the signal and two genes are present or absent/deleted) obtained by removing terms from Eq 1.4 when they are insignificant. A hypothetical topology is determined by matching this set of best-fit models to a set of corresponding models expected for one of sixteen topologies based on their assumptions. Expected models were obtained largely by applying the identification of epistasis analogous to Eq 1.4.

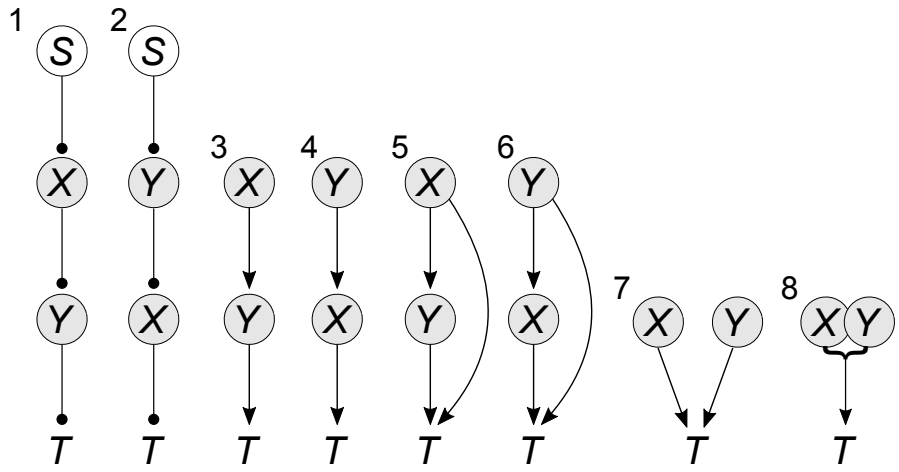

Figure 1.1: Examples of hypotheses derived from genetic interaction data. Topologies shown wherein gene X and Y regulate one another and a phenotypic trait T are among the hypotheses derived from genetic interaction data. In the case of genetic interaction experiments conducted in presence and absence of a stimulus S, the sign of regulatory interactions may also be derived. Derivation of these various topologies as hypotheses is reviewed in text. Symbols: arrow ending in circle, regulation may be activating or repressing; number, arbitrarily assigned topology number

The method of Aylor and Zeng is largely in agreement with the assumptions and rules of Avery and Wasserman, albeit adapted to quantitative phenotypes. The following paragraphs will describe methods which have disagreement with the assumptions or rules of Avery and Wasserman.

Drees et al. [57] proposed a method to analyze quantitative gene perturbation data that involved a statistical analysis of epistasis (§1.5). They described 75 possible patterns of relations among four phenotypes representing combinatorial gene perturbation data. Among these, they report 18 patterns to allow the hypothesis that gene X is an upstream activator of gene Y (Topology 3 in Figure 1.1), and an analogous 18 patterns to allow the hypothesis that gene Y is an upstream activator of gene X (Topology 4 in Figure 1.1). In each case, six of 18 patterns correspond to Rules 1 and 2 of Avery and Wasserman, for example the pattern $T_{wt} > T_{\Delta x\Delta y} = T_{\Delta x} > T_{\Delta y}$ allows the hypothesis that gene X is an upstream activator of gene Y.

For the same hypothesis, the remaining twelve of 18 patterns disagree with the Rules. I note the following disagreements. First, in one subset, the effect of deleting gene X has no effect, e.g. $T_{wt} = T_{\Delta x\Delta y} = T_{\Delta x} > T_{\Delta y}$. This pattern disagrees with Rule 1. In a second subset, there is no masking gene, e.g. $T_{\Delta y} < T_{\Delta x\Delta y} < T_{\Delta x} < T_{wt}$, therefore not reflecting Rule 2 nor 3. In a third subset, there is no masking gene and additionally deletion of gene Y has no effect, e.g. $T_{wt} = T_{\Delta y} < T_{\Delta x\Delta y} < T_{\Delta x}$, thereby additionally disagreeing with Rule 1. In the case of these twelve patterns, it is unclear what assumptions allow one to derive the same hypothesis as Rule 1 and Rule 2, yet from these alternative phenotypic patterns, as the authors did not describe their assumptions.

Two additional studies, similarly to Drees et al. [57], identify epistasis from quantitative combinatorial gene perturbation phenotypes to derive hypotheses about gene functions. First, Fischer et al. [39] generated a combinatorial gene knock-down

dataset wherein pairs of gene transcripts were targeted by RNAi. They used these data, after log transformation, to estimate the effects of gene perturbation effects with a linear regression model (Eq 1.4). Fischer et al. [39] identified epistasis according to Eq 1.13 by analyzing whether $\beta_I = \pm\beta_X$ or $\beta_I = \pm\beta_Y$ was better explained by the data by their criteria. When $\beta_I = -\beta_Y$ was a superior explanation, gene X was hypothesized to be an upstream activator of gene Y and when $\beta_I = \beta_Y$, gene X was hypothesized to be an upstream repressor of gene Y. While the former case is analogous to Rule 2 of Avery and Wasserman, i.e. the masking gene is hypothesized to be an upstream activator of the second gene for data obtained in one stimulus condition, the later case marks deviation.

Fischer et al. presents numerous deviations from the rules and assumptions of Avery and Wasserman. First, RNAi results in variable reduction of gene expression, whereas the assumptions of Avery and Wasserman are that perturbations cause complete loss or gain of gene function. The reduction of expression in high-throughput RNAi experiments has been estimated by quantitative RT-PCR, wherein Horn et al. [24] report transcript levels are reduced by $> 60\%$ for 83% of 192 RNAi experiments. A second deviation is the hypothesis of Fischer et al. [39] that a masking gene is a repressor of a downstream gene. This does not represent any of the rules of Avery and Wasserman, which require experiments in two conditions to hypothesize a repressing relationship between two genes, and additionally disagrees with hypotheses derived from gene interaction calculations. It is not clear to the author what assumptions are required to derive the inference rules of Fischer et al [39].

A study by Battle et al. [60], similarly to Fischer et al., examined quantitative genetic interaction data obtained for one stimulus condition. Battle et al. [60] identify epistasis when the data deviate less from the expectation value for epistasis (Eq 1.11) relative to two alternative expectation values. The alternative is either a multiplicative null expectation for gene interaction (Eq 1.1, see [36] for details), or an expectation

for partial epistasis. Wherein the expectation for epistasis is best explained relative to these alternatives, the masking gene is hypothesized to be downstream of the second gene. Therefore, Battle et al. interpret the genes to function in the opposite order as hypothesized by Rule 2 of Avery and Wasserman.

The hypotheses derived from the two alternatives to epistasis also represent deviations from previous studies. The first alternative, partial epistasis, is identified when genetic interaction data obtained for genes X and Y are best explained by the expectation that

$$T_{\Delta x \Delta y} = 0.5\Big(\frac{T_{\Delta x} \cdot T_{\Delta y}}{T_{wt}} + max(T_{\Delta x}, T_{\Delta y})\Big) \tag{1.14}$$

where $T_{\Delta x \Delta y}$, $T_{\Delta x}$, $T_{\Delta y}$ and T_{wt} are as defined previously. If this expectation best explains the quantity of $T_{\Delta x \Delta y}$, Battle et al. hypothesize the genes to function in Topology 5 or 6 in Figure 1.1, which are indistinguishable since gene order is not hypothesized. The second of two alternatives is the multiplicative null expectation, which if best explains the data, allows one to hypothesize genes function according to Topology 7 in Figure 1.1.

The study by Battle et al. presents many advances to the interpretation of genetic interaction data. First, it presents a synthesis of interpretations from both phenotypic equality and gene interaction calculations. Second, the study provides a method to estimate the architecture of large networks that incorporate hypotheses from all combinatorial gene perturbation analyses considered in a genetic interaction dataset. Nonetheless, the assumptions that allow their derived hypotheses are not clear to the author, particularly in the case of epistasis and partial epistasis, wherein the former hypothesis contradicts the rules of Avery and Wasserman. I address this ambiguity

in Chapter 3, wherein I analyze how well Battle et al.'s alternative expectation values meet genetic interaction data simulated with formalized assumptions.

1.6.1 Interpretation of phenotypic equality among sensitivity phenotypes

In the previous section, I provided examples of quantitative epistasis analysis for genetic interaction datasets obtained for one condition. In contrast, Avery and Wasserman's review indicates that additional hypotheses may be discerned when two conditions are analyzed. St Onge et al. [43] analyzed quantitative epistasis from genetic interaction data obtained in conditions with and without a DNA alkylating agent (0.002% v/v methyl methanesulfonate, MMS). In this study, they selected a set of 26 genes with hypothesized functions in the response to DNA damage to generate combinatorial gene deletion experiments. They reported over 50% of these genes had significant effects on the phenotype, growth rate, both in presence and absence of MMS. Therefore Rule 1 of Avery and Wasserman is not met in this study, and implying a subset of Avery and Wasserman's assumptions is not met.

St Onge et al [43] provided an alternative method to derive hypotheses from their dataset. They identified quantitative epistasis from sensitivity phenotypes as described in §1.3. A gene X is epistatic to gene Y when

$$S_{\Delta x \Delta y} = S_{\Delta x} \neq S_{\Delta y}, S_{wt} \neq \{S_{\Delta x \Delta y}, S_{\Delta x}, S_{\Delta y}\} \tag{1.15}$$

where S is as defined previously. Because S is the ratio of a phenotype obtained for the same genotype, in presence and absence of the stimulus, the analysis of S does not allow one to distinguish whether Rule 2 or Rule 3 is applicable. If epistasis

based on S is identified, St Onge et al hypothesized that the epistatic gene is an upstream activator of the second gene. It is not clear how modified assumptions from Avery and Wasserman would allow one to derive this hypothesis, and the underlying assumptions were not stated by the authors. I note that hypotheses made in this study, however, are supported by known functions of genes considered.

St Onge et al. [43] additionally consider a pattern of phenotypic equality wherein a masking gene cannot be identified,

$$S_{\Delta x \Delta y} = S_{\Delta x} = S_{\Delta y} \neq S_{wt}. \tag{1.16}$$

In this case, they hypothesize that genes X and Y function as a *unit*, for example in a *protein complex*. This hypothesis may be illustrated as Topology 8 in Figure 1.1. While this hypothesis was supported by known functions of the genes considered, it is not clear what assumptions underly this hypothesis.

1.7 Thesis motivation

In the previous sections, I introduced several approaches to interpret how genes regulate one another from the analysis of genetic interaction data. Avery and Wasserman's rules and assumptions are well-cited as evidence that these data may allow derivation of specific hypotheses about how genes functions in pathways. While these assumptions and rules are relatively well-defined, they are not met in more recent quantitative studies or in studies applying incomplete loss of function gene perturbations. In Chapter 2 I analyze how modifications to the assumptions of Avery and Wasserman affect the hypotheses one can derive. Numerous studies provide alternative interpretations compared to Avery and Wasserman, and in some cases contradictory interpretations. Because these studies do not report the assumptions underlying their interpretations, it is not clear which interpretation one should use. In Chapter 3, I consider one such method, by Battle et al. [60], for hypothesis generation in the form of hierarchical and non-hierarchical topologies, which I test on simulated and experimental stimulus and gene perturbation data.

The motivation for Chapter 4 emerges from the analysis of experimental perturbation data generated for genes in the DNA damage checkpoint in Chapter 3, and will be articulated therein.

1.8 Objectives

- The first objective is to develop a method for quantitative epistasis analysis that extends previous developments [41, 55, 58] and to benchmark this method on experimental data (Chapter 2).

- The second objective is to develop a method that generalizes epistasis analysis by allowing inference of any acyclic topology, and to determine the theoretical and practical limitations of this method on simulated and experimental data in relation to an alternative published method [60] (Chapter 3).

- The third objective is to conduct biochemical, cell division and gene expression experimental assays to test several alternative hypotheses that may explain repression of DNA damage-responsive transcription by high doses of DNA alkylating drug (Chapter 4).

2

Quantitative epistasis analysis

Summary

I describe a novel gene topology model that encapsulates a subset of assumptions of the classical epistasis analysis of Avery and Wasserman, while allowing modifications to improve applicability of classical assumptions to quantitative phenotypes. The model allows development of a method to infer topologies by comparing model-derived expected trait patterns to experimental data obtained for a well-characterized eukaryotic network. These advancements are important because of extensive application of classical epistasis analysis in biological research.

Contributions

The founding analyzes and experiments described in this chapter are published [61]. Dr. Theodore Perkins, Dr. Mads Kærn, Jacob Parker and I contributed to model and method development. Dr. Mads Kærn and I contributed to testing the method. Dr. Mads Kærn, Dr. Cory Batenchuk, Katy Morin and I contributed to experimental data analysis. Dr. Mads Kærn, Dr. Cory Batenchuk, Mila Tepliakova, Katy Morin and I contributed to experimental design. Mila Tepliakova, Katy Morin and I performed the experiments. Mila Tepliakova, Dr. Vida Abedi, Simon St-Pierre and I contributed to DNA construct and strain generation.

Contents

2.1 Background

Genetic interaction data can be analyzed to derive hypotheses about how gene products influence one another and the phenotype(s) considered. Classical epistasis analysis allows hypotheses about the order of gene functions relative to the phenotype, i.e. which of two genes is upstream of the other, and the nature of the regulatory function, i.e. if the upstream gene is a repressor or activator of the downstream gene.

Epistasis analysis requires the data to meet several criteria. However, these criteria were initially designed for qualitative phenotypes, and subsets of criteria are not applicable to quantitative ones. It remains an open question whether the assumptions underlying classical epistasis analysis can be fully adapted to quantitative genetic interaction datasets yet still allow equally powerful hypotheses.

This problem was illustrated by Aylor and Zeng [58], who formalized the assumptions of classical epistasis to enable analysis of quantitative microarray data, and found that subsets of hypotheses could no longer be derived. In particular, gene order could not be determined if neither gene was a repressor. In the present study, I begin by examining the assumptions of classical epistasis analysis and the work of Aylor and Zeng [58].

To address ambiguities in the interpretation of epistasis among previous studies, Avery and Wasserman [55] described four assumptions to predict epistasis and derive hypotheses about gene functions by classical epistasis analysis (Table 2.1). The assumptions describe the expectations for how a signal-responsive pathway with two genes regulates a phenotype. The assumptions also describe expectations for how perturbation experiments are conducted to deduce if genes function according to the assumptions. Of particular importance in the present study is the assumption that the activity of the stimulus and each gene can be ON or OFF, with no intermediate

levels of activity (Assumption 3), and the assumption that the activity of the signal determines the activity of an upstream gene, which determines the activity of a downstream gene (Assumption 4). The activities of this signal and the two genes are assumed to determine the phenotype (Assumption 2).

Table 2.1: Assumptions made by Avery and Wasserman [55] to derive rules for epistasis analysis

(1)	A signal state can be determined by the experimenter independent of the genotype and phenotype. The signal state affects the phenotype.
(2)	During the experiment, the signal and the two genes are the sole determinants of the phenotype.
(3)	The signal and the two genes are ON or OFF, there are no intermediate levels of activity.
(4)	In the wildtype, the signal determines whether the upstream gene is ON or OFF, the upstream gene in turn determines whether the downstream gene is ON or OFF.

Avery and Wasserman defined rules for epistasis analysis, derived from their assumptions (Table 2.1). The rules correspond to hypotheses for gene order and the sign of the regulatory signal between the two genes, that are determined by identifying mutually exclusive patterns among phenotypes from eight experimental conditions. The conditions correspond to the pairwise combinations of perturbations to a stimulus and each of two genes. To accommodate the assumptions, gene perturbations must cause complete loss of function (e.g. gene knockout, OFF) or complete gain of function (e.g. gene overexpression, ON). I summarize the rules below, contextualized to gene deletion perturbations for two genes X and Y,

- If deleting each gene X or Y affects the phenotype under a condition with or without the stimulus, but not both ($Rule1$), then
  - If epistasis is detected, and if each gene deletion affects the phenotype in the same stimulus condition, then the epistatic gene is an upstream activator ($Rule2$).
  - If epistasis is detected, and if each gene deletion affects the phenotype in the opposite stimulus condition, then the epistatic gene is downstream of a repressor ($Rule3$).

where epistasis was defined as follows,

- If deleting gene X or Y produces a different phenotype from the wildtype and from each other, and if deleting both genes produces a phenotype that *looks like* one of the phenotypes produced by deleting a single gene X, then X is epistatic to Y.

The rules therefore allow hypotheses about gene order and inter-gene regulatory function from interpretation of qualitative or discrete phenotypes which can be easily categorized. For example, *Dictyostelium discoideum* amoeba undergo a developmental

program in response to starvation which is accompanied by morphological changes in a transition from a unicellular to multicellular form. Deletions of genes with encoded functions in the PKA pathway cause obvious morphological defects that have been amenable to classical epistasis analysis [62] (Figure 2.1).

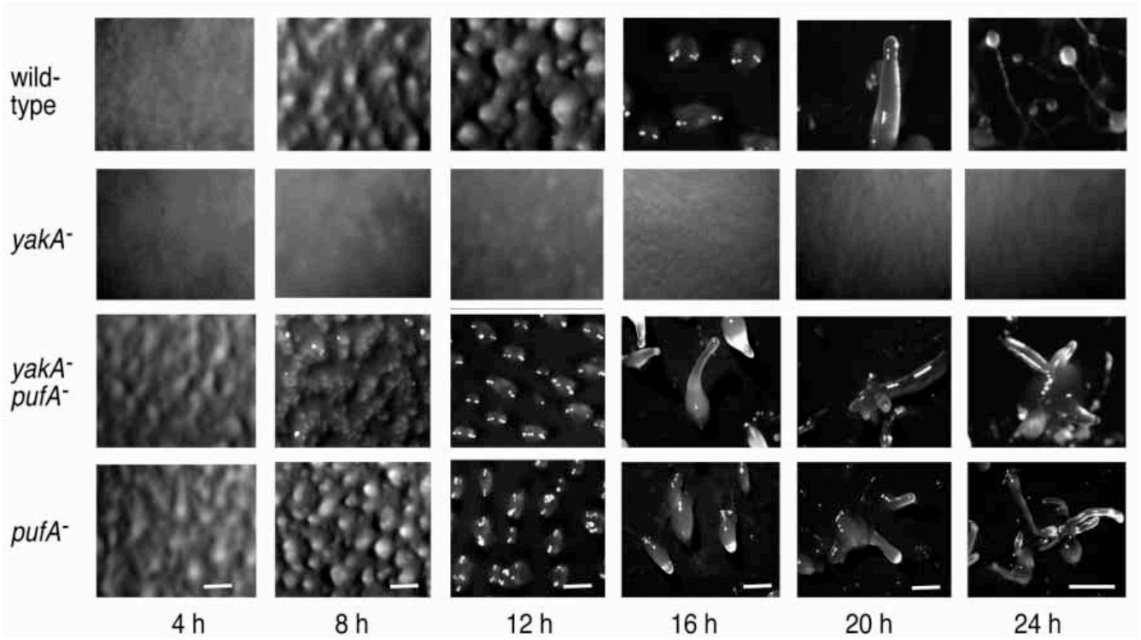

Figure 2.1: Phenotypes exemplifying classical epistasis analysis. Souza et al. [62] conducted epistasis analysis by examining the imaged morphological changes over time following starvation of *Dictyostelium discoideum* which in the wildtype marks a transition from unicellular (left) to multicellular form (right). Mutating pufA (pufA-) yields a morphological pattern which looks like the pattern when both yakA and pufA are mutated (pufA-,yakA-), i.e. pufA is epistatic to yakA. Reproduced from [62] with permission of the source journal Development.

To derive an analogous analysis for quantitative phenotypes, Aylor and Zeng [58] adapted the assumptions and rules and made several important advancements to the analysis of epistasis. First, Aylor and Zeng [58] incorporated the assumption that a signal and two genes with Boolean activities function in a hierarchy to derive sixteen possible pathways in which a stimulus S and two genes, X and Y, regulate a trait T. This number represents the possibilities that X or Y is upstream, and that each of three edges between S, X, Y and T may be activating or repressing (Figure 2.2, edges 1, 2 and 3). This advancement formalizes the classical assumptions as pathway diagrams.

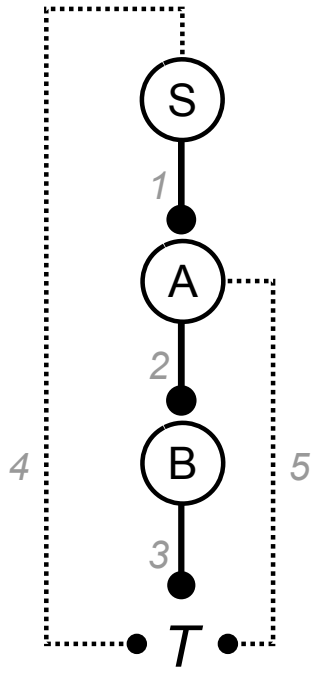

Figure 2.2: Hierarchical topologies used to interpret epistasis in previous studies. Studies [55] and [58] considered a signal S and two genes X and Y with Boolean activity to function in a hierarchical pathway regulating a phenotypic trait T. Avery and Wasserman [55] considered the edges 1 and 2, whereas Aylor and Zeng considered the edges 1, 2 and 3 [58]. The remaining edges (4-5, dotted lines) are consistent with the assumptions of Avery and Wasserman but were not explicitly considered. Symbols: edge ending in circle, may be activating or repressing.

Second, Aylor and Zeng [58] used quantitative and statistical estimates of gene perturbation effects to infer one of the sixteen pathways. Specifically, quantitative trait values, obtained by DNA microarray experiments, were regressed using a linear model as in Eq 2.1. This model incorporates estimates of the wildtype phenotype β_0, the effects of individual gene deletions (β_X, β_Y), and the effect of a gene interaction (β_I), approximating the deviation from the assumption that the effect of deleting both genes will be additive. Estimation of these terms by regression requires multiple measurements of a quantitative trait value T under all possible conditions wherein each gene is deleted or not, as follows

$$T(X,Y) = \beta_0 + \beta_X(1-X) + \beta_Y(1-Y) + \beta_I(1-X)(1-Y) + \epsilon, \tag{2.1}$$

where $X, Y \in \{0, 1\}$ indicates if the genes X and Y are deleted (zero) or not (one), respectively, for a given phenotype measurement, and ϵ is an error term.

To infer one of sixteen pathways, Aylor and Zeng compared Eq 2.1 estimated by regression of experimental traits, to Eq 2.1 theoretically predicted for each pathway. Whereas the estimated equation had dropped terms based on statistical significance, theoretical equations had dropped terms based on predicted effects to perturbations of genes in a given pathway. To make an inference, these two reduced models of equation Eq 2.1 must match.

I briefly summarize Aylor and Zeng's [58] method to derive predicted equations for a pathway. For each of sixteen pathways, Aylor and Zeng predicted four equations corresponding to the four possible gene perturbation conditions. Notably, they did not consider the perturbation of a stimulus, and rather assumed each pathway could have two possible states in the wildtype: one wherein the upstream gene X is ON, and a second wherein X is OFF. This state determines the Boolean states

of the downstream gene Y and the trait T, depending on the regulatory edges in an unperturbed pathway. These states then determine if a perturbation effect is expected to have a significant effect on the phenotype, i.e. perturbation causes a change in trait from ON to OFF or vice versa. For example, for a pathway wherein X represses Y, which activates T (i.e. $X \dashv Y \rightarrow T$) and X is ON in the wildtype, deleting Y is expected to have no effect ($\beta_Y = 0$) since it is already OFF (Table 2.2). Similarly, deleting both Y and X is expected to have no effect relative to wildtype since the effect of deleting X cannot be propagated to the the trait when Y is deleted (i.e. $\beta_I = -\beta_X$). Notably, this latter scenario corresponds to epistasis, when the effect of deleting one gene is masked or cancelled out by the effect of deleting a second gene. A best-fit model for these four conditions incorporates all coefficients predicted to be non-zero, i.e. $T = \beta_0 + \beta_X + \beta_I$. If a best-fit model for a pathway is unique among the sixteen possibilities, the pathway may be inferred as a hypothesis for how genes function.

Table 2.2: Trait equations predicted from a hierarchical pathway by Aylor and Zeng [58]. See text for interpretation. Symbols: WT, condition where no genes are deleted; ΔX or ΔY, condition where hypothetical gene X or Y is deleted; $\Delta X \Delta Y$, condition where both hypothetical genes are deleted. Adapted from [58].

State of X in WT	Genotype	Topology $X \dashv Y \rightarrow T$	T	Best-fit model
ON	WT	$ON \dashv OFF \rightarrow OFF$	β_0	
ON	ΔX	$OFF \dashv ON \rightarrow ON$	$\beta_0 + \beta_X$	
ON	ΔY	$ON \dashv OFF \rightarrow OFF$	β_0	$\beta_0 + \beta_X + \beta_I$
ON	$\Delta X \Delta Y$	$OFF \dashv OFF \rightarrow OFF$	β_0	
OFF	WT	$OFF \dashv ON \rightarrow ON$	β_0	
OFF	ΔX	$OFF \dashv ON \rightarrow ON$	β_0	
OFF	ΔY	$OFF \dashv OFF \rightarrow OFF$	$\beta_0 + \beta_Y$	$\beta_0 + \beta_Y$
OFF	$\Delta X \Delta Y$	$OFF \dashv OFF \rightarrow OFF$	$\beta_0 + \beta_Y$	

By extending such derivations to all sixteen pathways, Aylor and Zeng [58] made a number of important conclusions. First, pathways differing in the sign of regulation from the downstream gene to T had indistinguishable best-fit models. Similarly, this distinction was not made by the Avery and Wasserman rules. Second, Aylor and Zeng [58] could in some cases make an inference in absence of knowing the state of the signal. In contrast, Avery and Wasserman's rules had stricter requirements on the detection of perturbation effects in at least one stimulus condition. Finally, pathways having a positive regulatory edge between the two genes yielded indistinguishable best-fit models when X was upstream of Y versus when Y was upstream of X. This indicates that gene order cannot be identified for a subset of pathways. In contrast, the analogous distinction could be made by Rule 2 of Avery and Wasserman described above. It is not clear why Aylor and Zeng's adaptation of the Avery and Wasserman rules to quantitative phenotypes eliminated the ability to infer a subset of pathways.

By re-examining the Avery and Wasserman assumptions one can identify logical discrepancies between the assumptions of Avery and Wasserman and Aylor and Zeng. In particular, the assumptions of Avery and Wasserman do not indicate how the signal and two genes influence the phenotype. In contrast, Aylor and Zeng specify that the activity of the downstream gene in a pathway will determine the phenotype (Edge 3 in Figure 2.2). Indeed, there are two additional ways the trait can be regulated which are in full agreement with Avery and Wasserman's assumptions, but not considered by Aylor and Zeng. These correspond to the feedforward edges from the signal or the upstream gene to the trait (Figure 2.2, Edges 4 and 5).

The assumptions of Avery and Wasserman and Aylor and Zeng have a commonality which reduce applicability to quantitative data. In particular both studies consider gene activity to be ON or OFF, and activity to be determined by the state of an upstream gene's activity or an upstream stimulus. These assumptions result in the rule wherein gene perturbation has an effect in only when the stimulus is ON or OFF,

since deleting a gene with an activity that OFF is expected to have no effect on a phenotype.

In contrast to this rule, previously published studies [41, 43] that quantify the signal-dependency of gene perturbation effects report that a significant number of gene deletions (50-67%) yield significant perturbation effects both in presence and absence of a signal. These results therefore mark inapplicability of the assumptions of both Aylor and Zeng and Avery and Wasserman to these data. Such inapplicability would require modifications to the assumptions about gene activity used to derive expected phenotypes for different pathways.

In the present chapter, I describe a model to derive expectations for quantitative genetic interaction data that aims to address the limitations of these previous studies. Specifically, the model (1) considers gene activity to have a quantitative influence on a theoretical trait, (2) allows all feedforward influences in agreement with Avery and Wasserman's assumptions, and (3) allows gene activity to encompass a signal-dependent component and a basal-component. These modifications are important as they address limitations in one or both previous methods, respectively, towards deriving expectations for quantitative genetic interaction data that allows gene topology inference.

The structure of the present chapter is as follows. First, I describe a topology model that incorporates the above-described modifications. Second, I describe how this model can be used to derive an algorithm for topology inference applicable to experimental genetic interaction data. Third, I describe benchmarking of this algorithm, with experimental data obtained for a well-studied biological network of genes in *S. cerevisiae* encoding functions in galactose metabolism.

2.2 Objectives

- The first objective is to analyze the effects of modification to Avery and Wasserman's original assumptions to develop a topology model for deriving theoretical, quantitative phenotype expectations for hypothetical genetic interaction experiments.

- The second objective is to develop a topology inference method by analyzing how model-derived expectations can have topology-specific patterns.

- The third objective is to benchmark the inference method, by applying the method to experimental genetic interaction data obtained for a well-characterized eukaryotic network.

2.3 Results

2.3.1 Model to derive theoretical perturbation phenotypes

I begin by describing the model in Figure 2.3 which I use to derive theoretical, quantitative phenotype expectations for the effects of combinatorial stimulus and gene perturbations. In this hierarchical topology model, there is a signal S, which can be ON or OFF, and two genes, X and Y, which can be present (functional) or absent/deleted (non-functional). For the reader's reference, symbols used throughout the chapter are listed in Table 2.3.

In the model, S, X and Y can influence a quantitative theoretical phenotype ρ in a signal-dependent (σ influences) or signal-independent manner (α influences). These influences are determined by the states of two sets of Boolean variables for each gene. These variables correspond to whether the signal-dependent activity of a gene is ON or OFF (grey squares, Figure 2.3) or if the signal-independent activity of a gene is ON or OFF (white squares, Figure 2.3), respectively. Signal-independent influences correspond to the influence of basal gene activity whenever a gene is not deleted. In contrast, signal-dependent influences correspond to the additional activity of a gene when it is activated by the signal or the signal-dependent activity of its upstream gene.

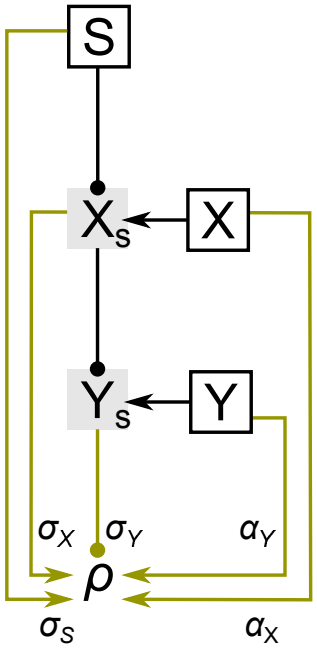

Figure 2.3: Hierarchical topology model. The model encompasses five Boolean variables (squares) and six quantitative influences (α and σ parameters) (α_I not shown) on a theoretical phenotype ρ according to Eq 2.3. The signal state S and the signal-dependent activities of the two genes X_S and Y_S regulate one another in a hierarchical topology. Symbols: white squares, Boolean variables determined by signal and gene perturbation states of a given experiment; grey squares, signal-dependent Boolean variables dependent on topology, signal and gene perturbation states; black directed edge, indicates state of upstream Boolean variable determines state of downstream Boolean variable; edge ending in circle, can be negative or positive regulation; green directed edge, quantitative influence on ρ from a Boolean variable.

Table 2.3: Descriptions of symbols used. Symbols referred to throughout text are provided with corresponding descriptions for ease of reference. Notably, descriptions provide quick reference and do not replace formal definitions provided in text. **A**: symbols related to the topology model to derive theoretical trait expressions. **B**: symbols related to the hypothetical or measured effects of signal or gene perturbations on an experimental phenotypic trait.

A	Description
X_s	Theoretical signal-dependent activity of gene X (Boolean)
Y_s	Theoretical signal-dependent activity of gene Y (Boolean)
ρ	Theoretical trait
ϕ	Difference of ρ between two conditions of the signal
ρ_0	ρ when the signal is absent and both genes X and Y are deleted
α_X	Signal-independent quantitative influence of gene X on ρ
α_Y	Signal-independent quantitative influence of gene Y on ρ
α_I	Signal-independent quantitative influence of the interaction between genes X and Y on ρ
σ_S	Quantitative influence of the signal S on ρ that is independent of genes X and Y
σ_X	Quantitative influence of gene X on ρ that is independent of gene Y
σ_Y	Quantitative influence of gene Y on ρ
B	Description
T	Experimental trait
β_0	Quantity of T when no gene is deleted and the signal is absent
β_i^0	Quantitative effect of deleting gene i when the signal is absent
β_j^0	Quantitative effect of deleting gene j when the signal is absent
β_I^0	Quantitative effect of the interaction of genes i and j when the signal is absent
β_0^1	Quantitative effect of the signal
β_i^1	Quantitative effect of deleting gene i when the signal is present
β_j^1	Quantitative effect of deleting gene j when the signal is present
β_I^1	Quantitative effect of the interaction of genes i and j when the signal is present
β'	Difference of β between two conditions of the signal
D	Difference of T between two conditions of the signal
δ_0	Quantity of D when no gene is deleted
δ_i	Quantitative signal-dependent effect of deleting gene i
δ_j	Quantitative signal-dependent effect of deleting gene j
δ_I	Quantitative signal-dependent effect of the interaction between genes i and j

By considering that a gene may have signal-independent activity, the model allows deviations from previous methods [55, 58] having the strict assumption that gene activity is only ON or OFF. Nonetheless, by considering that genes have a separate signal-dependent activity, the model also retains certain assumptions of these previous methods in the form of the hierarchical pathway, wherein the signal determines the state of the upstream gene's signal-dependent activity X_S, by activation or repression, which in turn determines the state of the downstream gene's signal-dependent activity Y_S, by activation or repression. Y_S then determines if the pathway will have an influence the trait (σ_Y). This hierarchical signal-dependent pathway has all the edges described in the model of Aylor and Zeng [58], as well as the additional feedforward influences from the signal σ_S or the upstream gene σ_S to the trait, that were not previously considered.

The deviation from strict ON/OFF gene activity, while maintaining a signal-dependent hierarchical pathway, corresponds to conservation of Assumption 4 of Avery and Wasserman, but not Assumption 3. In the model, Assumption 4 (Table 2.1) is encompassed by Eq 2.2. In this equation, Boolean variables X_S and Y_S denote the signal-dependent activities of the genes, respectively. The values of these variables are determined by whether the genes are deleted (Boolean variables X or Y are zero), the nature of upstream signal-dependent regulation (activating or repressing), and the states of upstream signal-dependent variables (S or X_S) according to Eq 2.2,

$$
\begin{aligned}
X_s(S,X) &= \begin{cases} X \cdot (1-S) & \text{if } S \text{ is a repressor,} \\ X \cdot S & \text{otherwise.} \end{cases} \\
Y_s(X_s,Y) &= \begin{cases} Y \cdot (1-X_s) & \text{if } X_s \text{ is a repressor,} \\ Y \cdot X_s & \text{otherwise.} \end{cases}
\end{aligned}
\tag{2.2}
$$

where $S, X, Y \in \{0, 1\}$ denote if the signal is present (one) or absent (zero), if the upstream gene X is present (one) or deleted (zero), or if the downstream gene Y is present (one) or deleted (zero), respectively.

Influences of the signal and genes in the topology depend on five Boolean variables in a manner dependent on the topology and on the state of signal and gene perturbation according to Eqs 2.2 and 2.3. In the model, we assume each influence contributes independently to the quantity of the theoretical phenotype (Eq 2.3) in addition to the basal phenotype quantity that is independent of the genes and the signal (ρ_0).

$$\rho(S, X, Y) = \rho_0 + \alpha_X(X) + \alpha_Y(Y) + \alpha_I(X \cdot Y) + \sigma_S(S) + \sigma_X(X_s) + \sigma_Y(Y_s). \quad (2.3)$$

where α_I accounts for non-additivity of the basal influences of x and y whenever both genes are not deleted.

The model in Figure 2.3 encompasses four possible regulatory topologies, differing in the whether the edge from the signal to X_S is activating or repressing, and whether the edge from X_S to Y_S is activating or repressing (Eq 2.2). For each of four possible regulatory topologies, I derive ρ for the eight hypothetical perturbation conditions (Table 2.4). These conditions correspond to whether the signal is present or not, and each gene is deleted or not. The derivations consist of quantitative predictions for trait values for the topologies if genes and stimuli function as assumed by the model. While one can observe certain patterns, for example the double gene deletion effect should be the same for a given signal condition independent of the topology, patterns of ρ useful for topology inference are not immediately obvious.

Table 2.4: Definitions of ρ for eight perturbation conditions for each of four topologies. Explanation of derivation provided in text.

$\rho(S,X,Y)$	(1) S $\downarrow$ X_S $\downarrow$ Y_S	(2) S $\perp$ X_S $\downarrow$ Y_S
$\rho(0,1,1)$	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y$	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y + \sigma_X + \sigma_Y$
$\rho(0,0,1)$	$\rho_0 + \alpha_Y$	$\rho_0 + \alpha_Y$
$\rho(0,1,0)$	$\rho_0 + \alpha_X$	$\rho_0 + \alpha_X + \sigma_X$
$\rho(0,0,0)$	ρ_0	ρ_0
$\rho(1,1,1)$	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y + \sigma_S + \sigma_X + \sigma_Y$	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y + \sigma_S$
$\rho(1,0,1)$	$\rho_0 + \alpha_Y + \sigma_S$	$\rho_0 + \alpha_Y + \sigma_S$
$\rho(1,1,0)$	$\rho_0 + \alpha_X + \sigma_S + \sigma_X$	$\rho_0 + \alpha_X + \sigma_S$
$\rho(1,0,0)$	$\rho_0 + \sigma_S$	$\rho_0 + \sigma_S$
$\rho(S,X,Y)$	(3) S $\downarrow$ X_S $\perp$ Y_S	(4) S $\perp$ X_S $\perp$ Y_S
$\rho(0,1,1)$	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y + \sigma_Y$	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y + \sigma_X$
$\rho(0,0,1)$	$\rho_0 + \alpha_Y + \sigma_Y$	$\rho_0 + \alpha_Y + \sigma_Y$
$\rho(0,1,0)$	$\rho_0 + \alpha_X$	$\rho_0 + \alpha_X + \sigma_X$
$\rho(0,0,0)$	ρ_0	ρ_0
$\rho(1,1,1)$	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y + \sigma_S + \sigma_X$	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y + \sigma_S + \sigma_Y$
$\rho(1,0,1)$	$\rho_0 + \alpha_Y + \sigma_S + \sigma_Y$	$\rho_0 + \alpha_Y + \sigma_S + \sigma_Y$
$\rho(1,1,0)$	$\rho_0 + \alpha_X + \sigma_S + \sigma_X$	$\rho_0 + \alpha_X + \sigma_S$
$\rho(1,0,0)$	$\rho_0 + \sigma_S$	$\rho_0 + \sigma_S$

2.3.2 Derivation of theoretical perturbation effects

Hypothetically, such patterns may be achieved by isolating individual perturbation effects. One approach to isolate individual perturbation effects is achieved by considering the linear regression model (Eq 2.4) to estimate experimental effects of signal and gene perturbation on a measured trait T. While deletion effects can be isolated by alternative means, linear regression has the advantage of considering all perturbation conditions and replicate data in one model (see Discussion), as follows:

$$\begin{aligned} T(S,i,j) = \beta_0 + (1-S)\cdot\Big(\beta_i^0(1-i) + \beta_j^0(1-j) + \beta_I^0(1-i)(1-j)\Big) \\ +S\cdot\Big(\beta_0^1 + \beta_i^1(1-i) + \beta_j^1(1-j) + \beta_I^1(1-i)(1-j)\Big) + \epsilon \end{aligned} \tag{2.4}$$

where $S, i, j \in \{0, 1\}$ denote if the experimentally determined stimulus S is present (one) or absent (zero), and if two hypothetical genes i and j are deleted (zero) or not (one), respectively, and ϵ is an error term. These perturbation conditions are analogous to those used to derive ρ, with the exception that it is not known which of genes i and j is upstream of the other or whether they function in a topology.

Eq 2.4 is analogous to regression performed by Aylor and Zeng [58] with the exception that Eq 2.4 encompasses all perturbation conditions in one model. In Eq 2.4, each genetic perturbation is associated with two β parameters, representing the effect of perturbation when the stimulus is absent (β^0) or present (β^1). Parameters β_I estimate the deviation from the assumption that the two genes' perturbation effects are additive, and therefore represent the gene interaction terms. In the model, the basal trait influence (β_0) is independent of other gene and stimulus perturbation effects, and correspondingly, the effect of the stimulus alone (β_0^1) is independent of other gene perturbation effects.

To isolate the theoretical definitions of isolated β perturbation effects, one can equate definitions of T and ρ for each of eight perturbation conditions and four topologies. This is achieved by arbitrarily setting gene i as the upstream gene X, and gene j as the downstream gene Y, as well as setting the error term ϵ to zero. This allows one to solve for the eight β parameters for each topology (Table 2.5).

Table 2.5: Definitions of β parameters for each of four topologies. Explanation of derivation provided in text.

	(1) S $\downarrow$ X_S $\downarrow$ Y_S	(2) S $\perp$ X_S $\downarrow$ Y_S
β_0	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y$	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y + \sigma_X + \sigma_Y$
β_i^0	$-\alpha_I - \alpha_X$	$-\alpha_I - \alpha_X - \sigma_X - \sigma_Y$
β_j^0	$-\alpha_I - \alpha_Y$	$-\alpha_I - \alpha_Y - \sigma_Y$
β_I^0	α_I	$\alpha_I + \sigma_Y$
β_0^1	$\sigma_S + \sigma_X + \sigma_Y$	$\sigma_S - \sigma_X - \sigma_Y$
β_i^1	$-\alpha_I - \alpha_X - \sigma_X - \sigma_Y$	$-\alpha_I - \alpha_X$
β_j^1	$-\alpha_I - \alpha_Y - \sigma_Y$	$-\alpha_I - \alpha_Y$
β_I^1	$\alpha_I + \sigma_Y$	α_I
	(3) S $\downarrow$ X_S $\perp$ Y_S	(4) S $\perp$ X_S $\perp$ Y_S
β_0	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y + \sigma_Y$	$\alpha_I + \rho_0 + \alpha_X + \alpha_Y + \sigma_X$
β_i^0	$-\alpha_I - \alpha_X$	$\sigma_Y - \alpha_X - \sigma_X - \alpha_I$
β_j^0	$-\alpha_I - \alpha_Y - \sigma_Y$	$-\alpha_I - \alpha_Y$
β_I^0	α_I	$\alpha_I - \sigma_Y$
β_0^1	$\sigma_S + \sigma_X - \sigma_Y$	$\sigma_S - \sigma_X + \sigma_Y$
β_i^1	$\sigma_Y - \alpha_X - \sigma_X - \alpha_I$	$-\alpha_I - \alpha_X$
β_j^1	$-\alpha_I - \alpha_Y$	$-\alpha_I - \alpha_Y - \sigma_Y$
β_I^1	$\alpha_I - \sigma_Y$	α_I

2.3.3 Inference of gene order using sensitivity phenotypes

Although theoretical predictions of β parameters are simplified compared to the consideration of ρ alone, this approach is insufficient to reveal rules of inference as derived by Avery and Wasserman. One does, however, observe that the theoretical values of β_k parameters ($k \in \{i, j, I\}$) have a pattern wherein one of $\{\beta_k^0, \beta_k^1\}$ is equal to a sum of α and σ influences and the other is equal to the sum of corresponding α influences only. This pattern suggests that perturbation effects can be further simplified by examining the change in gene perturbation effects between the two signal conditions.

To examine this possibility, one can consider a theoretical trait ϕ, equal to the difference between ρ of the two signal conditions. Specifically, $\phi(X, Y) = \rho(0, X, Y) - \rho(1, X, Y)$. Analogous treatment of the experimental trait is $D(i, j) = T(0, i, j) - T(1, i, j)$, wherein regression of D would allow one to estimate the signal-dependent effects of genetic perturbations (δ) (Eq 2.5) as follows:

$$D(i, j) = \delta_0 + \delta_i(1 - i) + \delta_j(1 - j) + \delta_I(1 - i)(1 - j) + \epsilon. \tag{2.5}$$

where i, j and ϵ are defined as in Eq 2.4.

To derive theoretical trait definitions for δ parameters, one can equate ϕ and D to solve for the four δ parameters for each topology (Table 2.6). To achieve this, I arbitrarily set gene i in Eq 2.5 as the upstream gene X and ϵ to zero, as performed above for the analysis of β parameters.

Table 2.6: Definitions of δ parameters for each of four topologies. Explanation of derivation provided in text.

	(1) S $\downarrow$ X_S $\downarrow$ Y_S	(2) S $\perp$ X_S $\downarrow$ Y_S
δ_0	$-\sigma_S - \sigma_X - \sigma_Y$	$\sigma_X - \sigma_S + \sigma_Y$
δ_i	$\sigma_X + \sigma_Y$	$-\sigma_X - \sigma_Y$
δ_j	σ_Y	$-\sigma_Y$
δ_I	$-\sigma_Y$	σ_Y
	(3) S $\downarrow$ X_S $\perp$ Y_S	(4) S $\perp$ X_S $\perp$ Y_S
δ_0	$\sigma_Y - \sigma_X - \sigma_S$	$\sigma_X - \sigma_S - \sigma_Y$
δ_i	$\sigma_X - \sigma_Y$	$\sigma_Y - \sigma_X$
δ_j	$-\sigma_Y$	σ_Y
δ_I	σ_Y	$-\sigma_Y$

The theoretical definitions of δ parameters reveal predictions of quantitative epistasis. When analyzing δ parameters, quantitative epistasis takes the form of $\delta_i = -\delta_I$ where $\delta_i \neq \delta_j$. Because I arbitrarily set j as the downstream gene to derive theoretical definitions of δ, the model therefore predicts that for all topologies the effect of the downstream gene (δ_j) will be masked when both genes are deleted (i.e. $\delta_I = -\delta_j, \delta_i \neq \delta_j$) (Table 2.6). This prediction contrasts that of Avery and Wasserman, whose rules reflect the upstream gene is masked for a subset of topologies, but agrees with the rules described by St Onge et al. [43] for the interpretation of epistasis identified from sensitivity phenotypes. Sensitivity phenotypes are analogous to δ parameters (see §1.3), and therefore the model in Figure 2.3 provides a set of assumptions which explain cases where the rules of St Onge et al. are expected to work.

I note that this pattern of masking depends on the assumption that σ_X is non-zero (Table 2.6). This requirement to detect masking is consistent for all topologies, because in all cases $\delta_i = \delta_j$ when $\sigma_X = 0$ and therefore the effect of deleting the upstream gene cannot be distinguished from that of the downstream gene. The consequence of this finding can be related to Avery and Wasserman's assumptions. If the model in Figure 2.3 had maintained all previous assumptions to derive quantitative trait predictions, one would consider σ_Y the only quantitative influence of the signal and genes, and δ parameters would not allow gene order inference. The contrast between derivations of the two sets of Assumptions is considered further in Discussion.

2.3.4 Inference of topology

While the consideration of δ parameters in the previous section allows the identification of gene order, it does not allow one to distinguish topologies. To test an alternative strategy to distinguish topologies, I reconsider observation that a pair of β parameters deviate predictably when compared between two signal conditions.

This predictable pattern suggests that if one considers an alternative parameter, β', that represents the difference in gene perturbation effects across conditions yet maintains information about the signal state of each β parameter. In contrast, this information is lost when considering δ parameters.

The parameters β' maintain information about signal state as they are obtained from β parameters by conditional subtraction. To obtain a parameter for each signal state $z \in \{0, 1\}$, $\beta_k'^z = \beta_k^z - \beta_k^1$ if β_k^1 is defined by α parameters only, otherwise $\beta_k'^z = \beta_k^z - \beta_k^0$. By this definition, $\beta_k'^z$ is equal to zero or is defined by σ parameters only (Table 2.7).

Table 2.7: Definitions of $\beta^{'}$ parameters, obtained from conditional subtraction of β^1 and β^0. Explanation of derivation provided in text.

	(1) S $\downarrow$ X_S $\downarrow$ Y_S	(2) S $\perp$ X_S $\downarrow$ Y_S
$\beta_i^{'0}$	0	$-\sigma_X - \sigma_Y$
$\beta_j^{'0}$	0	$-\sigma_Y$
$\beta_I^{'0}$	0	σ_Y
$\beta_i^{'1}$	$-\sigma_X - \sigma_Y$	0
$\beta_j^{'1}$	$-\sigma_Y$	0
$\beta_I^{'1}$	σ_Y	0
	(3) S $\downarrow$ X_S $\perp$ Y_S	(4) S $\perp$ X_S $\perp$ Y_S
$\beta_i^{'0}$	0	$\sigma_Y - \sigma_X$
$\beta_j^{'0}$	$-\sigma_Y$	0
$\beta_I^{'0}$	0	$-\sigma_Y$
$\beta_i^{'1}$	$\sigma_Y - \sigma_X$	0
$\beta_j^{'1}$	0	$-\sigma_Y$
$\beta_I^{'1}$	$-\sigma_Y$	0

Theoretical definitions of β' parameters reveal a pattern of predictions useful for topology inference (Table 2.7). Specifically, each of four topologies has a unique set of predicted β' parameters that is equal to zero.

Theoretical definitions of β' also reveal predictions of quantitative epistasis. In contrast to the patterns observed for δ parameters described above, the patterns of β' resemble the rules for gene order as described by Avery and Wasserman. Specifically, if genes have a non-zero perturbation effect (β') in the same signal state, the masked gene is the downstream gene (Topologies 1 and 2). In contrast, if genes have a non-zero effect in different signal states, the masked gene is the upstream gene (Topologies 3 and 4). I note that these statements however have opposite theoretical requirements. The influence σ_X must be non-zero for the first statement, yet equal to zero for the second statement.

These opposite requirements for identifying epistasis correspond to a requirement for a topology have a feedforward loop from the upstream gene to the trait in order to infer Topology 1 or 2, yet the requirement for a topology to lack this feedforward loop in order to infer Topology 3 or 4. I note that that the inability to infer gene order for Topology 1 and 2 in absence of a feedforward loop is consistent with the model and results of Aylor and Zeng [58]. This consistency suggests that the requirement can be made obvious by formalizing assumptions about gene topology functions in a quantitative model.

2.3.5 Method for topology and gene order inference

In the present section, I amalgamate the predictions of theoretical phenotype patterns discovered in the previous section to outline a step-wise algorithm to infer gene order and topology from genetic interaction data obtained by quantitative phenotype measurements of signal and gene perturbation experiments. The algorithm therefore involves first a calculation of δ and β' parameters by linear regression of experimental phenotype measurements using Eqs 2.5 and 2.4, respectively, followed by the identification of theoretically predicted patterns among corresponding experimentally-derived parameters.

Notably, the theoretical predictions cannot be directly compared to experimental parameters without additional considerations. First, in the case of theoretical parameters, there is knowledge of which β parameter is defined by α and σ parameters versus σ parameters only. Because this information is absent in experimentally-derived β parameters, one must assume there might be a quantitative difference between β parameters of the two signal states. In particular, I consider the assumption that a β_k^z parameter equal to α and σ parameters will be greater in absolute magnitude than the corresponding β_k^z predicted to be defined by α parameters only. This assumption allows a rule for obtaining β' from experimentally-derived β parameters:

$$\beta_k'^z = \begin{cases} \beta_k^z - \beta_k^0 & \text{if } |\beta_k^0| < |\beta_k^1|, \\ \beta_k^z - \beta_k^1 & \text{otherwise.} \end{cases} \tag{2.6}$$

where $k \in \{i, j, I\}$ and $z \in \{0, 1\}$.

The second consideration specific to experimental parameter analysis is the identification of epistasis. From theoretical parameters, epistasis is identified when

trait expressions are equivalent in terms of α and σ variables. In contrast, from corresponding experimental parameters, identification of epistasis can only be identified from the magnitudes and signs of δ or β' parameters estimated by linear regression. These parameters are quantities having experimental uncertainty. To address the problem of identifying epistasis in this case, I consider the assumption that experimental parameters representing theoretically equivalent parameters will deviate less in absolute magnitude than experimental uncertainty, measured as the standard error of parameters. This assumption is described in more detail in Methods and compared with alternative assumptions in Discussion.

By incorporating the two above-described considerations and suggested compensating assumptions, one can obtain a step-wise method to infer gene order and Topology from measurements of combinatorial signal and gene perturbation experiments. The algorithm considers as input a dataset corresponding to all replicate experimental trait measurements T for eight conditions corresponding to the eight possible combinations wherein each of two hypothetical genes i and j is deleted or not, and wherein an experimental stimulus is present or not. The steps of the algorithm for each input dataset are as follows:

1. Estimate δ parameters by linear regression (Eq 2.5) of T.
2. Infer gene i upstream of j if $\delta_i = -\delta_I$, $\delta_j \neq -\delta_I$, and $\delta_i, \delta_j, \delta_I \neq 0$.
3. Estimate β parameters by linear regression (Eq 2.4) and calculate β' parameters (Eq 2.6).
4. If *Step 2* failed because $\delta_i = \delta_j = -\delta_I$, infer gene i upstream of j if in one condition $z \in \{1, 0\}$, $\beta_i'^z \neq 0$, $\beta_I'^z \neq 0$, and $\beta_j'^z = 0$.
5. If *Step 2* or *Step 4* is successful, order β_i' and β_j' variables according to which gene is upstream, as in Table 2.7.
6. Infer Topology 1 to 4 if the pattern of zeros of ordered β' is equivalent to a pattern in Table 2.7.

7. If *Step 6* is successful, identify the sign of σ_Y from δ_I according to the theoretical definition of δ_I in the inferred Topology in Table 2.6.

In the final step of the method, the inference made in Step 6 is applied to deduce the sign of the edge between the inferred downstream gene and the phenotype. Therefore one can make a further distinction between one of eight topologies, rather than four. Because one of two genes may be upstream, there is a total of sixteen possible topologies inferred from a given set of experimental data. The step-wise method includes several requirements which must be made for the inference of a topology, including the statistical significance of δ and β regression coefficients and the identification of theoretically-predicted patterns among significant coefficients.

The inference of an additional edge in Step 7 marks the ability of the algorithm to deduce additional quantitative information about topology edges in addition to inference of the topology alone. In Step 7, the magnitude and sign of the edge from the inferred downstream gene to the trait (σ_Y) can be deduced directly from the experimental estimate of δ_I, since for any topology the theoretical definition of δ_I is equal to either σ_Y or $-\sigma_Y$ (Table 2.6). Analogously, one can deduce the magnitudes and signs of all α and σ influences for a pathway if Steps 1 through 6 of the method are successful for a given dataset. In Table 2.8 I provide the calculation for each of these influences from β parameters obtained from experimental data. The calculation of each influence is topology-specific with the exception of ρ_0 and σ_S. This is not surprisingly as these are topology-independent variables in the model of Figure 2.3.

Table 2.8: Definitions of α and σ parameters. Parameters can be calculated from β parameters estimated from experimental data, when experimental data allow the inference of topology and gene order. This Table describes the calculation when gene i is inferred as the upstream gene. Topology numbers (1) through (4) are as defined in Table 2.7

	(1)	**(2)**
ρ_0	$\beta_0 + \beta_I^0 + \beta_i^0 + \beta_j^0$	$\beta_0 + \beta_I^0 + \beta_i^0 + \beta_j^0$
α_X	$-\beta_I^0 - \beta_i^0$	$-\beta_I^1 - \beta_i^1$
α_Y	$-\beta_I^0 - \beta_j^0$	$-\beta_I^1 - \beta_j^1$
α_I	β_I^0	β_I^1
σ_S	$\beta_0^1 - \beta_I^0 + \beta_I^1 - \beta_i^0 + \beta_i^1 - \beta_j^0 + \beta_j^1$	$\beta_0^1 - \beta_I^0 + \beta_I^1 - \beta_i^0 + \beta_i^1 - \beta_j^0 + \beta_j^1$
σ_X	$\beta_I^0 - \beta_I^1 + \beta_i^0 - \beta_i^1$	$\beta_I^1 - \beta_I^0 - \beta_i^0 + \beta_i^1$
σ_Y	$\beta_j^0 - \beta_j^1$	$\beta_j^1 - \beta_j^0$
	(3)	**(4)**
ρ_0	$\beta_0 + \beta_I^0 + \beta_i^0 + \beta_j^0$	$\beta_0 + \beta_I^0 + \beta_i^0 + \beta_j^0$
α_X	$-\beta_I^0 - \beta_i^0$	$-\beta_I^1 - \beta_i^1$
α_Y	$-\beta_I^0 - \beta_j^1$	$-\beta_I^1 - \beta_j^0$
α_I	β_I^0	β_I^1
σ_S	$\beta_0^1 - \beta_I^0 + \beta_I^1 - \beta_i^0 + \beta_i^1 - \beta_j^0 + \beta_j^1$	$\beta_0^1 - \beta_I^0 + \beta_I^1 - \beta_i^0 + \beta_i^1 - \beta_j^0 + \beta_j^1$
σ_X	$\beta_I^0 - \beta_I^1 + \beta_i^0 - \beta_i^1$	$\beta_I^1 - \beta_I^0 - \beta_i^0 + \beta_i^1$
σ_Y	$\beta_j^1 - \beta_j^0$	$\beta_j^0 - \beta_j^1$

In this section, I have defined criteria which theoretically should allow one to infer one of sixteen possible pathways from experimental data. Inference requires the estimation of β and δ parameters by linear regression of experimental traits, followed by a stepwise analysis of these parameters. In the following section, I apply this algorithm to a well-characterized metabolic and gene-regulatory network in *S. cerevisiae*.

2.3.6 Inference method benchmarking

To benchmark the method described in the previous section, I consider an experimental dataset [61] obtained for strains of *S. cerevisiae* having individual or double deletions of eight *GAL* genes. This highly connected set of genes is considered one of the best understood eukaryotic networks. *GAL* genes encode functions for galactose-induced transcription (Figure 2.4B) or galactose metabolism (Figure 2.4A). The dataset consists of experimental replicate phenotype measurements for wildtype and gene deletion strains cultured in presence or absence of galactose (2% v/v) as the stimulus.

The dataset consists of two alternative quantitative phenotype measurements for all strains and stimulus conditions. Rate of exponential population growth (min^{-1}) is estimated from culture absorbance measurements over time, whereas mean cellular expression level (arbitrary fluorescence units) is estimated from flow cytometry measurements of cellular fluorescence by flow cytometry. Wildtype and *GAL* gene deletion strains have a $P_{GAL10}-yEGFP$ expression cassette at the *ade2* locus (Figure 2.4C), allowing estimation of galactose-induced transcription. Details of cell culture conditions and trait measurements are provided in Methods (§2.5.1).

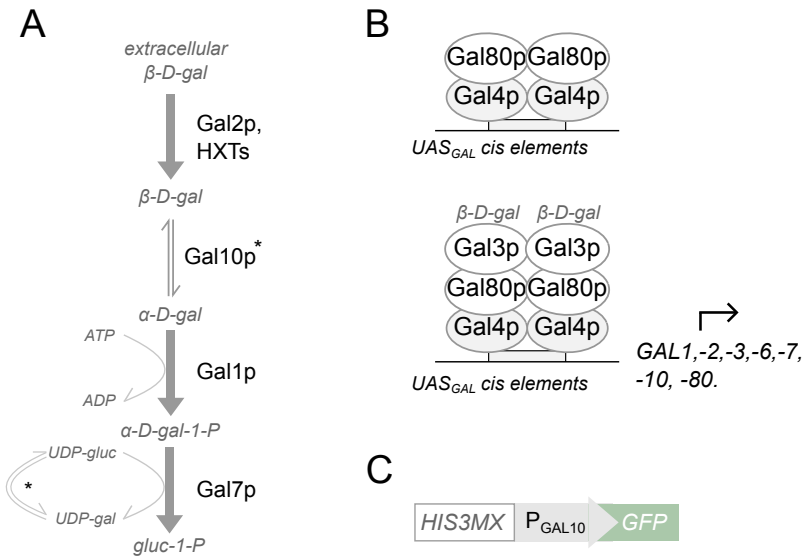

Figure 2.4: Schematic of intracellular functions encoded by eight *GAL* genes. **A.** Model of Leloir metabolic pathway adapted from [63]. Gal2p permease or other hexose transporters (HXTs) import galactose. Gal10p mutarotase and epimerase (*) [64], Gal1p galactokinase, and Gal7p uridyl-transferase activities sequentially convert intracellular galactose to glucose-1-phosphate. **B.** Model of galactose-induced transcription adapted from [65]. Gal3p, if bound to intracellular galactose and ATP (not shown), relieves Gal80p inhibition of DNA-binding protein Gal4p, allowing de-repression of transcription of seven *GAL* genes. The model considers protein dimerization of Gal80p, Gal4p and Gal3p, as well as two positive (Gal2p, Gal3p) and one negative (Gal80p) feedback loops on transcription. Gal6p is hypothesized to negatively regulate transcription of *GAL1/10*, *GAL7* and *GAL2* [66] (not shown). Symbols: gal, galactose; gluc, glucose; -P, -phosphate.

The phenotypes of individual gene deletions (Figure 2.5) deviate from the Avery and Wasserman assumptions. Such deviations justify the need for the inference method developed. First, phenotypes of the various genotypes cannot be categorized into qualitative groups, regardless of whether growth or (Figure 2.5A) or expression (Figure 2.5 B) is the phenotype. This finding indicates the requirements to have quantitative predictions for topology inference, and to assign statistically significant deviation, relative to wildtype, for categorizing whether a deletion has an *effect* on a given phenotype in a given condition. I test the assignment of significance by two methods, t-test ($p < 0.05$) and 90% confidence intervals (Figure 2.5). While the confidence intervals allow more conservative assignment of significance, most deletion effects are significant by both measures.

The above-described analysis of statistical significance reveals a second deviation from the Avery and Wasserman assumptions, that is the finding that gene deletions have effects in both stimulus conditions. For the growth phenotype, five of eight genes have significant effects in both conditions (t-test, Figure 2.5A). For the expression phenotype, deletion of *GAL80* has a significant effect in both conditions (t-test, Figure 2.5B).

Growth phenotypes in the dataset are therefore more likely to deviate from Avery and Wasserman's assumptions than expression for the same genotypes. This may due to growth reporting on less specific biological processes compared to expression, or due to bistability of the *GAL* gene regulatory network [67] regulating expression. The conditions of galactose in the dataset correspond to concentrations wherein the network is estimated to be monostable *OFF* or monostable *ON* [68], respectively, which may reduce sensitivity to non-specific genetic perturbation. Notably, bistability is hypothesized to exemplify developmental regulatory networks [69], for which Avery and Wasserman's rules were contextualized [55].

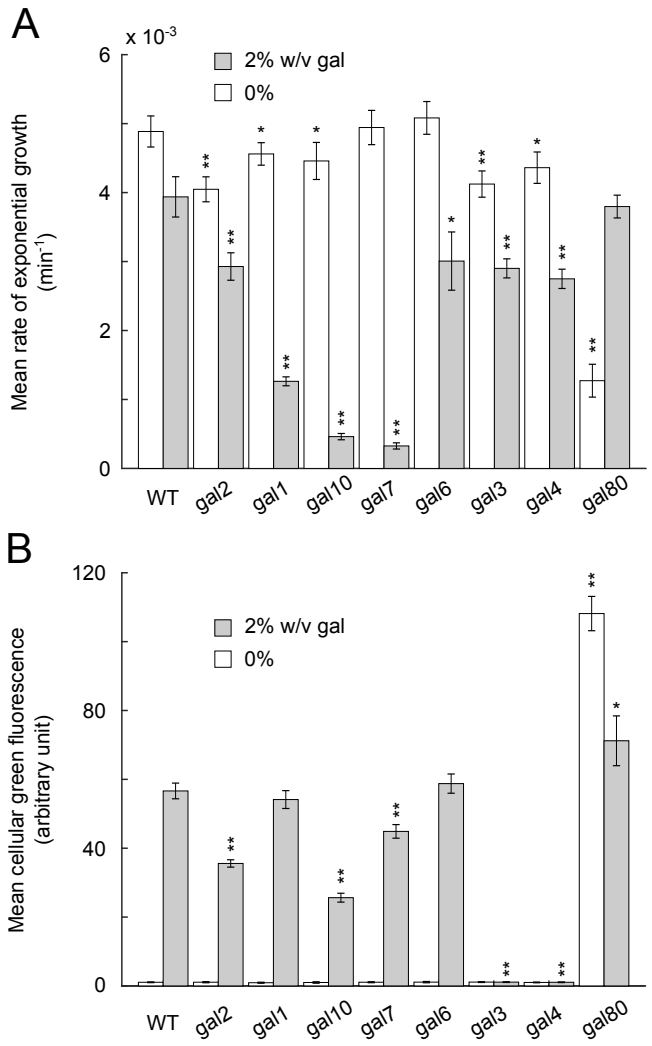

Figure 2.5: *GAL* gene deletion phenotypes deviate from Avery and Wasserman assumptions. We find two deviations. First, deletions yield changes in phenotype which cannot obviously be categorized into distinct qualitative groups. Second, some second deletions have effects in both conditions of the stimulus. Replicate growth rate data (**A**) or gene expression data (**B**) of each single *GAL* gene deletion strain are compared to wildtype data in the corresponding condition and phenotype to determine if deletion has a significant effect. We indicate if both (**) 90% confidence intervals (90% CI $= \mu \pm \sigma * z$, $z = 1.645$) and t-test ($p < 0.05$) result in significance, or t-test alone (*). Symbols: gal, galactose; σ, standard deviation; bar height, mean (μ) of replicate data; error bar, σ.

To benchmark the method, I consider 28 true positive topologies. These topologies, schematized in Table 2.9 correspond to one topology per gene pair in the dataset, obtained from previous literature. The regulatory edges may be direct (e.g. Gal80p regulation of Gal4p) or indirect (e.g. Gal80p regulation of Gal1p). I exclude multiple topologies possible when genes have feedback edges or multiple functions. For example, Gal10p mutarotase and epimerase functions are required in different steps of the metabolic pathway, yet we consider the mutarotase function (*GAL10* upstream of *GAL1*) to be the true positive because this function occurs most upstream (Figure 2.4A). I consider true positives to be the same for both phenotypes. While assigning the regulatory edge from the downstream gene to ρ is clearly defined by positive or negative gene regulatory functions for the expression phenotype, the assignment is less clear for growth since I anticipate that gene perturbation effects on growth may occur through many alternative bioprocesses involving both positive and negative regulation. To obtain true positive edges in this case, I consider gene i to be an activator of ρ if the literature-derived function of gene i is to enhance galactose-induced gene expression or galactose metabolism, and otherwise I consider gene i to be a repressor. With these criteria, all *GAL* genes with the exception of *GAL6* and *GAL80* are considered activators of ρ, and the edge from the downstream gene to ρ in the true positives is consistent for both phenotypes.

To benchmark the above-described algorithm I analyze each step of the algorithm towards inferring these true positive interactions. I first estimate δ parameters for each pair by linear regression (Step 1). Steps 2 and 4 of the method require criteria for determining if regression parameters are significantly different from zero. I consider parameters are significant with a p-value less than 0.05 (t-statistic). I ensured p-values were not related to failure of the assumptions of linear regression, by inspection of residuals.

By this criterion of significance, more gene pairs can be analyzed for growth phenotypes than for expression. For growth, δ_i is significant for all genes, and δ_I is significant for 23 out of 28 gene pairs (Figure 2.6A). The five gene pairs with insignificant δ_I are *GAL3/GAL80*, *GAL6/GAL80*, *GAL10/GAL80*, *GAL2/GAL3*, and *GAL2/GAL4*. For expression, 14 of 28 gene pairs can be analyzed (Figure 2.6B), as $\delta_i = 0$ for *GAL1* and *GAL6* genes, thereby excluding any gene pairs including these genes for further analyses.

Among gene pairs which can be analyzed in Step 2 (i.e. $\delta_i, \delta_j, \delta_I \neq 0$), I tested if epistasis could be detected. The identification of epistasis, i.e. $\delta_i = -\delta_I$ or $\delta_j = -\delta_I$, requires a criterion for equality. While many criteria are possible, I consider the assumption that a threshold for epistasis detection should be based on measurement noise and parameter estimation error of regression, and that the identification of epistasis should not be biased towards datasets with high measurement noise. This allows a single threshold for each phenotypic dataset, determined from the data and in this sense is not biased or arbitrary. I first analyze the data with these assumptions, in later sections, I will analyze how the threshold for epistasis detection influences the rate of true positive topology inference.

With these assumptions, gene i is identified as epistatic to gene j (i.e. $\delta_j = -\delta_I$) when

$$\frac{|\delta_j + \delta_I|}{max(|\delta_j|, |\delta_I|)} \leq \mu_{rse}, \tag{2.7}$$

where μ_{rse} is the mean of all relative standard errors (rse) among significant δ_I values for a given phenotype, $rse = se_{\delta_I}/|\delta_I|$, and se_{δ_I} is the standard error of δ_I obtained by regression analysis.

With this criterion for epistasis detection, the majority of significant gene interactions in the network exhibit epistasis. For the growth phenotype, the epistasis threshold is a deviation less than 13% ($\mu_{rse} = 0.13$), allowing 18 of 23 gene pairs to be considered epistatic (Figure 2.6A). For expression, the epistasis threshold is a deviation less than 20% ($\mu_{rse} = 0.20$), allowing six of 14 gene pairs to be considered epistatic (Figure 2.6B) and three of 14 gene pairs to be considered co-equal.

Co-equality reflects the case where $-\delta_I = \delta_i = \delta_j$ and therefore neither gene may be inferred as upstream (see Step 2 of method). In the model (Figure 2.3, co-equality by this measure is predicted to occur when the upstream gene influences the phenotype solely by influencing the downstream gene (i.e. $\sigma_X = 0$, Table 2.6). St Onge et al. [43] identified co-equality analogously, from the analysis of sensitivity phenotypes, and found 9/10 co-equal identifications in their dataset were gene pairs were documented to function in the same physical complex. Indeed, the three co-equal relationships identified in the present dataset correspond to those among genes *GAL3*, *GAL4* and *GAL80*, which are hypothesized to function as a complex [65] (illustrated in Figure 2.4B). Notably, co-equality for these genes is only identified by expression phenotypes. For growth phenotypes, the method infers that these genes either do not exhibit epistasis (*GAL3/GAL80*) or exhibit epistasis rather than co-equality (*GAL4/GAL80*,*GAL4/GAL3*).

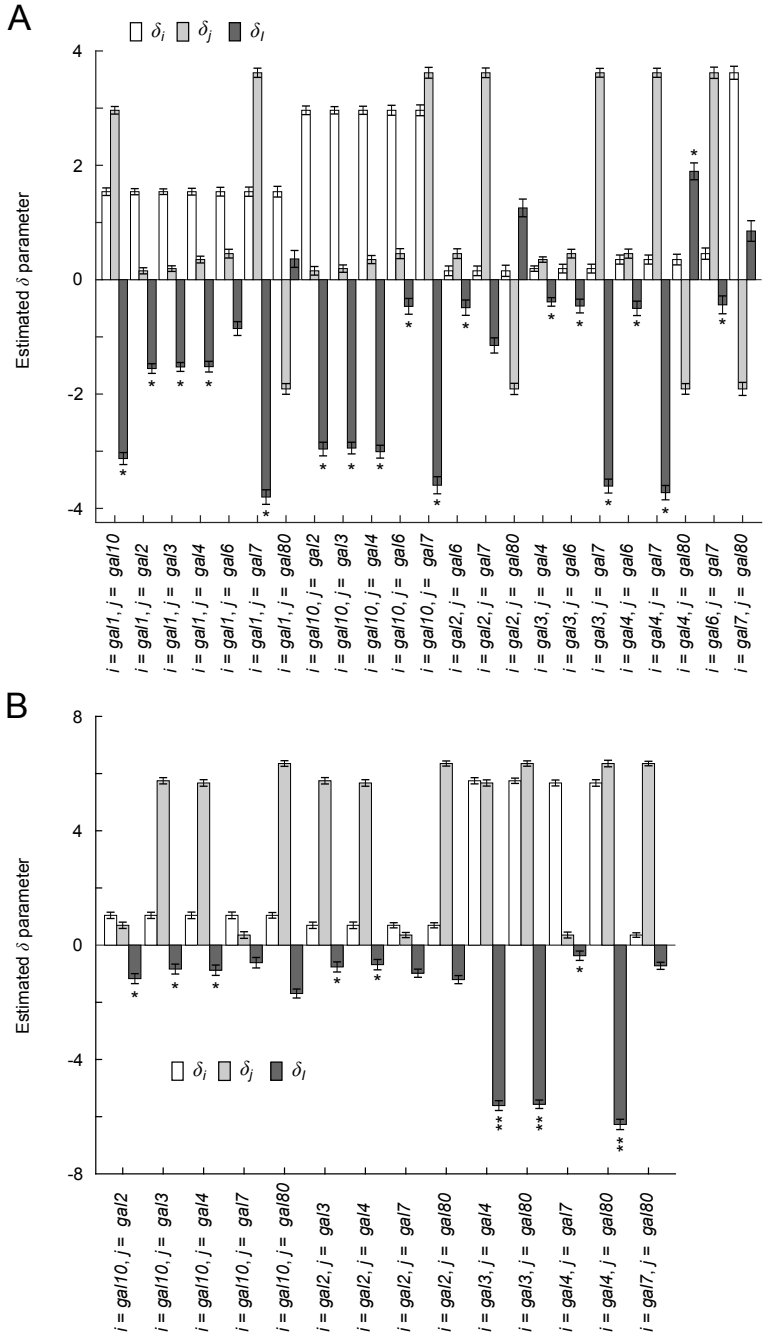

Figure 2.6: Stimulus-dependent gene interactions and epistasis among pairs of *GAL* genes. δ parameters for gene pairs wherein all three delta parameters are significant (p-value < 0.05) are shown for growth (**A**) and expression (**B**) phenotypes. Symbols: error bar, standard error; *, epistasis is detected by Eq 2.7; **, co-equality is detected by Eq 2.7

Step 4 of the method provides means to infer gene order even when genes exhibit co-equality. Following Step 3, i.e. estimation of β parameters by linear regression and calculation of β' parameters required for topology inference, I analyzed the extent to which co-equal gene pairs could be ordered based on β' parameters. Indeed, two of three such gene pairs can be ordered, namely *GAL3* is inferred upstream of *GAL80*, and *GAL80* upstream of *GAL4*. One of three gene pairs does not allow gene order to be inferred, namely *GAL3* and *GAL4* because all β' terms are non-zero in the same stimulus condition, characteristic of Topology 1 or 2 (Table 2.7). Therefore a total of 18 gene pairs can be analyzed by Step 5 for growth, and a total of 8 gene pairs for expression.

Following ordering of β' parameters by Step 5, I determined if patterns among ordered β' parameters matched one of four pathways (Step 6), and used this information to identify the sign of the edge corresponding to σ_Y in the model (Step 7). In the case that topology could be inferred, I compared the inferred gene order and topology to the corresponding true positive, for each phenotype (Table 2.9).

Table 2.9: Schematic of *GAL* gene topologies and their inference. Inference of gene order (**A**) and topology (**B**) when growth rate is the trait. Inference of gene order (**C**) and topology (**D**) when gene expression is the trait. Results reflect inference when the equivalence threshold is set to the mean rse (see Methods). Symbols: S, signal; T, phenotypic trait; ✓, correct inference; ✓*, correct inference of gene order by β' terms; ✗, wrong inference; blank, inconclusive inference.

n	Topology	A	B	C	D
1	$S \rightarrow GAL10 \rightarrow GAL1 \rightarrow T$	✗	✓		
2	$S \rightarrow GAL2 \rightarrow GAL1 \rightarrow T$	✓	✓		
3	$S \rightarrow GAL3 \rightarrow GAL1 \rightarrow T$	✓	✓		
4	$S \rightarrow GAL4 \rightarrow GAL1 \rightarrow T$	✓	✓		
5	$S \rightarrow GAL6 \dashv GAL1 \rightarrow T$				
6	$S \rightarrow GAL1 \rightarrow GAL7 \rightarrow T$	✓	✓		
7	$S \dashv GAL80 \dashv GAL1 \rightarrow T$				
8	$S \rightarrow GAL2 \rightarrow GAL10 \rightarrow T$	✓	✓	✓	✓
9	$S \rightarrow GAL3 \rightarrow GAL10 \rightarrow T$	✓	✓	✓	✓
10	$S \rightarrow GAL4 \rightarrow GAL10 \rightarrow T$	✓	✓	✓	✓
11	$S \rightarrow GAL6 \dashv GAL10 \rightarrow T$	✗	✗		
12	$S \rightarrow GAL10 \rightarrow GAL7 \rightarrow T$	✓	✓		
13	$S \dashv GAL80 \dashv GAL10 \rightarrow T$				
14	$S \rightarrow GAL2 \rightarrow GAL3 \rightarrow T$			✗	✓
15	$S \rightarrow GAL2 \rightarrow GAL4 \rightarrow T$			✗	✓
16	$S \rightarrow GAL2 \rightarrow GAL6 \dashv T$	✓	✗		
17	$S \rightarrow GAL2 \rightarrow GAL7 \rightarrow T$				
18	$S \rightarrow GAL2 \dashv GAL80 \dashv T$				
19	$S \rightarrow GAL3 \rightarrow GAL4 \rightarrow T$	✓	✓		✓
20	$S \rightarrow GAL3 \rightarrow GAL6 \dashv T$	✓	✗		
21	$S \rightarrow GAL3 \rightarrow GAL7 \rightarrow T$	✓	✓		
22	$S \rightarrow GAL3 \dashv GAL80 \dashv T$			✓*	✓
23	$S \rightarrow GAL4 \rightarrow GAL6 \dashv T$	✓	✗		
24	$S \rightarrow GAL4 \rightarrow GAL7 \rightarrow T$	✓	✓	✓	✓
25	$S \dashv GAL80 \dashv GAL4 \rightarrow T$			✓*	✓
26	$S \rightarrow GAL6 \dashv GAL7 \rightarrow T$	✗	✗		
27	$S \dashv GAL80 \dashv GAL6 \dashv T$				
28	$S \dashv GAL80 \dashv GAL7 \rightarrow T$				

For growth phenotypes (Table 2.9A,B), 17 of 18 epistatic gene pairs have β' parameters which match the pattern of one of four topologies. The gene pair corresponding to *GAL4* and *GAL80* has parameters inconsistent with any topology we model. Therefore a total of 11/28 gene pairs (40%) are false negatives for one of many reasons, including non-significant δ parameters (5/11), no epistasis detected (5/11), or inconsistent β' parameter pattern (1/11). Among the gene pairs where an inference could be made, $\sim$ 65% correspond to correct gene order and topology.

The majority of wrong inferences (5/6) correspond to gene pairs with *GAL6*. No topology involving *GAL6* is inferred correctly, either the inference is wrong or no inference can be made. The function of *GAL6* is therefore not supported by the present analysis or the dataset. I note that *GAL6* is not as extensively characterized as the other *GAL* genes considered. The previous study from which I based determination of true positive topology [66] involving *GAL6* had conducted analyses of transcript levels of *GAL2,GAL1,GAL7* (Northern blot) or reporter expression driven from a $P_{GAL1/GAL10}$ promoter, and found in all cases transcripts were higher when *GAL6* was deleted relative to wildtype, allowing the conclusion that *GAL6* is a repressor of transcription of these metabolic genes of galactose metabolism. However, the measurements in the previous study were made between 20 and 36 h following change in growth medium from a glucose to galactose carbon source. This contrasts the dataset I analyze, wherein reporter expression was measured at 6 h following a change in growth medium from a raffinose to raffinose plus galactose carbon source. Additionally, the $P_{GAL1/GAL10}$ is bidirectional, and the strains in the dataset I consider correspond to the direction of expression of *GAL10* whereas the previous study had used the direction corresponding to expression of *GAL1*. The inability to infer previously suggested functions of *GAL6* may therefore be explained by differing experimental designs.

If I exclude topologies with *GAL6* as true positives, only one error in inference is detected by analysis of growth phenotypes. I note that this error is nonetheless still consistent with known biology of the network, namely that *GAL10* encodes a protein with both mutarotase and epimerase activities. Whereas mutarotase activity is a metabolic step upstream of *GAL1*, epimerase activity is downstream of *GAL1*. The method infers the latter scenario (Table 2.9A). Had I considered this alternative gene order as the true positive, there would be zero false inferences based on growth, and 57% true positives inferred. I note that the remaining false negatives often correspond to topologies with *GAL80*, none of which can be inferred by growth phenotype.

In contrast, a subset of topologies involving *GAL80* can be inferred correctly by expression phenotypes. These topologies correspond to the correct inference among genes *GAL3*, *GAL4* and *GAL80*. Of the eight gene pairs for which gene order and topology could be inferred, most are correct (6/8). I note that the two errors, corresponding to wrong order among *GAL2*, *GAL3* and *GAL4* are nonetheless consistent with known biological functions of these genes, which include feedback loops. Specifically, *GAL2* encoded galactose permease activity is upstream of *GAL3* and *GAL4*, however *GAL2* is also downstream of *GAL3* and *GAL4* since the activities of these genes are required for transcription of *GAL2*. The method infers the latter scenarios. Had I considered these alternative gene orders as true positives, there would be zero false inferences based on expression, and 38% true positives inferred when topologies with *GAL6* are excluded. False negatives (13) are due to non-significant δ parameters (7/13), failure to detect epistasis (5/13) or to resolve co-equality (1/13) by analysis of β' parameters.

2.3.7 Impact of threshold for epistasis detection

The results reported in the previous section are based on detection of epistasis by a threshold determined using Eq 2.7. I anticipate that modifying the threshold may substantially change these results. To analyze the effect of epistasis threshold on inference, I re-analyzed the data at 21 possible thresholds representing increments of 5% in the range of 0% deviation among δ parameters to 100% (Figure 2.7), and calculated the number of true positive topologies that are inferred correctly.

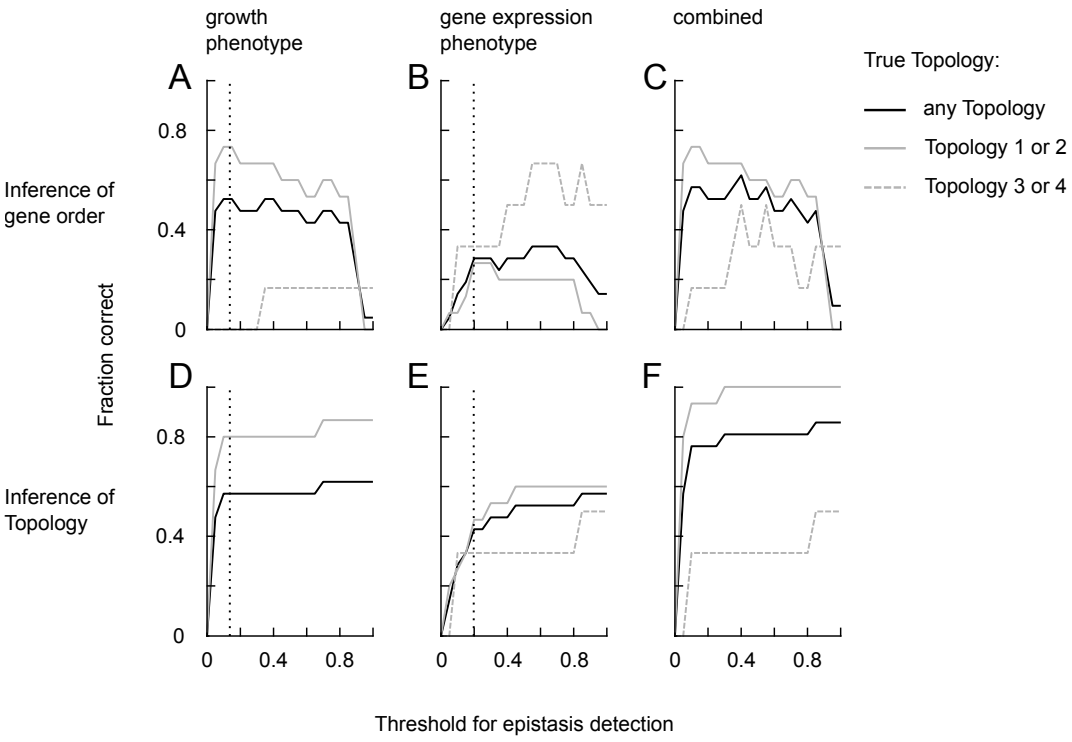

Figure 2.7: Phenotype influences topology and gene order inference success. Fraction of topologies with gene order correctly inferred based on growth (**A**), gene expression (**B**), or combined results (**C**). Fraction of topologies with edge signs inferred correctly based on growth (**D**), gene expression (**E**), or combined results (**F**). True positives are illustrated in Table 2.9, excluding any topology with *GAL6*. Symbols: vertical dotted line, equivalence threshold determined by Eq 2.7 with corresponding inferences in Table 2.9A-D.

This analysis of epistasis threshold modification illustrates several trends. First, no inference can be made when the threshold is zero (Figure 2.7A-F). This indicates measurement and biological noise must be taken into account when analyzing phenotypic equivalence. Second, I find the inference of topology can only be improved as the threshold is increased (Figure 2.7D-F). In contrast, I find the correlation between correct inference of gene order and equivalence threshold depends on the type of gene regulation (Figure 2.7A-C). For negative inter-gene regulation (Topology 3 or 4) the number of correct inferences increases as a function of threshold, whereas for positive inter-gene regulation (Topology 3 or 4) the number of correct inferences is biphasic, having a maximal chance of correct gene order inference when the threshold is at an intermediate level between zero and 100%. This may be explained by increased thresholds corresponding to increased rates of identifying co-equality ($\delta_I = -\delta_i = -\delta_j$, which can be resolved to determine gene order by alternatively analyzing β' parameters in topologies with negative inter-gene regulation but not positive regulation (Step 4 of method). Even in case of positive inter-gene regulation, identification of correct gene order is not sensitive to changes in threshold in the range between $\sim$ 10 and 80% deviation (Figure 2.7A-C).

The analysis of epistasis threshold modification additionally shows that phenotype influences inference success. Specifically, correct topology and gene order are more likely to be inferred for positive inter-gene regulatory topologies when growth is the phenotype, whereas correct inferences for negative inter-gene regulatory topologies are more likely when expression is the phenotype. Notably, this trend was also present for a single threshold (Table 2.9), however, the modification of threshold indicates this trend is threshold-independent. The trend may, however, be specific to the genes used for benchmarking. Specifically, regulation among genes encoding metabolic functions is positive, and deletions of these genes have greater effect on growth, whereas regulation among transcriptional regulators is primarily negative,

and deletions of these genes have greater effects on the expression phenotype. All together, the highest chances of correctly inferring gene order and topology are obtained by combining the results from both phenotypes (Figure 2.7C,F).

2.3.8 Deducing a network from pairwise inferences

By considering the inference of gene order and topology from combined results from growth and expression phenotypes, the highest rates of correct inferences are near the threshold based on uncertainty in δ_I values (Eq 2.7, Table 2.9). At this threshold, for each respective phenotype, the majority of errors correspond to false negatives, wherein no inference can be made. Upon examination of the gene pairs, I find false negatives primarily correspond to indirect regulatory relationships. For example, the indirect regulation of *GAL1*, *GAL7* or *GAL10* transcription by Gal80p cannot be inferred from analysis of either phenotype (Table 2.9A-D). There is a possibility that such indirect relationships could be deduced by combining all inferences corresponding to direct relationships in one network.

To analyze the extent to which these indirect relationships could be deduced from inferences the method can make, I incorporated all edges inferred in Table 2.9A-B or Table 2.9C-D into a large gene network N. To simplify these networks, I examined the transitive reduction of N (N_R). Transitive reduction maintains reachability among nodes in the original network, while minimizing the number of edges. While N_R obtained for each individual phenotype (Figure 2.8A-B) fails to capture all true positive relationships, N_R obtained from N wherein edges from inferences from both phenotypes does contain nearly all true positive relationships (Figure 2.8C). If I consider alternative orders of *GAL2* relative to genes with transcriptional regulatory functions, alternative order of *GAL10* explained by epimerase function, and exclude

topologies with *GAL6*, one can recover all gene pair inferences correctly without false negatives.

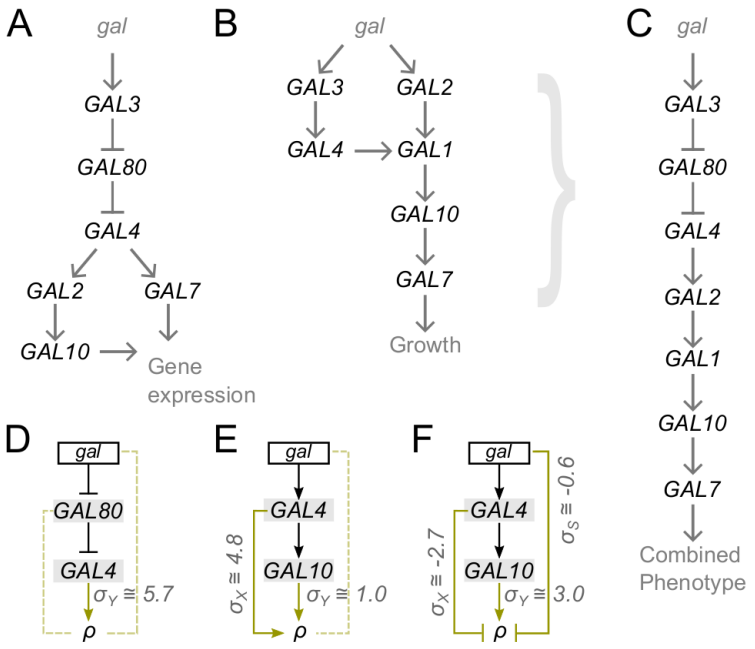

Figure 2.8: *GAL* gene networks are obtained by transitive reduction. A Transitive reduction of graph consisting of inferences of Table 2.9A&B. **B** Transitive reduction of graph consisting of inferences of Table 2.9C&D. **C** Transitive reduction of graph consisting of inferences of Table 2.9A-D. **D-F.** Quantitative influences deduced from inferred topologies may be disagree with networks obtained by transitive reduction. The topology inferred for *GAL80* and *GAL4*, when ρ is the expression phenotype (**D**), does not disagree since feedforward influences σ_S and σ_X are insignificant (95% confidence interval). For the same phenotype, the topology for *GAL4* and *GAL10* (**E**) would be in disagreement since the majority of influence of the signal is propagated by the feedforward influence of *GAL4* (σ_X). For this gene pair wherein growth is the phenotype, the inferred topology is the same but the quantitative influences are different in magnitude and sign (**F**). σ parameters are calculated from Table 2.8. Symbols: *gal*, galactose stimulus; dotted line, influence that is not significantly different from zero (95% confidence interval).

Transitive reduction nonetheless poses problems in the assumptions to derive N_R in relation to the assumptions used to infer topologies for gene pairs. Specifically, for the inference of topologies wherein the upstream gene is an activator (Topology 1 or 2), the feedforward loop from the upstream gene to the phenotype must be non-zero in order to infer gene order (i.e. $\sigma_X \neq 0$, Tables 2.6 and 2.7). In contrast, the assumptions of transitive reduction allow one to obtain a network by minimizing edge count while maintaining reachability, therefore removing every feedforward loop.

The contradiction between the assumptions of the inference method and transitive reduction may depend on the gene pair. Because the method allows for influences in inferred topologies to be quantified, one can assess the extent to which a feedforward influence is required to explain a given set of perturbation phenotypes. For example, in the case of the topology of *GAL4* and *GAL80* inferred from expression phenotypes, signal-dependent feedforward influences are insignificant ($\sigma_X, \sigma_S \approx 0$, 95% confidence interval), therefore agreeing with the assumption of transitive reduction. In contrast, for the topology of *GAL4* and *GAL10*, the feedforward influence on expression (σ_X) is over four fold greater than the influence of the linear pathway (σ_Y), and therefore a network obtained by transitive reduction would fail to accurately represent the influences among genes and the phenotype. A further complication may be that feedforward influences of inferred topologies are different in sign and magnitude when obtained from different phenotype data. This is exemplified by the topologies inferred for *GAL4* and *GAL10* (Figure 2.8E-F). Such conflicts complicate the formation of a network which combines inferences from both phenotypes (for e.g. Figure 2.8C). Notably, to form the network in Figure 2.8C, I excluded any edges wherein there was disagreement between the inference of the two phenotypes.

2.4 Discussion

In this chapter, I described a topology model (§2.3.1) developed in [61] that allows the interpretation of quantitative epistasis even when previously described models are not applicable. Specifically, because this model allows genes to have basal activity, the model allows interpretation of phenotypes that cannot be analyzed by Avery and Wasserman's rules nor Aylor and Zeng's method. Additionally, because the model formalizes the additional two possible feedforward influences of genes on a phenotype, the model allows interpretation of gene order in positive regulatory pathways that cannot be achieved with Aylor and Zeng's method.

By formalizing the assumptions to derive theoretical phenotype expressions from the topology model in §2.3.1, I showed how rules for inference of topology and gene order emerge. This analysis increases knowledge about the assumptions that may be consistent in different previously published methods for analyzing epistasis. For example, I showed how the rules of Avery and Wasserman and St Onge et al. are related. Specifically, both sets of distinct rules can be derived from the model described in §2.3.1, but for distinct phenotypic parameters. To isolate the rules of Avery and Wasserman, one must examine β' parameters, whereas to isolate the rules of St Onge et al. one must examine δ parameters. To maximize the ability to infer gene order and topology, the inference method I describe in §2.3.5 incorporates the analysis of both β' and δ parameters.

In addition to allowing such clarification, the model described increases the applicability of gene topology inference to quantitative genetic interaction data. Specifically, no previous methods had allowed genes to have basal activity and therefore significant effects of gene perturbation independent of the stimulus condition. The analysis I describe is applicable in this case, and recent quantitative datasets indicate that these effects are prevalent in quantitative datasets. In general, I assume that quantitative

phenotypes are more sensitive to smaller perturbations than qualitative phenotypes. In addition, the previous method of Aylor and Zeng could not be applied to infer gene order in positive regulatory pathways, whereas the method I describe can allow such an inference.

While a subset of assumptions are modified, the model of §2.3.1 maintains several assumptions originally described by Avery and Wasserman, and extended by Aylor and Zeng. Specifically, the signal and genes are assumed to function in a hierarchical topology. In a hierarchical topology, each gene activity is assumed to be determined by a single upstream input. Because I did not test modification of these assumptions in the present chapter, the extent to which the theoretical phenotype patterns I describe may be modified or mimicked by a model incorporating alternative assumptions is not clear. I test modification of the assumption that topology is hierarchical in the following chapter of this thesis.

Because several assumptions of classical epistasis analysis are maintained, the method is limited in similar ways. For example, because of the assumption that gene perturbation causes full loss gene function, one cannot infer topologies with feedback edges. Similarly, because only two genes are considered per topology, one cannot infer multiple alternative topologies among the same genes, which may be possible when genes encode multiple functions corresponding to, for example, multiple protein domains or enzymatic active sites, or feedback loops.

I illustrated these limitations in the context of the experimental dataset used to benchmark the inference method. The network of eight *GAL* genes I consider is documented to contain at least three feedback loops, and one gene with multiple enzymatic functions (*GAL10*). To isolate true positive topologies for the purpose of benchmarking, I assumed that among the possible topologies that may be inferred for two genes, the true positive consists of the topology having the order and regulatory of

functions of the genes which is documented to occur first in the network after addition of the stimulus (i.e. extra-cellular galactose). In the inference results, however, I found the rule was not consistent, wherein a subset of our true positives having multiple possibilities was correctly inferred (*GAL80/GAL4*, *GAL80/GAL3*) but not the other subset (*GAL2/GAL4*,*GAL2/GAL3*,*GAL1/GAL10*).

False negatives often corresponded to indirect regulation between genes in the known network. To determine if these edges could be recovered, I integrated all pairwise inferences into one network, followed by transitive reduction of this network. With this analysis, I found one could obtain nearly all true positives when the network incorporates edges inferred from both phenotypes. This also illustrated a number of problems that arise when integrating pairwise data into a global network. In particular, transitive reduction automatically removes all feedforward edges in a network. In contrast, the inference method requires a feedforward edge to infer gene order whenever the inter-regulatory function is positive, i.e. most inferences for the *GAL* network. In addition, the calculation of feedforward influence σ_X indicates that for some pathways, the magnitude of the feedforward influence contributes more to the phenotype than the influence that propagates through the downstream gene (σ_Y), indicating that feedforward is essential to accurately reflect the dataset. In these cases, transitive reduction is limited as it makes no attempt to assess when a feedforward loop should be maintained or not in a larger network.

Battle et al. [60] developed a method to obtain large networks from genetic interaction data that may address part of this problem. In particular, presence of a feedforward edge in a larger network is maintained based on phenotypic expectations for this topology relative to a topology lacking a feedforward loop, thus allowing the formation of network structure based on agreement with experimental data. I explore the utility of Battle et al.'s method in the next chapter of this thesis.

2.5 Methodology

2.5.1 Strain generation

All strains in the library are derived from haploid *S.cerevisiae* strain BY4742 [70] (Open Biosystems), which I refer to as the autofluorescence strain. The wildtype strain was generated by replacing the *ade2* locus of BY4742 with an expression cassette reporting on galactose-induced transcription. The cassette has the transcription promoter (-1 to -668 bp relative to *GAL*10 start codon) of the *GAL10* gene (P_{GAL10}) fused upstream of DNA encoding green fluorescent protein (yEGFP3) [71] and the terminator of *ADH1* (T_{ADH1}, +922 to +1213 bp relative to *ADH*1 start codon +1), downstream of DNA encoding a histidine expression cassette (P_{HIS3}-*HIS3*-T_{HIS3}) (-188 to +864 bp). The *HIS3* expression cassette and reporter cassette were assembled in a plasmid by standard restriction digest and bacterial cloning protocols, by Simon St-Pierre. Single gene deletion strains were generated by gene replacement of entire open reading frames (ORFs) of the wildtype. Combinatorial gene deletion strains were generated by gene replacement of ORFs of a single gene deletion strain having a different *GAL* gene deleted. Specifically, the first *GAL* gene was replaced with a Geneticin drug resistance cassette (P_{TEF1}-*KANMX*-T_{TEF1}), and the second gene with a cassette conferring uracil prototrophy (P_{TEF1}-*URA3*-T_{TEF1}). Gene replacement was conducted by standard methods as described in [15]. Briefly, cells were transformed with DNA consisting of the replacement cassette with 40 bp on the 5' and 3' flanking ends homologous to the 40bp region directly upstream and downstream, respectively of the ORF to be replaced. DNA for transformation was obtained by PCR amplification of plasmid DNA with 60 bp primers. A detailed description of the transformation protocol used is provided as supplementary information in [72].

DNA sequences

yEGFP3 from [71]:

ATGTCTAAAGGTGAAGAATTATTCACTGGTGTTGTCCCAA
TTTTGGTTGAATTAGATGGTGATGTTAATGGTCACAAATT
TTCTGTCTCCGGTGAAGGTGAAGGTGATGCTACTTACGGT
AAATTGACCTTAAAATTTATTTGTACTACTGGTAAATTGC
CAGTTCCATGGCCAACCTTAGTCACTACTTTCGGTTATGG
TGTTCAATGTTTTGCTAGATACCCAGATCATATGAAACAA
CATGACTTTTTCAAGTCTGCCATGCCAGAAGGTTATGTTC
AAGAAAGAACTATTTTTTTCAAAGATGACGGTAACTACAA
GACCAGAGCTGAAGTCAAGTTTGAAGGTGATACCTTAGTT
AATAGAATCGAATTAAAAGGTATTGATTTTAAAGAAGATG
GTAACATTTTAGGTCACAAATTGGAATACAACTATAACTC
TCACAATGTTTACATCATGGCTGACAAACAAAAGAATGGT
ATCAAAGTTAACTTCAAAATTAGACACAACATTGAAGATG
GTTCTGTTCAATTAGCTGACCATTATCAACAAAATACTCC
AATTGGTGATGGTCCAGTCTTGTTACCAGACAACCATTAC
TTATCCACTCAATCTGCCTTATCCAAAGATCCAAACGAAA
AGAGAGACCACATGGTCTTGTTAGAATTTGTTACTGCTGC
TGGTATTACCCATGGTATGGATGAATTGTACAAATAA.

2.5.2 Cell culture

Cell culture experiments were designed to gather gene expression and growth rate measurements in parallel. From a solid agar plate with appropriate selection, single colonies of each strain in the library were used to inoculate 400 ul yeast peptone

media with raffinose as the carbon source (YPR) media in a 96 deep-well plate. This plate of stock cultures contained two replicates of wildtype, autofluorescence and each single gene deletion strain, and one replicate of each combinatorial gene deletion strain. Following 48 h growth (30°C, 250 rpm shaking), stock culture plates were stored at 4°C for four days, to be used for four replicate experiments. For each experiment, 60 ul of each stock culture was used to inoculate 400 ul YPR in a deep-well plate. After 16 h growth (30°C, 250 rpm shaking), cell density of cultures was estimated by absorbance (596-604 nm light, Victor 3V Plate Reader, Perkin Elmer). Following absorbance measurement, each culture was divided into two, corresponding to conditions without (YPR) or with the galactose stimulus (YPR with 2% w/v galactose, YPRG). The use of raffinose as a neutral carbon source in galactose-induced transcription experiments is reported previously in [67, 73]. For each condition, a given culture in a well had a total volume of 300 ul and absorbance adjusted to approximately 0.15. After 3 h growth, absorbance of each culture was re-measured and re-adjusted to an absorbance reading of 0.02 in a volume of 750 ul. To obtain growth rate measurements, a 300 ul aliquot of each adjusted culture was transferred to a 100-well plate. The remaining cells were grown for 3 h (30, 250 rpm shaking) and were then used for gene expression measurements. Four such experiments resulted in eight measurements of each autofluorescence, wildtype and single gene deletion strain, and four measurements of each double gene deletion strain, for each condition.

Growth media

Yeast peptone medium with raffinose as the carbon source (YPR) contains the following in H_2O: 10 g/l yeast extract (Wisent), 20 g/l bacteriological peptone (Wisent, Cat. no. 800-157-LG), 20 g/l raffinose, and 0.084 g/l adenine hemisulfate (Sigma, Cat. no. A9126). Adenine is added because strains have the *ADE2* gene

deleted. Yeast peptone medium with raffinose and galactose as the carbon source (YPRG) contains the ingredients of YPR and 20 g/l galactose.

2.5.3 Population growth rate measurements

Growth rate was estimated from time-lapse measurements of cell culture absorbance of light between 450 nm and 580 nm, acquired in a specialized 100 well plate (Growth Curves USA) maintained at 30°C without shaking, in a spectrophotometer equipped with an incubator (Bioscreen C Analyzer, Growth Curves USA). Measurements were acquired every 15 minutes for 22 h. Absorbance measurements over time were fit to an exponential growth model $Abs(t) = Abs_0 \cdot e^{k \cdot t}$ with parameters estimated by minimizing the sum of squared residuals, where $Abs(t)$ is the absorbance measurement at time t (min), Abs_0 is absorbance of the first time point considered in a given fit, e is Euler's number, and k is the rate (min^{-1}) by which Abs increases by a factor e. Fitting was performed using a custom script in MATLAB (The MathWorks, Inc., Natick, Massachusetts, United States) by Mads Kærn and Cory Batenchuk. The phenotypic growth rate measurement used for inference corresponds to the estimate of k.

2.5.4 Reporter gene expression measurements

Single cell fluorescence corresponding to GFP emission was acquired by a flow cytometer (Beckman Coulter FC500) with a 488nm excitation laser. Cell green fluorescence is estimated as the Log Pulse Height of a signal corresponding to 510-550 nm filtered emission. I obtained a minimum of 60,000 events per sample. Of these events, some were filtered to limit analysis to cells of similar size and morphology. Events were filtered based on Log Height Pulse of signals corresponding to forward

(FS) and side-scattered (SS) 488nm light. For each sample, analysis was limited to 50% of acquired events per sample belonging to 2D FS-SS area that maximized event density per area. These filtered events were linearized and normalized. I consider the mean of these filtered events to be the quantitative gene expression trait (arbitrary units) for a given sample. Flow cytometry data analysis was conducted with custom (Mads Kærn) and open source scripts for reading FCS files (*fca_readfcs.m*, Ver 1.5, 2006-2009, University of Debrecen, Institute of Nuclear Medicine, Laszlo Balkay). All scripts are written in MATLAB (The MathWorks, Inc., Natick, Massachusetts, United States).

2.5.5 Theoretical trait derivation

Symbolic computation (MATLAB 2017a, The MathWorks, Inc., Natick, Massachusetts, United States) was used to derive theoretical traits as well as expectation values for parameters of linear regression equations (Tables 2.4 to 2.7).

2.5.6 Linear regression of trait measurements

I obtained β parameters and corresponding standard errors by linear least squares regression of T according to Eq 2.4 with 48 degrees of freedom where $T = log_2(T_{exp})$ and T_{exp} is a gene expression or growth rate measurement, as described above). I obtained δ parameters and corresponding standard errors by regression of D according to Eq 2.5, having 24 degrees of freedom. Linear regression statistics were obtained using the MATLAB function *regstats.m* (The MathWorks, Inc., Natick, Massachusetts, United States).

2.5.7 Transitive reduction of gene networks

Gene networks in Figure 2.8A-C were obtained by forming a graph N representing the network of all inferred directed edges for a given phenotype (Figure 2.8A-B) or both phenotypes (Figure 2.8C). Since inferences may differ between phenotypes, I applied the following rules: (1) if an edge or gene order cannot be inferred for one of two phenotypes, incorporate the edge that can be inferred in N; (2) if the inferences obtained for the two phenotypes differ in gene order or sign of edge, exclude this edge in N. I performed transitive reduction of N with a MATLAB function (*transreduction.m*, MATLAB 2017a, The MathWorks, Inc., Natick, Massachusetts, United States).

3

Generalized epistasis analysis

Summary

I describe how the the rules of epistasis analysis in Chapter 2 can be generalized to allow inference of any acyclic topology from quantitative genetic interaction experiments. This analysis increases knowledge about how discernible alternative topology models are for such experiments, both theoretically and in practice. This advance is important because these data are extensively gathered and interpreted without thoroughly tested methodologies.

Contributions

The founding analyses presented in this chapter are published [74]. I performed initial testing and analysis of Battle et al's method [60]. Dr. Theodore Perkins developed the model and derived equivalence classes. Dr. Mads Kærn developed the model and generated simulated data. Dr. Mads Kærn and I tested the model predictions on the simulated data. For the experimental dataset, Dr. Mads Kærn, Mila Tepliakova and I designed the experiments. Mila Tepliakova, Daniel Jedrysiak and I contributed to strain generation, and I conducted the experiments. Dr. Mads Kærn developed algorithms for flow cytometry analysis. I extended [74] by performing the analyses described in this chapter, with the advice of Dr. Mads Kærn.

Contents

3.1 Background

Genetic interaction data can be used to propose hypotheses about how gene product activities regulate each other and the phenotype for which the data is obtained. Such hypotheses can take the form of topology models with directed edges between gene nodes. It remains an open question whether a topology inferred from genetic interaction data represents the most accurate hypothesis, or if there are alternative topologies that explain the data as well or better. Encompassed in this question is the problem of how certain assumptions influence the set of topologies considered as plausible hypotheses in various studies.

To illustrate this problem, Azpeitia et al. [75] simulated genetic interaction data with the assumptions that gene activities regulate one another through cyclic Boolean networks. When one is blinded to original assumptions of the simulation and instead considers the assumptions of classical epistasis analysis, one wrongly infers the cyclic network to be acyclic and hierarchical. This illuminates a fundamental limitation to epistasis analysis for the inference of feedback loops.

Assumptions of classical epistasis analysis, as described by Avery and Wasserman [55], allow one to consider the set of hierarchical topologies between a signal and two genes (Figure 3.1A) as plausible hypotheses. This set includes sixteen topologies, differing in which of two genes is upstream of the other, and differing in whether each of the edges between the three nodes is a positive or negative regulatory edge. Notably, the remaining edges illustrated in Figure 3.1A are allowed by classical assumptions but not identifiable by classical rules. This is a major problem that can only be addressed by creating new methods that generalize epistasis analysis.

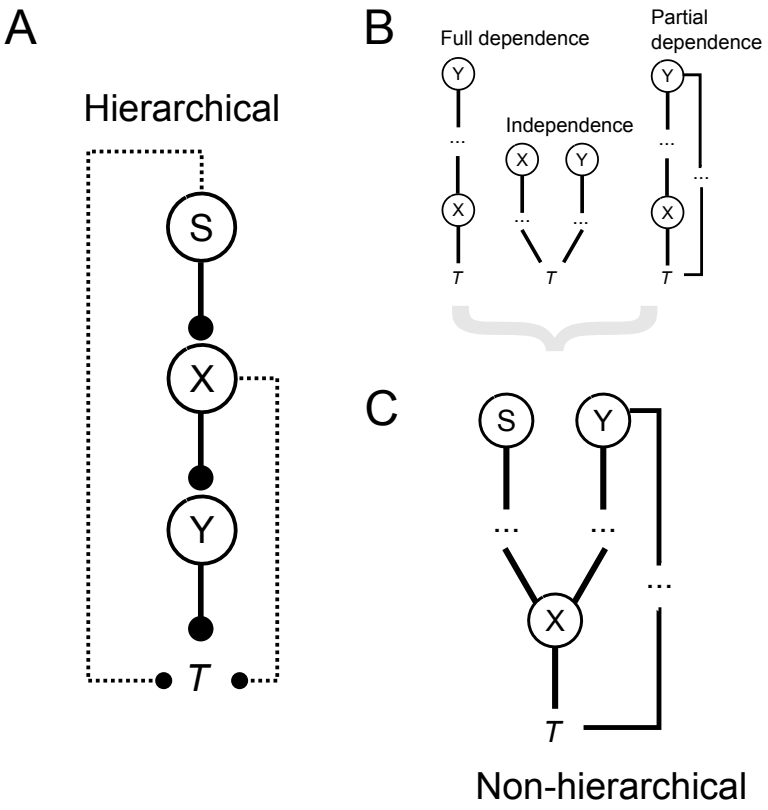

Figure 3.1: Signal and gene pathway hierarchies are among the possible networks inferred by [60]. A regulatory hierarchical pathway (**A**), as defined by Avery and Wasserman [55], requires a set of three nodes where the stimulus S has no input, genes X and Y have exactly one input, and where a minimum of one gene influences the observed phenotype ρ. In contrast, Battle et al [60] considered three possible dependence types between two genes X and Y (**B**). These structures can combine as any size acyclic topology (**C**), which may or may not contain 'hierarchical' node groups. Symbols: ..., any number of nodes.

Classical assumptions include gene and signal activities to be Boolean, i.e. "ON" or "OFF". This assumption was also present in the simulation of Azpeitia et al [75]. An important difference between [75] and [55] is the consideration of feedback. It can be argued that genetic interaction experiments are inherently limited to acyclic topology hypotheses since the effects of loss of feedback versus non-feedback influences are not separable when genetic mutations cause full loss of gene function. Azpeitia et al. [75] therefore outlined how cyclic Boolean topology simulations could be conducted post-hoc, that is following classical epistasis analysis, to identify alternative plausible models testable by secondary experimental designs.

In the present chapter, I address the question of discernible hypotheses that are encompassed within the inherent limitations of genetic interaction data when full loss of function mutations are employed. I begin with the research of Battle et al. [60] who described a method for inferring both hierarchical and non-hierarchical acyclic topologies from genetic interaction data and made several advancements.

I emphasize the following three advancements. First, Battle et al. [60] consider that a pair of genes can regulate one another in three possible ways. These distinct regulatory types are in the form of node dependencies, wherein the activity of one gene node can be independent, fully or partially dependent on the other gene node (Figure 3.1B). Second, they derived expectation values for each of three dependencies and assumed that one could identify the best-fit dependency by minimizing the difference between observed data and the possible expectation values. Third, they considered that such dependencies among gene pairs could combine in the form of large networks of greater than two genes, and described a method for searching the space of possible networks, guided by minimizing the sum of differences between observed data and the expectation values encompassed in a hypothetical network. I contrast these approaches to preceding research limited to, for example, consideration of hierarchical topologies [39, 43, 55, 57, 58], the assumption that genes have Boolean

activities [55, 58], consideration of topologies no larger than two nodes [43, 55, 57, 58], or lack a quantitative measure of agreement between alternative hypotheses and the observed data [43, 55]. In summary, the approach of [60] allowed more topologies to be quantitatively weighed against one another for a particular dataset than ever before.

In the present chapter, I first aim to systematically test how well the approach of Battle et al. [60] allows one to discern alternative hypotheses about how genes function in hierarchical and non-hierarchical topologies. To achieve this I analyze simulated combinatorial stimulus and gene perturbation data. Battle et al. originally tested the method on a large experimental dataset [36].

Simulated data has many advantages over experimental data for benchmarking topology inference methods. Experimental data can compound the question of whether alternative topologies can be discerned as a superior hypothesis or not. For example, experimental error may be higher than the differences in expectation values of multiple alternative topologies, or the genes considered in an experimental dataset may overrepresent certain topologies and therefore limit the knowledge of whether underrepresented topologies are identifiable [76]. Moreover, using experimental data for benchmarking is difficult because the *true* topology for a native biological network is often subject to change with further research. Such uncertainties are well-documented and common to any biological data-driven topology inference problem. A simulated genetic interaction dataset allows one to circumvent these complications and also allows one to systematically test causes of method failure through modification to the simulation, such as measurement noise, for example, or assumptions underlying the method.

By testing a scoring method based on Battle et al. on simulated data, I reveal limits in identifiability among all possible acyclic topologies from genetic genetic interaction

data. To increase knowledge about the origin of these limits, I describe an approach to derive theoretical identifiability of the same topologies. This approach generalizes classical epistasis analysis by considering any possible patterns of phenotype equalities as a means to identify topologies from their theoretical phenotypes. I then apply knowledge of theoretical trait patterns to derive a new method for inference.

I compare both the scoring method and this new method based on phenotype equivalence identification against increasingly noisy data. First, I analyze the effect of measurement noise in simulated data values. These data allow me to further develop the new method. Lastly, I apply the new method and the scoring method towards an experimental test case. I find that the stimulus applied in the experimental test case has unexpected effects, and therefore a subset of these analyses yield new hypotheses about the network, to be tested in the next chapter of this thesis.

3.2 Objectives

- The first objective is to benchmark a topology scoring method based on Battle et al., using simulated data.
- The second objective is to develop a model to derive theoretical stimulus and gene perturbation phenotypes.
- The third objective is to derive a method for topology inference method based on theoretical phenotype equivalence patterns.
- The fourth objective is to optimize this method by testing it against simulated data with added noise.
- The fifth objective is to test this optimized method on an experimental combinatorial gene and stimulus perturbation dataset.

3.3 Results

3.3.1 Simulating genetic interaction data

To systematically analyze acyclic topology inference allowed by the method of Battle et al. [60], I used simulated quantitative genetic interaction data described in [74]. These data correspond to quantitative phenotype values for eight hypothetical perturbation experiments, wherein a stimulus S is present or not, and each of two genes X or Y is deleted or not. These hypothetical experiments correspond to the experiments analyzable by classical epistasis analysis, as discussed in the previous chapter. In contrast to classical analysis, which only considers fully hierarchical pathways, the simulation considers that nodes S, X and Y may regulate a theoretical trait ρ in any acyclic topology that may contain both hierarchical and non-hierarchical edges.

To accommodate assumptions about the nature of full loss of function perturbation experiments there are some restrictions to the acyclic topologies considered in the simulated dataset. Specifically, in all acyclic topologies considered, the following restrictions are met: (1) the trait is be reachable from S, X and Y; (2) S is never downstream of other nodes; and (3) ρ is never upstream of other nodes. Notably, these restrictions in part correspond to analyzing perturbation data wherein perturbation of S, X or Y, is expected to have a significant effect on a hypothetical phenotype measurement in at least one condition. Within these restrictions, all permutations of edges and gene order (Figure 3.2) result in having simulated data for 35 possible topologies, as shown in Figure 3.3.

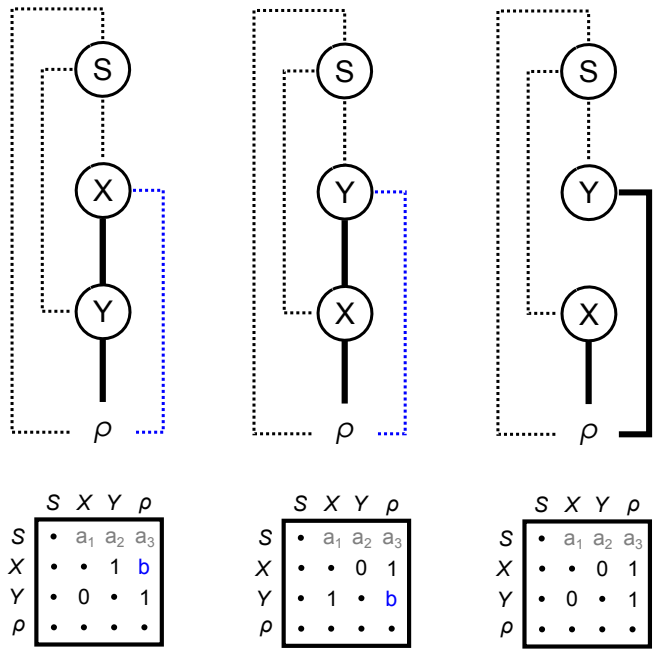

Figure 3.2: Derivation of thirty-five acyclic topologies. In the topologies I consider, node X and Y may not regulate one another (right), X may regulate Y (left), or Y may regulate X (middle). In the latter two scenarios, the upstream X or Y node may directly influence the theoretical trait ρ ($b \in 0, 1$, blue lines). For these five scenarios, S may regulate one, two or all of X, Y, ρ (dotted lines). In corresponding binary edge matrices, these regulatory influences correspond a the sum of a_1, a_2 and a_3 equal one or greater. Symbols: S, signal node; X, gene node; Y, second gene node; ρ, phenotype; one in matrix, edge is present; zero in matrix, edge is absent; dot in edge matrix, edges absent in all topologies.

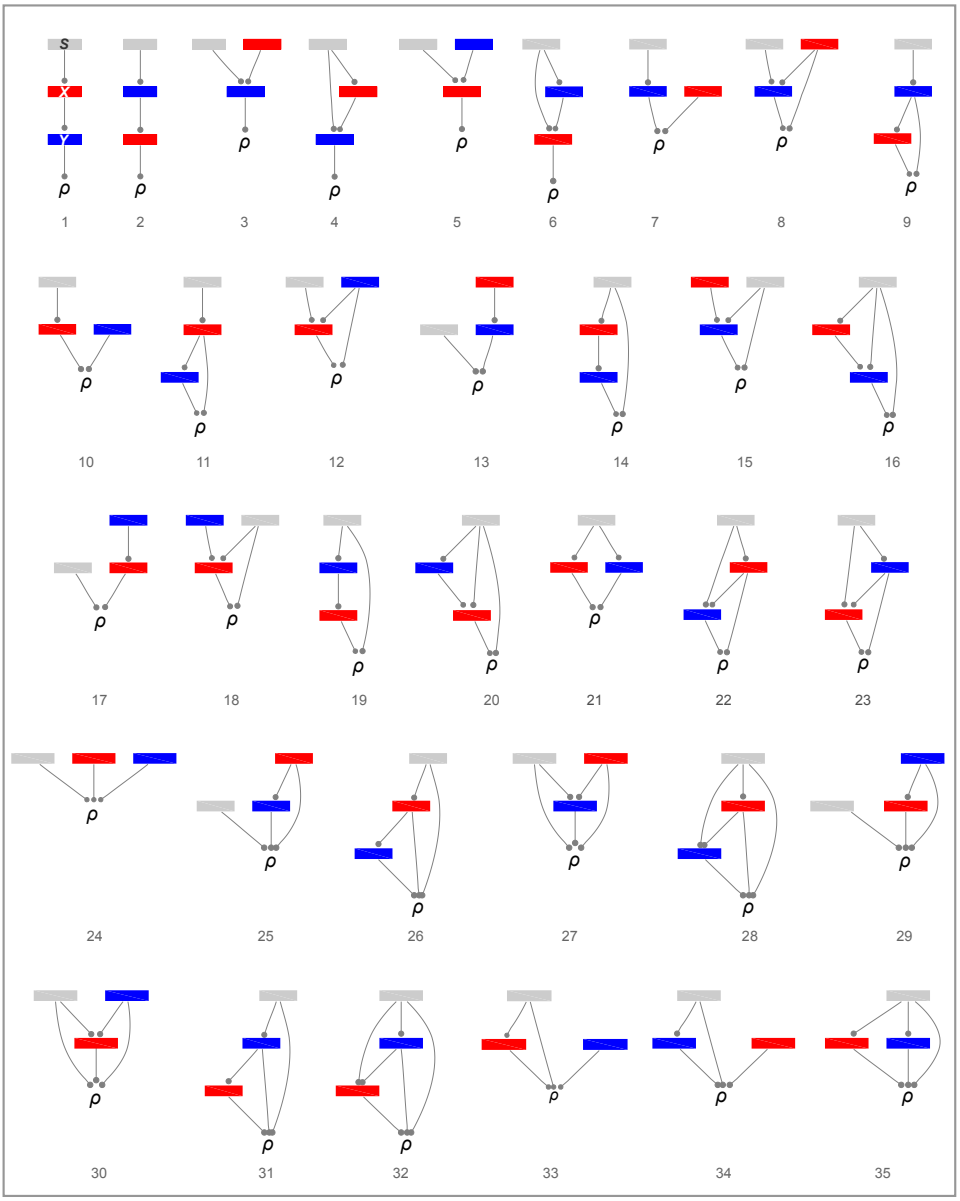

Figure 3.3: Thirty-five acyclic topologies considered in the present study. In the topologies I consider, a signal S (grey), gene X (red) and gene Y (blue) influence a theoretical phenotypic trait ρ in one of thirty-five ways according to the rules in Figure 3.2. Note: topology enumeration is consistent in all figures.

For each topology, I consider 100 simulations of genetic interaction data as described in [74]. These simulations were obtained by modifying the method of [76, 77], which has been applied to benchmark many network inference methods [76]. In this method, a simulated dataset is obtained by assigning to each gene node ordinary differential equations (ODE) to calculate the rate of change in mRNA and protein abundance over time. Node activity is modelled by the time derivative of protein abundance, and topology edges are modelled as the dependence of the time derivative of mRNA of a downstream node on the activity of its upstream node(s).

Full details of the simulation method are provided in §3.5.1. Briefly, the method assumes that nodes correspond to genes in a transcriptional regulatory network wherein the protein of an upstream gene is a regulator of transcription of each of its downstream nodes connected by an edge. When a downstream node has multiple upstream regulators, we assume that each upstream regulator functions as transcription factor which can independently regulate the downstream gene j by binding to an independent cis-element on a hypothetical transcriptional promoter of j.

For each topology, the 100 simulated datasets differ in parameter magnitudes influencing mRNA or protein production and decay, as well as the Hill coefficients and constants that modify the functions representing topology edges. By considering multiple simulations of the same topology, one can analyze whether topology inference by the method of Battle et al. [60] is influenced by data or parameter values which are topology-independent.

3.3.2 Isolating a Topology Score

The method of Battle et al. [60] allows topology inference by scoring how well each hypothetical topology explains a genetic interaction dataset, and then inferring the best-scoring hypothetical topology. More specifically, the method of Battle et al. was designed for large networks wherein there any many hypothetical topologies and therefore the score is also the basis for an optimization process to find the best-scoring topology in a search space of all possible hypothetical topologies based on the number of genes. The original score incorporates many terms, some of which are required to estimate scores when there are missing experimental data, or to incorporate uncertainty in experimental measurements. Of particular interest in the present study are the terms that specifically correspond to the assumptions underlying the discernment of hierarchical and non-hierarchical topologies, to analyze if we can discern the 35 acyclic topologies based on these assumptions.

I briefly summarize these assumptions. First, there are three mutually exclusive possibilities for how two gene nodes X and Y depend on one another to influence a phenotype ρ. These possibilities are called dependencies. The first dependency type is full dependency, wherein X is fully dependent on Y to influence ρ. The second dependency type is independence, wherein X and Y independently influence ρ. The third dependency type is partial dependence, wherein X is partially dependent on Y to influence ρ. Each dependency type corresponds to a specific node relationship in a topology, identifiable based on topology edges and reachability (described in Methods §3.5.2). The node relationship corresponding to each dependency type is depicted in Figure 3.1.

Each dependency type corresponds to a unique expectation value ($\epsilon_{\Delta x \Delta y}$) for a hypothetical measured experimental phenotype T when both nodes X and Y are

deleted ($T_{\Delta x \Delta y}$). The expectation value $\epsilon_{\Delta x \Delta y}$ is calculated from T when neither X nor Y is deleted (T_{xy}) or one node is deleted ($T_{\Delta X}, T_{\Delta Y}$), as follows

- Full dependency of X on Y: $\epsilon^{f}_{\Delta x \Delta y} = T_{\Delta y}$,
- Independence of X and Y: $\epsilon^{i}_{\Delta x \Delta y} = T_{\Delta x} * T_{\Delta y} / T_{xy}$,
- Partial dependence of X on Y: $\epsilon^{p}_{\Delta x \Delta y} = [\epsilon^{i}_{\Delta x \Delta y} + \max(T_{\Delta x}, T_{\Delta y})]/2$,

where nodes $X \neq Y$, and $X, Y \neq \rho$.

Each of these dependencies identified in a topology contributes to the overall score of the topology. For each dependency identified of type m for two hypothetical gene nodes X and Y, the dependency is scored as $\delta^{m}_{x,y} = |\epsilon^{m}_{\Delta x \Delta y} - T_{\Delta x \Delta y}|$. For a hypothetical topology γ, the topology is then scored as the sum of all dependency scores, $TS_{\gamma} = \sum_{1}^{n} \delta_{n}$, where n is the number of unique node pairs. The best scoring topology is assumed to be one which encompasses a combination of dependency types that allow minimization of the sum of differences between expected phenotypes and measured phenotypes.

I specifically analyzed if the above-described assumptions, isolated from the original scoring method of Battle et al., would allow one to correctly infer one of 35 topologies from the above-described simulated combinatorial stimulus and gene perturbation data. To conduct this analysis, I computed TS_{γ} for each of 35 hypothetical topologies ($\gamma = 1$ to 35), and determined how many times the topology with the lowest TS_{γ} score corresponded to the actual topology for which the simulated data was obtained. I refer to TS_{γ} as a Topology Score, and application of TS_{γ} for inference as the topology score method.

In this first analysis, I additionally tested alternative forms of the above-described dependency-specific expectation calculations ($\epsilon_{\Delta x \Delta y}$). I found that normalizing $\epsilon_{\Delta x \Delta y}$

based on magnitudes of T considered in the calculation reduced success of topology inference. In addition, I found a slight improvement in inference when I altered the expectation calculation for partial dependency to the following

- Partial dependence of X on Y: $\epsilon^{p}_{\Delta x \Delta y} = [\epsilon^{i}_{\Delta x \Delta y} + \epsilon^{f}_{\Delta x \Delta y}]/2$,

An important difference between this calculation and the alternative described by Battle et al. is that this calculation is different when X is partially dependent on Y, versus when Y is partially dependent on X.

3.3.3 Inferences based on dependency types are often incorrect

I report the topology inference results, when I use this modified calculation of partial independence, as a heatmap in Figure 3.4. I find that certain topologies can be identified correctly for nearly all 100 simulations of the topology. These topologies correspond to Topologies 1,2,7,10,13, and 17 (Figure 3.3). In contrast, other topologies can never be identified correctly, corresponding to Topologies 16, 20, 26, 28, 31, 32 and 35 (Figure 3.3). The remaining topologies are inferred in a subset of simulations but not reliably identified. Often, wrong inferences are predictable. For example, Topologies 14, 15 and 16 are almost always wrongly inferred as Topology 13 (Figure 3.3). I refer to these patterns of errors as a bias in topology inference, since the inference of certain topologies appears to be favoured.

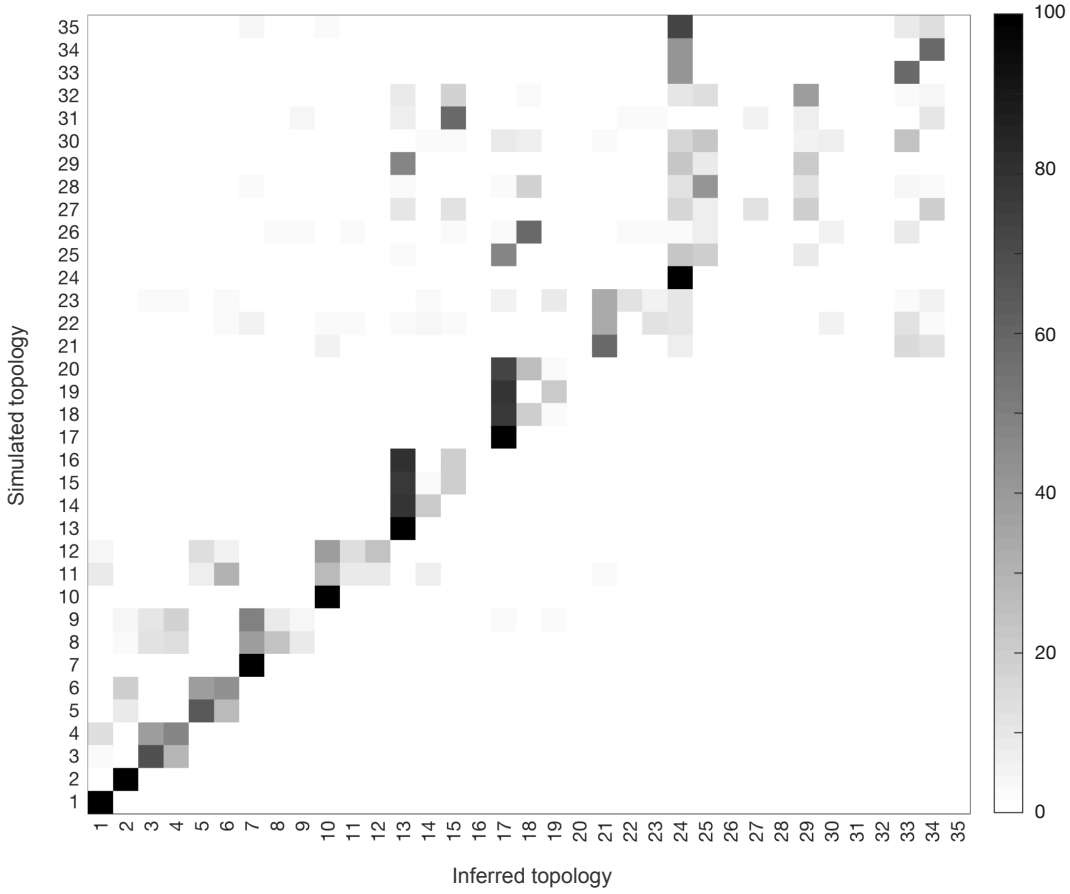

Figure 3.4: Topology scoring exhibits bias in topology inference from simulated data. The heatmap allows visual comparison of simulated and inferred topology when 100 simulations for each of topology are analyzed by topology score. Perfect inference for all simulations and topologies would correspond to a black diagonal line, whereas unbiased incorrect inferences for a given topology would correspond to a horizontal, uniformly grey line. See text and Methods.

3.3.4 Theoretical analysis reveals that certain topologies are indistinguishable

The bias seen for certain topologies could be imposed by simulation parameter values, assumptions of the simulation method, assumptions of the topology score, or a combination thereof. To test if exclusion of the first two possibilities eliminates bias, I consider the approach developed in [74] to obtain theoretical trait expressions for each topology. Specifically, I derive theoretical trait expressions by assuming that topology edges can be modelled by arbitrary functions, wherein a downstream node's activity is an arbitrary function of the activity of its upstream node or nodes. Theoretical trait derivation is described in Methods (§3.5.3), and is exemplified for Topology 6 (Figure 3.5) in Table 3.1.

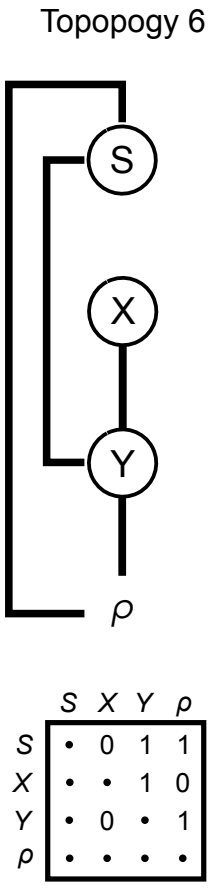

Figure 3.5: Theoretical trait expressions are derived from topology edges. Direct edges (lines) are defined in a binary matrix where unity at a given row and column represents presence of an edge from the upstream node (row) to the downstream node (column). The theoretical trait expressions for this Topology (6) are provided in Table 3.1. Symbols: as defined in Figure 3.2.

For each topology, I derive eight trait expressions reflecting the eight unique node perturbation conditions ($C1 - C8$) (Table 3.1), analogous to the conditions corresponding to simulated genetic interaction data. Notably, trait derivation imposes none of the assumptions of the simulated data.

Table 3.1: Example of topology-derived theoretical trait expressions. I provide as an example the theoretical trait expressions for Topology 6 (Figure 3.2). To derive the expression of the theoretical trait ρ I define each node i's activity ($i \in X, Y, \rho$) as an arbitrary function f_i of the activities of its upstream nodes connected by an edge. In absence of input a node activity may be equal to a basal term b_j ($j \in X, Y$) or a term representing the activity of the signal S. The eight node perturbation conditions C indicate which of S, X, and Y is deleted (0) or not (1), and may alter the expression. If a node is deleted in C, its activity is set to zero in the theoretical trait expression corresponding to C.

C	S	X	Y	ρ
1	0	1	1	$f_\rho(0, f_y(0, b_x))$
2	0	0	1	$f_\rho(0, f_y(0, 0))$
3	0	1	0	$f_\rho(0, 0)$
4	0	0	0	$f_\rho(0, 0)$
5	1	1	1	$f_\rho(s, f_y(s, b_x))$
6	1	0	1	$f_\rho(s, f_y(s, 0))$
7	1	1	0	$f_\rho(s, 0)$
8	1	0	0	$f_\rho(s, 0)$

By examining the trait expressions of Topology 6, one can find a number of interesting patterns. In particular, certain combinations of perturbations are predicted to yield the same trait values. This is of course related to the phenomenon of epistasis wherein a perturbation has no effect when another is present. For Topology 6, I find the following combinations of perturbations yield equivalent trait expressions: conditions $C3$ and $C4$ are equivalent, and separately those for conditions $C7$ and $C8$ are equivalent. These equivalences are readily explained by examination of the topology edges. For example, the upstream gene X has no edge input from the signal, and can only influence the theoretical trait ρ through its influence on the downstream gene Y. Thus, the effect on ρ of deleting X and Y ($C4$,$C8$), is always the same as when only Y is deleted ($C3$,$C7$). Because ρ also has an additional input from the signal S independent of Y, the dependency of X on Y corresponds to two separate sets of equivalences, i.e. $\rho_{C3} = \rho_{C4} \neq \rho_{C7} = \rho_{C8}$, where each set corresponds to the different conditions of S.

Because the equivalence pattern among theoretical trait expressions for Topology 6 clearly yields information that relates to the edges and nodes of the Topology, I analyzed to what extent this pattern of equivalence may be unique to Topology 6. Such a finding would indicate that the topology may identifiable from its phenotype patterns alone, obtained with very few assumptions.

To conduct this analysis, one derives eight theoretical trait expressions for each of 35 topologies and records the pattern of trait equivalences. If an equivalence pattern is unique to a topology, this topology is considered identifiable. Otherwise, all topologies having the same equivalence pattern belong to an equivalence class (EQC) of topologies that cannot be distinguished from one another.

Only two of 35 possible topologies are identifiable. The remaining 33 topologies are partitioned into eight EQCs. There are therefore ten EQCs which have identifiable

equivalence patterns, indicating that in general individual topologies may not be identifiable, but classes (EQCs) of topologies are.

By visualizing the topologies in each EQC (Figure 3.6) and the corresponding equivalence patterns in Table 3.2, one can make several conclusions. First, topologies in the same class have many conserved edges and conserved order among the gene nodes X and Y. For example, in *EQC8*, all topologies in the class have gene Y upstream of X, and have three out of four edges conserved. Second, I find that equivalences correspond to node dependencies in the corresponding topology. This finding has implications for identifiability since many pathways between classes can have a subset of the same dependencies (Table 3.2). For example, topologies of *EQC3* and *EQC1* all have a dependency relationship wherein X is fully dependent on Y, and the trait ρ is fully dependent Y, corresponding to the equivalence pattern $\rho_{C7} = \rho_{C8} = \rho_{C3} = \rho_{C4}$. *EQC1* has an additional equivalence that corresponds to the dependency of X on S, which allows *EQC1* to be distinguished from *EQC3*. Third, the number of topologies in an EQC is correlated with the number of equivalences. *EQC10* has no node dependencies, and no equivalences, and contains the largest number of topologies (12 out of 35).

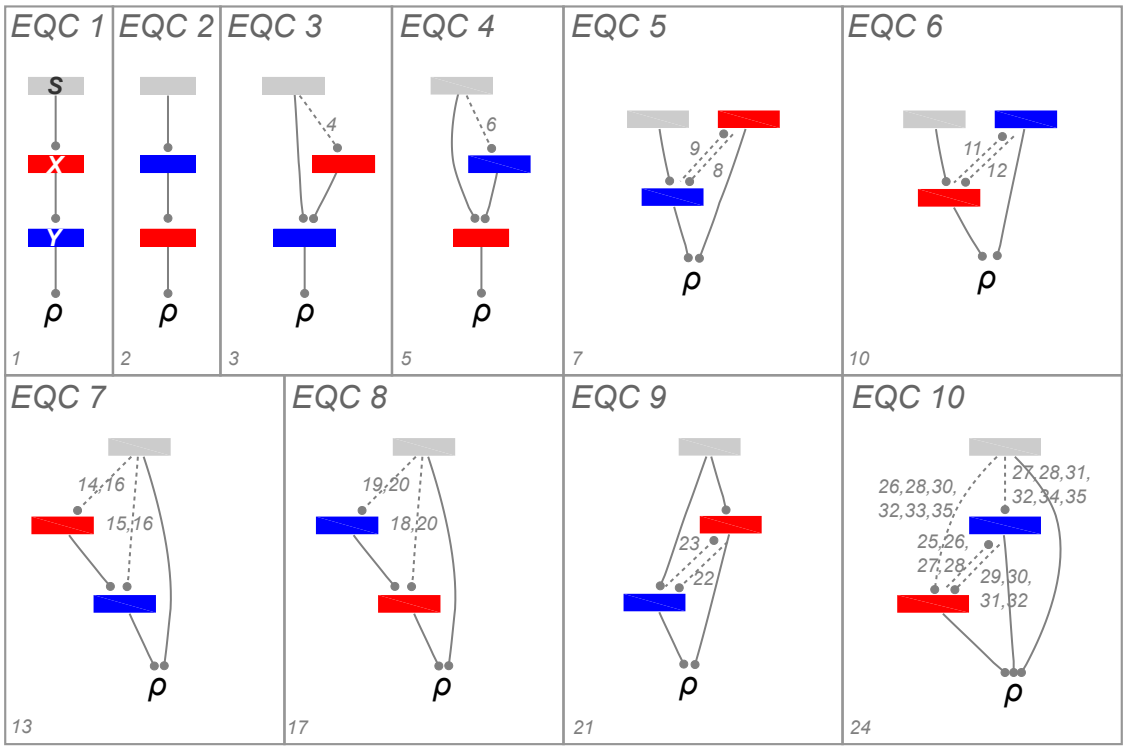

Figure 3.6: Topology edges conserved within equivalence classes. An equivalence class (EQC) contains topologies wherein a signal (grey) and two genes, X (red) and Y (blue), influence a trait (ρ). An edge (line) indicates the node nearest the end circle is influenced by the node where the edge begins. Edges conserved in all topologies of a class are solid, variable edges are dotted. To indicate topology numbers defined by variable edges, we enumerate the topology with conserved edges only at bottom left, and enumerate the topology with one or more variable edges beside each of its variable edges.

Table 3.2: Unique patterns of theoretical trait expression equivalences among conditions C define an EQC. Thirty-five topologies are partitioned into ten EQC, with equivalence patterns illustrated here. Equivalent trait expressions within a set are all indicated by the same ρ_i symbol ($i \in A, B$, $A \neq B$). Symbols are not comparable between classes (i.e. ρ_i of class j is not equivalent to ρ_i of class k ($k \neq j$). EQC order is non-numerical to show similarities between EQCs. Symbols: $-$, trait not belonging to an equivalent set.

C	$\{S, X, Y\}$	EQC 1	3	6	9	2	4	5	7	8	10
1	$\{0, 1, 1\}$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
2	$\{0, 0, 1\}$	ρ_A	$-$	ρ_A	$-$	ρ_A	ρ_A	$-$	$-$	ρ_A	$-$
3	$\{0, 1, 0\}$	ρ_B	ρ_B	$-$	$-$	ρ_B	$-$	ρ_B	ρ_A	$-$	$-$
4	$\{0, 0, 0\}$	ρ_B	ρ_B	ρ_B	ρ_A	ρ_A	ρ_A	ρ_A	ρ_A	ρ_A	$-$
5	$\{1, 1, 1\}$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
6	$\{1, 0, 1\}$	ρ_A	$-$	ρ_A	$-$	ρ_A	ρ_A	$-$	$-$	ρ_B	$-$
7	$\{1, 1, 0\}$	ρ_B	ρ_B	$-$	$-$	ρ_B	$-$	ρ_B	ρ_B	$-$	$-$
8	$\{1, 0, 0\}$	ρ_B	ρ_B	ρ_B	ρ_A	ρ_A	ρ_A	ρ_A	ρ_B	ρ_B	$-$

3.3.5 Trait equivalences limit epistasis analysis

The low number of identifiable topologies suggests that inference bias of the topology score may be imposed by theoretical limits of identifiability, and none of the three possibilities I initially proposed. To test this hypothesis I re-analyzed topologies inferred by topology scores. I categorized the topology of each simulated dataset according to its theoretical *EQC*, as well as the topology inferred. Correct EQC inference is identified when inferred and simulated topologies are in the same class.

If topology inference bias is solely imposed by theoretical limits, I expect topology scores to allow perfect inference of EQC despite a low rate of correct topology inference. Although the inference of EQC is not perfect (Figure 3.7A), I find EQCs are correctly identified with a substantially higher probability (80% of simulations) than individual topologies (15% of simulations).

In other words, the data suggests that most of the errors arise within a given equivalence class, and aiming to identify which *EQC* is consistent with a given dataset will in most cases be more accurate. Additionally, I find that within an *EQC* the topology score is biased to identify the topology of the class with the fewest edges, corresponding to a low false discovery rate of edge detection (Figure 3.6, Table 3.3).

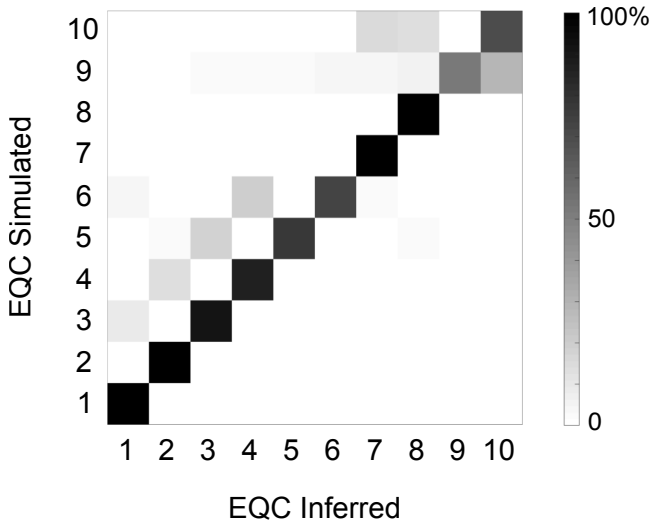

Figure 3.7: Topology score inference errors reflect limits of theoretical identifiability. For all topologies and all simulations, I quantify the number of times the topology score method allowed inference of the correct theoretical EQC (see Text).

The results of Figure 3.7 illustrate that limits of theoretical identifiability largely explain topology inference errors by a topology score, with some remaining influence on errors unaccounted for. While certain classes are more likely to be inferred correctly (*EQC1*, 100% of simulations) than others (*EQC9*, 52% of simulations) (Table 3.3), this remaining bias may be imposed topology-dependent assumptions of the topology score or the simulations. Topology-dependent assumptions cannot be ruled out in this case since topologies of *EQC*s have many conserved edges and node dependencies (Figure 3.6). For example, the bias to infer topologies of *EQC9* incorrectly may be explained by these topologies having a node dependency relationship not detectable by a topology score. This dependence, of the signal on both genes (Figure 3.6), is observable by comparing triple node perturbation to dual node perturbation (i.e. $C4$ and $C8$ in Table 3.2). In contrast, in a topology score the expectation calculation for full dependency considers only two nodes at a time.

Table 3.3: Statistical analysis of Topology score inference errors. For all simulations and topologies of each equivalence class I calculated the percentage of simulations where the *EQC* (A) or topology (B) was inferred correctly. In cases of wrong inferences I calculated (C) the average number of wrong edges (± standard deviation), (D) the false negative rate of edge detection (% edges) and (E) the false discovery rate of edge detection (% edges).

EQC	A	B	C	D	E
1	100	100	–	–	–
2	100	100	–	–	–
3	92	58	1 ± 0.2	8	5
4	87	54	1.1 ± 0.3	9	5
5	78	43	1.7 ± 1	19	9
6	73	44	1.9 ± 1.2	19	11
7	100	35	1.3 ± 0.5	21	0
8	100	35	1.3 ± 0.5	21	0
9	52	22	2.3 ± 1.1	27	13
10	70	23	2.2 ± 1	30	9

3.3.6 Identification of numerical equivalences

To rule out the influence of the assumptions of topology scoring on *EQC* inference errors I examine the extent to which theoretical trait equivalences for each *EQC* can be identified in simulated traits. This would allow *EQC* inference in absence of a topology score. I refer to this alternative inference approach as the identification of numerical equivalences.

Numerical equivalence identification requires a measure to assess if trait values deviate by a small enough value to be considered equivalent, and a method to correct for transitivity between numerical equivalence groups. This correction ensures equivalent sets are transitive, for example, if I identify two numerical equivalence groups $\{a, b\}$ and $\{a, c\}$, the correction assumes that $\{a, b, c\}$ is the true equivalence group.

I first consider the simplest measure of trait deviation, the absolute difference between traits. Specifically, this corresponds to trait values t_i and t_j being considered equivalent when $|t_i - t_j| \leq \theta$, where θ is the threshold value for equivalence. I examine the effect of varying the threshold value θ on *EQC* inference errors which determines whether the deviation is sufficiently small.

This analysis allows one to identify that with an optimal threshold, over 99% of simulations can be identified in the correct *EQC* (Figure 3.8). This result suggests that the $\sim$ 20% of *EQC* inference errors obtained by topology score are imposed by the assumptions of the topology score. For example, the topology score assumes that the trait is defined by pairwise dependencies, while the equivalence testing method makes no such assumptions.

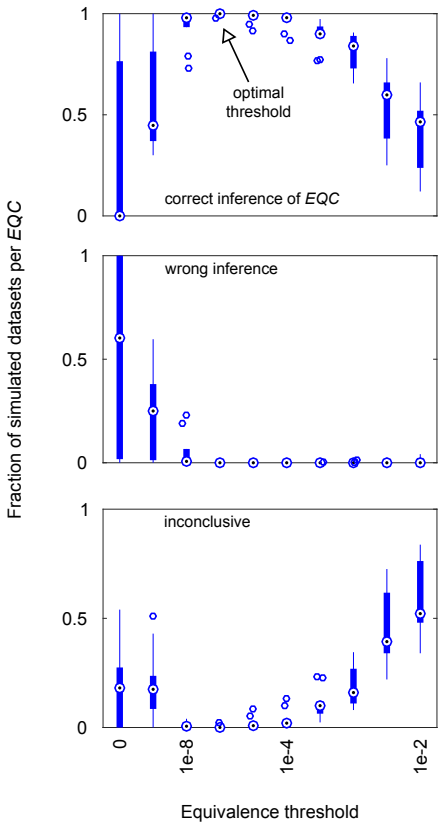

Figure 3.8: Success of numerical equivalence method of inference depends on threshold. Numerical equivalence occurs whenever the threshold $\theta \geq |t_i - t_j|$ where t is a simulated trait value and $i \neq j$. Numerical equivalence groups may match true theoretical equivalence groups (correct inference), match theoretical groups of the wrong EQC (wrong inference), or fail to match equivalences of any EQC (inconclusive). Symbols: black circle of boxplot, median; boxplot extremes, 25th and 75th percentiles; arrow, inference at optimal threshold of $8.5e-6$.

3.3.7 Evaluating the effect of measurement noise

The analysis in the previous section suggests that the equivalence method is more successful than the topology scoring method in identifying the correct *EQC* of a simulated topology. The success of the numerical equivalence method is, however, contextual. First, success requires identification of an optimal threshold. This is achieved by knowledge of the true topology for each simulation. Thresholds above or below optimal result in *EQC* inference errors (Figure 3.8). With lower thresholds, there is a failure to identify all numerical equivalences of an *EQC*, resulting in wrong inferences (Figure 3.9). With higher thresholds, spurious numerical equivalences result in inconclusive patterns not found in any *EQC* (Figure 3.9). Notably, inconclusive patterns have a subset of equivalences that do correspond an *EQC*, as well as additional equivalences that are not found in any *EQC* (Figure 3.9). For example, at a minimum every inconclusive pattern agrees with *EQC10*, as *EQC10* lacks any equivalence.

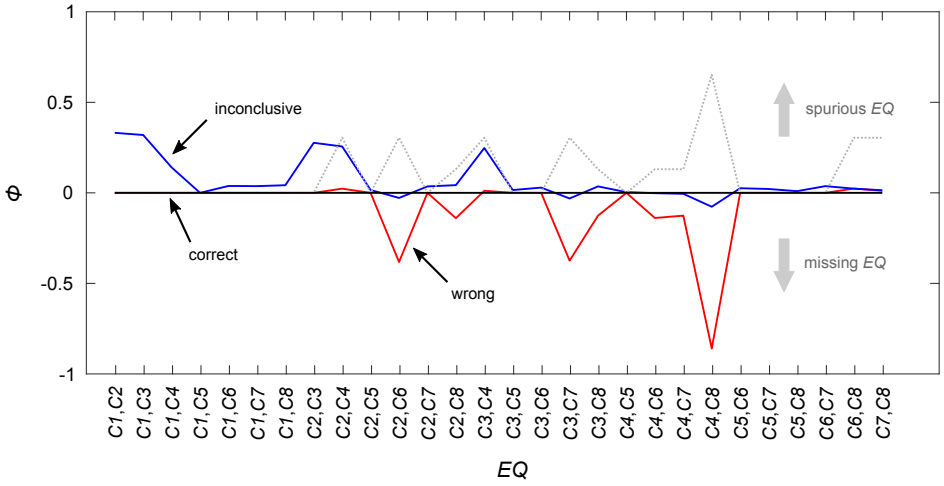

Figure 3.9: Differences between theoretical and numerical equivalences characteristic of inference error types. Inferences shown in Figure 3.8 are re-analyzed in terms of their numerical equivalences detected. Numerical equivalences are statistically compared to theoretical equivalences following categorization of inference type (correct, wrong, inconclusive). Measure $\phi^{e}_{EQ} = [\sum_n P - Q]/n$ is calculated for each equivalence test EQ across all simulations n resulting in inference type e. P is a logical indicating if the test was deemed numerically equivalent, whereas Q is a logical indicating if the corresponding theoretical equivalence is expected for the simulation. A ϕ^{wrong}_{EQ} value of -1, for example, indicates that all simulations resulting in wrong EQC inference had the EQ test fail numerical equivalence identification despite being theoretically equivalent. ϕ values can be compared to the dotted grey line, i.e. the theoretical probability of an EQ test being equivalent if one samples topologies from $EQC1$ to $EQC9$.

Second, the success achieved by an optimal threshold requires all pairs of theoretically equivalent traits, across all topologies and perturbation types, to differ by a value equal or less than the threshold. Differences between theoretically inequivalent traits, however, should exceed the threshold to limit false positive inference or inconclusive inference.

In the analysis of simulated traits, the optimal threshold ($\sim$ 8.5 x 10^{-6}) may correspond to a number of causes that limit the accuracy in simulated phenotypes. For example, theoretically equivalent phenotypes may differ in the time needed to reach steady-state in the simulation. In more realistic datasets, measurement noise is expected to complicate the identification of the appropriate or optimal threshold.

To analyze the effect of measurement noise on numerical equivalence identification, I impose uncertainty on simulated trait values. I test two scenarios: one wherein the same increasing level of noise is added to simulated traits (Figure 3.10A), and a second where different levels of noise are randomly assigned to simulated traits (Figure 3.10B). In both scenarios, I find the probability of a correct *EQC* inference to decline steeply when the uncertainty (or uncertainty maximum) exceeds the threshold. The simulation of traits with uncertainty is described in Methods (§3.5.5).

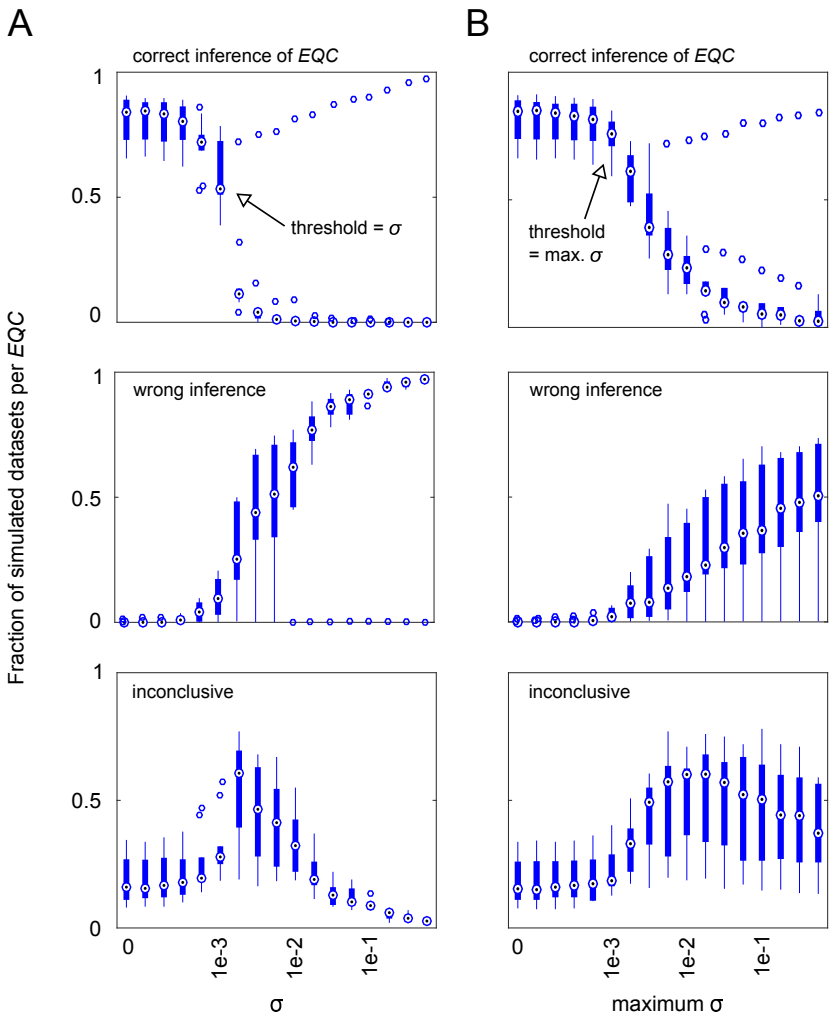

Figure 3.10: Noise added to trait values modifies inference errors. A. When increasing noise is added to simulated traits, correct inferences decline steeply when standard deviation of noise exceeds the threshold. **B.** When traits have different degrees of noise added, the steep decline of correct inferences occurs when the maximum standard deviation of randomly added noise exceeds the threshold. Numerical equivalence is identified as described in Figure 3.8. Symbols: as defined in Figure 3.8.

3.3.8 Adapting the method to allow inference with noisy simulated data

I hypothesized that numerical equivalence detection with measurement noise may be improved by incorporating the uncertainty of traits specific to each pairwise equivalence comparison. I refer to this analysis as noise-dependent threshold testing. Indeed, previous studies by Drees et al. [57] and St Onge et al. [43] have incorporated estimates of experimental error in the identification of epistasis from experimental genetic interaction data.

I test three alternative approaches to noise-dependent threshold testing. Each approach incorporates two estimated measures of uncertainty, one for each trait in a pairwise comparison. In my simulation of measurement noise, each trait is resampled with added noise (see §3.5.5). Estimates of uncertainty correspond to the calculated deviations from the means or medians of these resampled values. The three variants of noise-dependent threshold testing correspond to the following calculations for each pairwise trait comparison,

- **Variant 1:** traits t_i and t_j are equivalent when $t_i + z \cdot \bar{\sigma}_i \geq t_j - z \cdot \bar{\sigma}_j$, where $t_i \leq t_j$, t_i is the estimated mean and $\bar{\sigma}_i$ is the estimated standard deviation of values sampled to obtain the value of trait i with measurement noise, and z is a defined confidence threshold.

- **Variant 2:** traits t_i and t_j are equivalent when $t_i + \overline{MAD}_i \geq t_j - \overline{MAD}_i$, where $t_i \leq t_j$, t_i is the estimated median and $\overline{MAD}_i$ is the estimated median absolute deviation of values sampled to obtain the value of trait i with measurement noise.

- **Variant 3:** traits t_i and t_j are equivalent when $\frac{|t_i - t_j|}{\sqrt{\bar{\sigma}_i^2 + \bar{\sigma}_j^2}} \leq 1$, where t_i and $\bar{\sigma}_i$ are as defined for the first variant.

Notably, Variant 2 corresponds to the method by Drees et al. [57], whereas Variant 3 corresponds to the method by St Onge et al. [43], which is similar to a z-score.

To analyze if noise-dependent thresholds improve the inference of EQC from simulated data with added measurement noise, I apply each variant to make inferences from the same dataset analyzed by a single threshold in Figure 3.10B. I find Variant 1 allows the highest success in EQC inference (Figure 3.11A). This method identifies numerical equivalence when there is an overlap of confidence intervals based on estimated standard deviations. Nonetheless, all variants (Figure 3.11) allow substantially reduced numbers of wrong inferences compared to the inference from the same dataset using a single threshold.

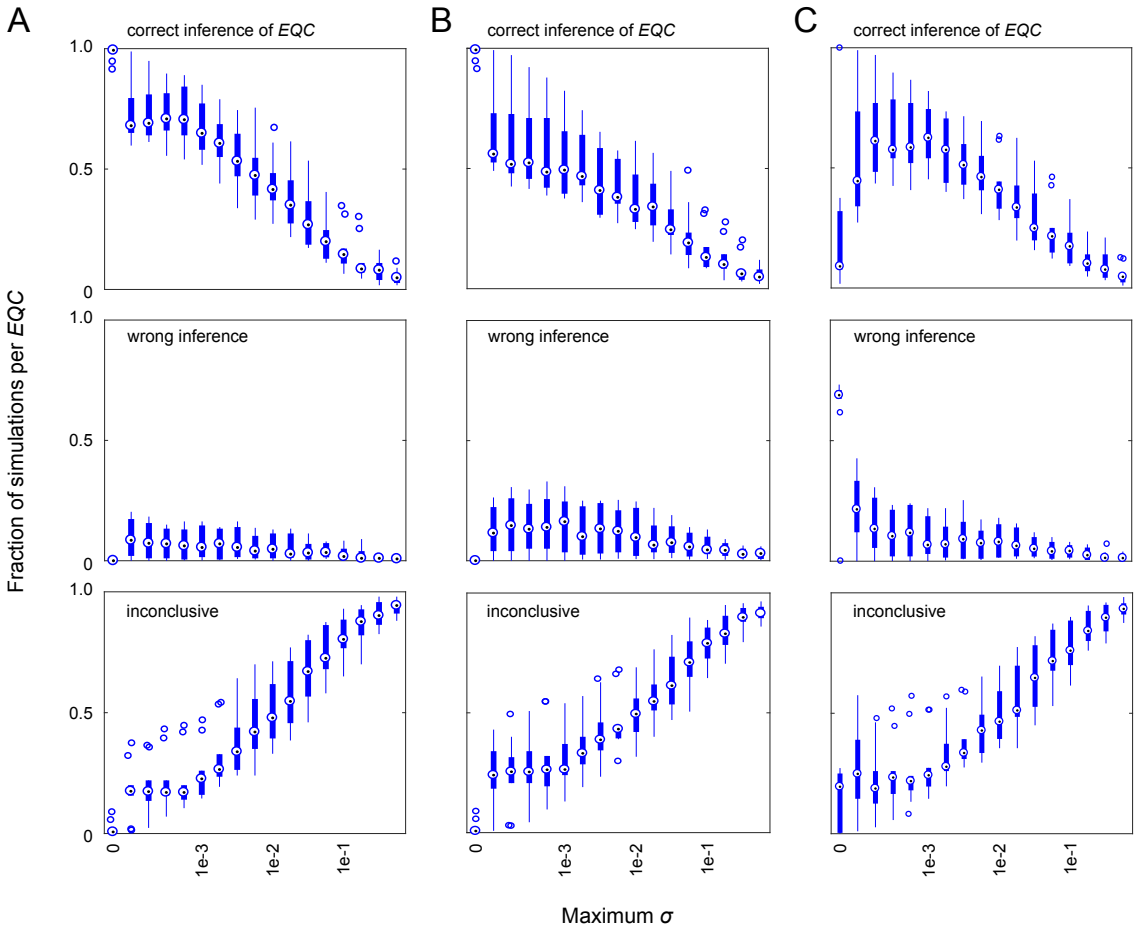

Figure 3.11: Noise-dependent thresholds reduce noise-induced inference errors. Data from Figure 3.10B is re-analyzed with modified criteria for numerical equivalence detection. Equivalence is identified when confidence intervals of one standard deviation from the mean (**A**) or one MAD from the median (**B**) overlap for a given pair of traits, or when the ratio is less than one of absolute trait difference to an estimate of combined standard deviation (**C**). Symbols: boxplot quantiles, as defined for Figure 3.8.

Interestingly, the decrease in wrong inferences by noise-dependent threshold testing is nonetheless accompanied by an increase in inconclusive inferences. One may view this as a strength because it is expected to lower false discovery rates. In the context of deriving hypotheses from the analysis of experimental data, a wrong inference may in many cases be worse than stating that no conclusion can be drawn.

To further explore the origin of inconclusive inferences, I re-analyze the inference results with the following assumption: inconclusive inferences correspond to patterns of numerical equivalence wherein all theoretically equivalent traits are numerically equivalent with additional spurious numerical equivalences. In application to inconclusive inference results, the assumption implies that a *best-fit* inferred topology is the topology with all theoretical equivalences found to be numerically equivalent in an inconclusive inference.

If the *best-fit* assumption is true, re-analysis of the inconclusive inferences in Figure 3.11A should result in the change of every inconclusive inference to a correct inference. In application to these data, I find the expectation is correct for $\sim 36\%$ of inconclusive inferences at all levels of uncertainty, allowing an improvement in inference compared to the application of noise-dependent thresholds alone.

The success of *best-fit* inference is maximized when the noise-dependent threshold comprises confidence intervals of estimated standard deviations multiplied by an optimal z factor (Figure 3.12A, inset). I find this factor to be near 2.5. The inference results with a z factor of 2.58 are shown in Figure 3.12A. I find that the assumption of transitivity among numerical equivalence groups reduces false positive inferences at low noise, but increases false positive inferences at high noise (Figure 3.12C). I compare these results to inferences achieved with a topology score applied to the same dataset (Figure 3.12B).

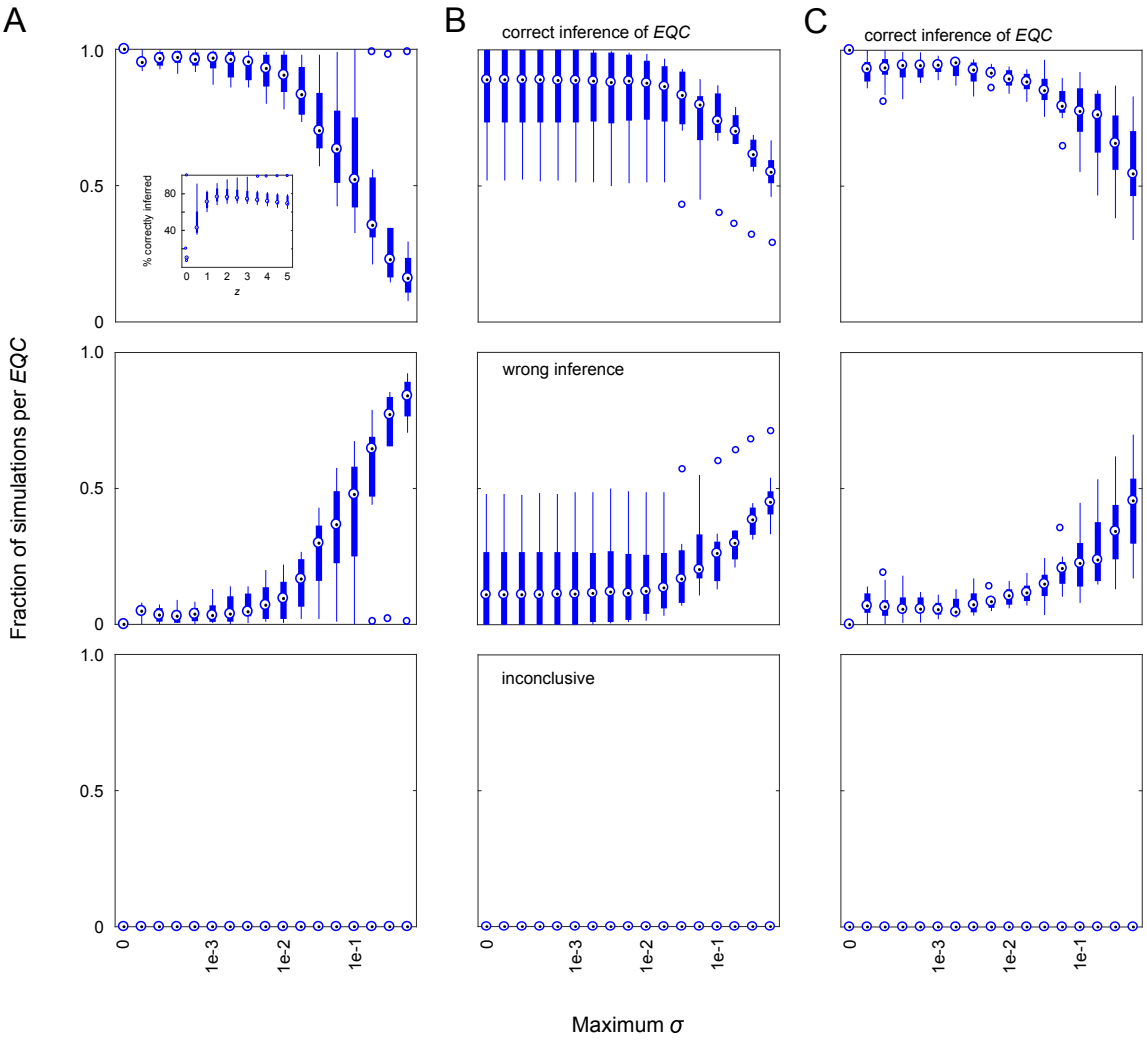

Figure 3.12: Combining noise-dependent thresholds and best-fit assumptions further reduces noise-induced inference errors. Data from Figure 3.10B is re-analyzed with modified criteria for inference. **A.** Numerical equivalence is identified when confidence intervals of 2.58 standard deviations from the mean overlap for two simulated traits with noise, and inconclusive inferences by this criterion is forced to a wrong or correct inference by making the *best-fit* assumption (See text). The z value of 2.58 was selected because it obtains the highest number of correct inferences with *best-fit* assumptions across all grades of trait noise (inset). This inference can be compared to inference of the same data by topology score (**B**), or when the assumption of numerical equivalence transitivity is not applied (**C**).

3.3.9 Topology inference from experimental data

To further compare *best-fit* and topology score methods, I apply the methods to experimental genetic interaction data generated for a well-studied biological system. In this system, I consider transcription of the *RNR3* gene as a trait reporting on activity of the DNA damage checkpoint pathway. I consider a subset of four genes in this pathway, encoding three kinases and one transcription factor. A model of the pathway, which will act as a set of three true positive inferences, is shown in Figure 3.13A. I do not consider the inference of the three auto-regulatory feedback edges.

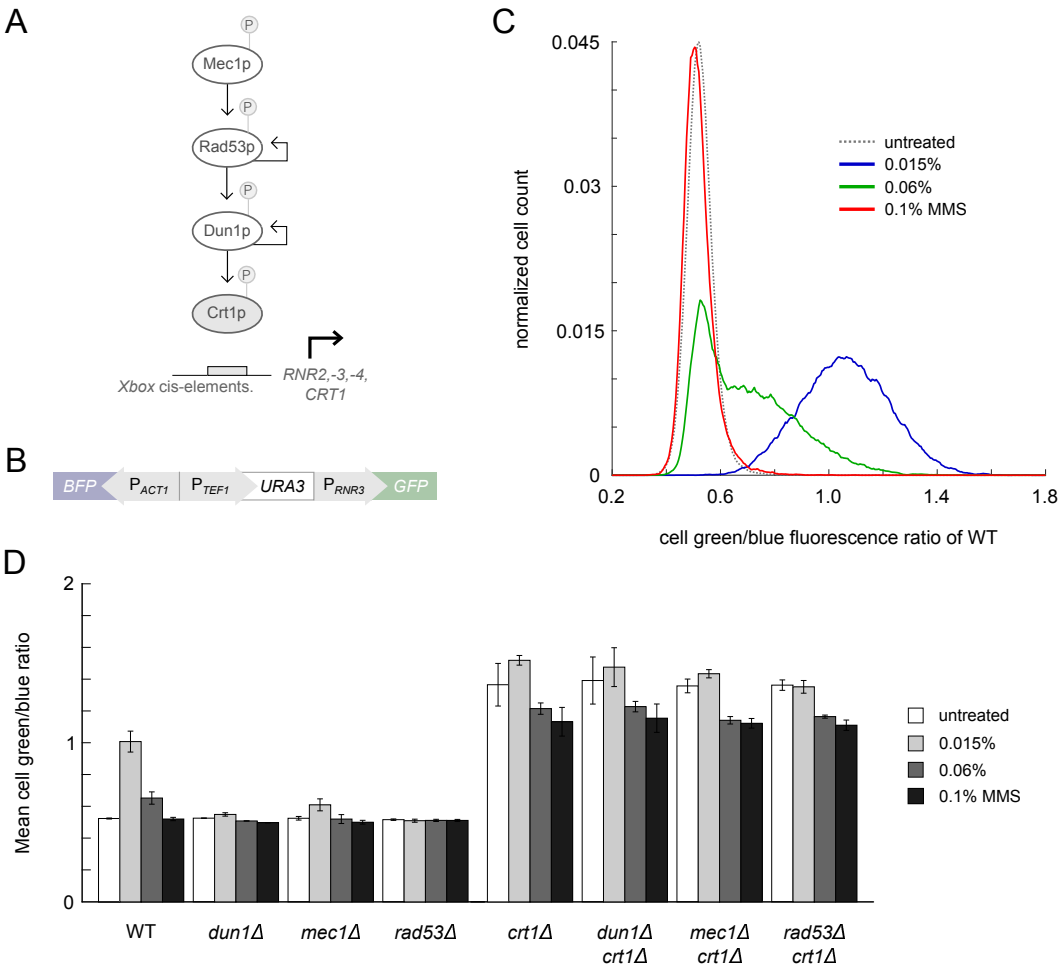

Figure 3.13: DDC gene network model as an experimental test case. Four-gene model (**A**) for DNA damage-induced derepression (bold arrow) of *RNR3* transcription as represented in current literature. Derepression requires gene-encoded kinase (white oval) or DNA binding (grey oval) activities, regulated by phosphorylation (P) by upstream kinases (arrow) or auto-phosphorylation (feedback arrow). **B.** Our quantitative phenotype to report on network activity is *RNR3* transcription, achieved by integrating the DNA construct shown at the *sml1*Δ locus of haploid cells. The construct contains a growth selection marker (*URA3*) and transcriptional control (*ACT1*). **C.** Ratios of green to blue fluorescence of cells with one copy of **B** indicate that low MMS treatment (240 min) causes cells to derepress *RNR3* transcription relative to untreated, whereas high MMS does not, and intermediate MMS causes variable transcriptional responses that result in a multimodal population distribution between that of low and high MMS-treated cells. Data are obtained by flow cytometry for wildtype cells, after gating and linearization of log fluorescence measurements (see Methods §3.5.7). **D** Genetic interaction data consists of means (bar height) and standard deviations (± error bar) of population means for replicate measurements of corresponding gene and signal perturbation conditions (data obtained as in **C**).

I consider the mean cellular ratio of green to blue cell fluorescence as the trait. This measure estimates the mean ratio of transcription from the *RNR3* promoter to the control *ACT1* promoter (Figure 3.13B), and is analogous to a previous analysis by Jonikas et al [36]. Trait value calculation from flow cytometry acquired fluorescence distributions is described in Methods.

I consider methyl methanesulfonate (MMS) as the stimulus of the pathway. I test three concentrations applied in previous studies. I find that 240 min exposure to a low concentration of MMS results in derepression of *RNR3* transcription, as expected. However, this response is reduced when cells are exposed to higher concentrations (Figure 3.13C).

Intermediate and high concentrations of drug yield responses in disagreement with findings of previous studies. These responses suggest one or all of following assumptions is false: the model of the pathway in Figure 3.13A is correct, derepression of *RNR3* transcription is positively correlated with activation of the pathway, activation of the pathway is positively correlated with drug concentration. While the present chapter will examine the first assumption, the next chapter of this thesis describes experiments to test the latter two.

The set of genetic interaction data I analyze consists of the mean and standard deviation of replicate measurements of the population mean green to blue fluorescence ratio (Figure 3.13D), obtained for each combination of gene pair deletion and stimulus condition.

To examine the first assumption, I infer *EQC* by *best-fit* or topology score from these data when low or high drug concentration is the stimulus. If the first assumption is true, I expect to infer *EQC2* for each of three gene pair analyses, and both stimulus conditions. In contrast, I find that *EQC2* can be inferred by both methods when low

drug concentration is the stimulus (Figure 3.14A), but not when high concentration is the stimulus.

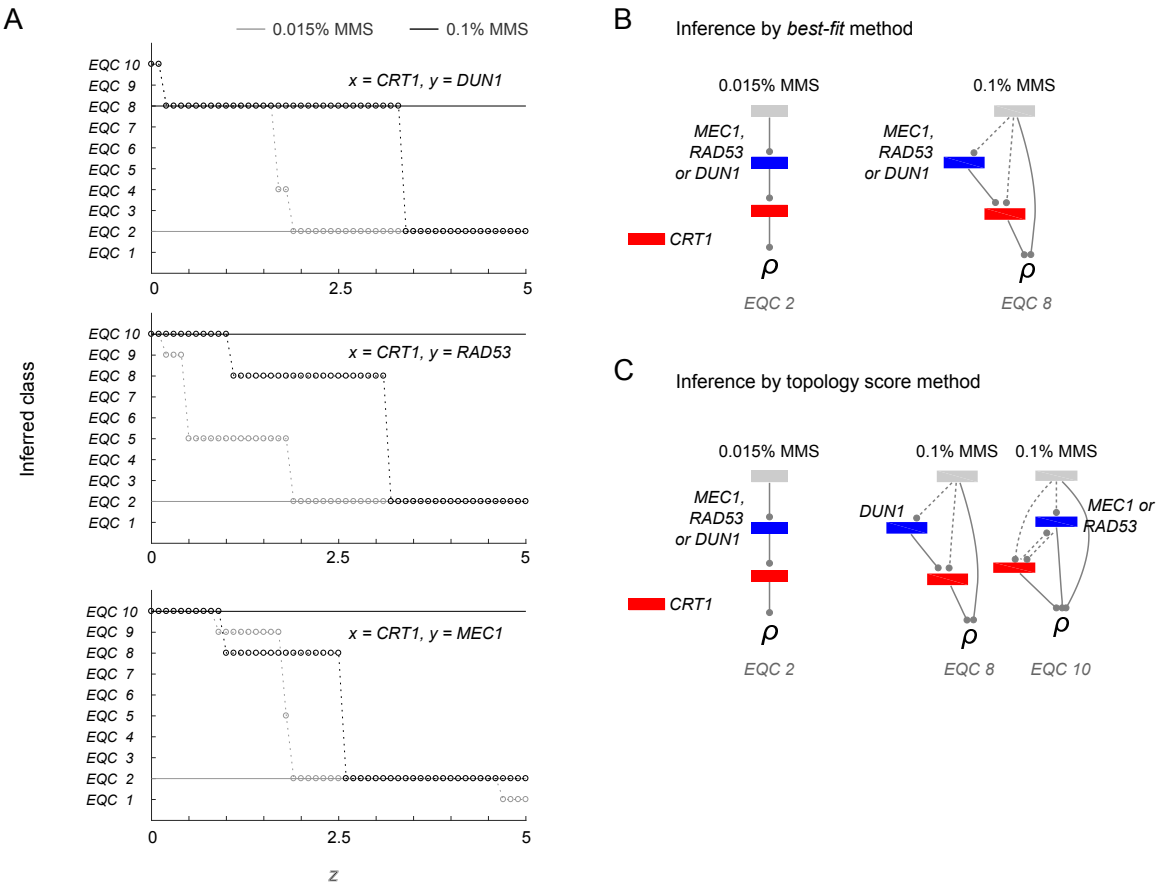

Figure 3.14: DDC gene network model inferences. A. *EQC* Inferences by *best-fit* with varying z value (circles) are compared to inferences by topology score (solid lines) when low MMS (grey) or high MMS (black) is the signal tested. **B.** Diagrams corresponding to inferences by *best-fit* when z has a value of 2.58. **C.** Diagrams corresponding to inferences by topology score.

Although both methods allow inference of *EQC2* with low drug concentration, inference by *best-fit* requires a high z factor. If I consider the inference at a z value optimal for the analysis of simulated traits with uncertainty (Figure 3.12A), I find that *EQC2* can be inferred for all three gene pairs (Figure 3.14B).

In contrast, corresponding analyses of high drug concentration data result in the inference of *EQC10* or *EQC8*. The consistent edge among both inference methods and all gene pairs is the influence of the stimulus on the trait independent of the pathway. This hypothesis is tested in the next chapter of this thesis.

3.4 Discussion

The starting point of the present study is the analysis of topology inference from simulated genetic interaction data when applying a topology score derived from the method of Battle et al. [60]. This analysis revealed limits in how well one could identify an acyclic topology among the 35 possibilities.

To discern potential causes of the limits of identifiability by topology score, I developed an approach to determine theoretical topology identifiability. This approach generalizes classical epistasis analysis rules in derivation and application. I derive theoretical trait expressions by assuming node activity is any continuous function of all of its upstream nodes' activities. In contrast classical epistasis analysis rules are derived by assuming node activity is a Boolean variable equal or opposite to the node activity of its single upstream node. In application, I consider a topology to be identifiable when the topology has any unique equivalence pattern among topology-derived trait expressions of the eight conditions (see Tables and 3.1 and 3.2). In contrast, classical epistasis analysis involves the identification of one of two possible pairs of equivalent traits, with mutually exclusive identification. I tested

whether theoretical patterns of trait equivalences could also be identified numerically in simulated trait data, with or without trait uncertainty, thereby allowing topology class *EQC* inference.

I tested the effect of trait uncertainty on simulated trait equivalence identification by two approaches. First, I test the addition of the same uncertainty to all eight traits of a simulated topology. This approach ignores the potential for traits having different uncertainty levels, as well as noise propagation to downstream nodes [78] (since noise is only added to trait), noise propagation over time (since noise is added to steady-state trait values), and noise propagation among different processes underlying each node activity (e.g. mRNA and protein production). Second, I test the addition of different uncertainty to eight traits of a simulated topology, by randomly sampling standard deviations from 17 values between zero and 0.75. The second approach may overestimate differences in uncertainty in a dataset but nonetheless served as benchmark for inference errors induced by trait uncertainty. Based on this benchmark, I found applying a noise-dependent threshold for equivalence testing combined with a *best-fit* assumption, referred to as the *best-fit* method, to allow lowest inference errors across the range of uncertainties I test.

The *best-fit* method for *EQC* inference is highly reliant on accuracy of equivalence identification. Applying the method requires 28 pairwise numerical equivalence tests. I note that these tests are conceptually analogous to the identification of equivalent or non-inferior medical treatments by an equivalence or non-inferiority clinical trial, respectively. In these trials, the standard of reporting [79] is to compare differences in patient outcomes following treatment to a clinically relevant margin of difference. It is the author's opinion that false positive errors of equivalence detection are sensitive to the selected method for incorporating statistical measures of confidence into the difference or ratio of two treatments' outcomes. For example, a method which considers two treatments to be equivalent when the confidence

interval of ratio of the treatments' outcomes is lower a margin of clinically relevant difference [80], is less prone to identify equivalence with increasing uncertainty. In contrast, identifying two treatments as equivalent when their confidence intervals overlap is more prone to identify equivalence with increasing uncertainty. I base my selection of method of noise-dependent threshold testing on true positive inferences from simulated data. In the context of experimental data, I do note that I examine uncertainty in the context of confidence intervals representing uncertainty in means of replicate fluorescence distributions (discussed below). Methods directly comparing probability or cumulative sample distributions rather than means may improve the analysis of equivalence.

I applied to experimental genetic interaction data the methods of inference allowing highest chance of correct *EQC* from simulated data with different uncertainties. I compared results of *EQC* inference by topology score or *best-fit* method ($z = 2.58$). The methods allowed true positive inference for all gene pairs with low MMS as the signal. Both methods allowed *EQC8* inference for gene pair $\{CRT1, DUN1\}$ with high MMS as the signal, wherein I do not have a known true positive. Inference by the two methods diverged for gene pairs $\{CRT1, MEC1\}$ and $\{CRT1, RAD53\}$ with high MMS as the signal, wherein I do not have a known true positive.

I further examine the cases wherein the two methods diverge in inference and no true positive is available. Figures 3.15A and 3.15B provide visual inspection of fluorescence distributions underlying traits which should be equivalent if *EQC8* is the correct *EQC*. I contrast the case wherein *EQC8* is inferred by both methods or (Figure 3.15A) versus only inferred by the *best-fit* method (Figure 3.15B). It is the author's opinion that differences in inference between the alternative methods may be explained by the differences in underlying assumptions. For example, the *best-fit* method only considers equivalence patterns, whereas the topology score method weighs equivalence

(full dependency) to alternative dependency types such as the multiplicative product rule (independence).

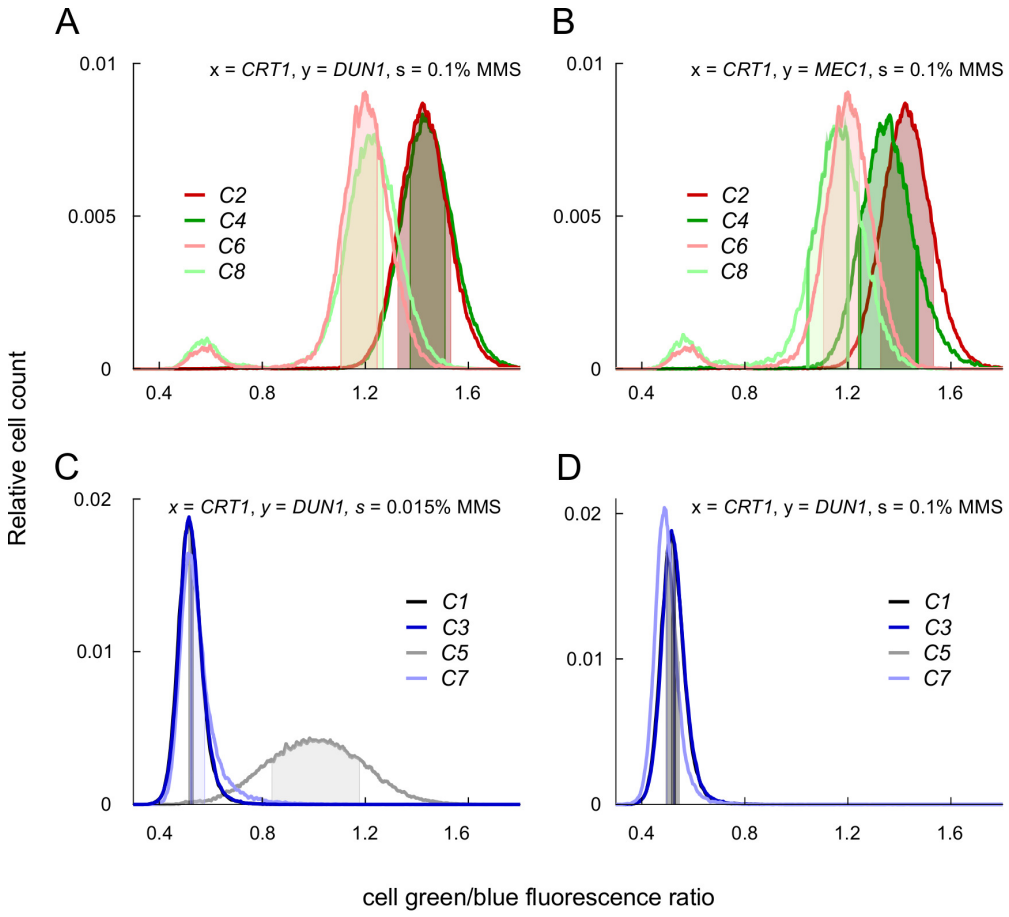

Figure 3.15: Cell fluorescence distributions underlying experimental trait equivalences identified. I examine the distributions underlying inferences shown in Figure 3.14. For data in **A** *EQC8* was inferred using topology score and *best-fit* method. Theoretically equivalent trait groups of *EQC8* are $\{C2, C4\}$ and $\{C6, C8\}$. In contrast, for data in **B** *EQC8* was inferred using *best-fit* method but *EQC10* with a topology score. There are no theoretically equivalent trait groups for *EQC10*. For data in **C**, equivalence group $\{C1, C3, C7\}$ was identified although the *best-fit EQC* inferred (*EQC2*) has theoretical equivalence group $\{C3, C7\}$. For data in **D** equivalence group $\{C1, C3, C5, C7\}$ was identified although the *best-fit* inferred *EQC* (*EQC8*) has none of these equivalences. Symbols: plot line, mean relative histogram count of replicate experiment distributions with bin size $2\text{x}10^{-3}$; shaded area, confidence interval used to calculate noise-dependent equivalence threshold when z is 2.58.

The *best-fit* assumption applied to infer *EQC*s from inconclusive numerical equivalence patterns relies on the expectation that data are derived from a topology among the ten *EQC*. The assumption is that spurious false positive equivalences are caused by either trait uncertainty or imperfect method of equivalence detection. When this expectation is false or unknown (e.g. for experimental data), the *EQC* inference can be wrong. Figures 3.15C-D provide visual inspection of fluorescence distributions underlying *best-fit* inferences from experimental data when there was disagreement between identified and theoretical equivalences of the inferred *EQC* (Figure 3.15C-D). Visually, it appears that "spurious" equivalences filtered by the *best-fit* method correspond to distributions more or as similar as distributions underlying "true" equivalences. This suggests that the *best-fit* assumption is wrong, and the data is not representative of any *EQC* topology. I note an alternative possibility. I find that including additional assumptions for the derivation of theoretical trait expressions allows additional theoretical equivalences as well as higher topology identifiability across all topologies. This finding suggests that additional numerical equivalences may not be spurious, but reflect a need to consider additional assumptions to derivation of trait expressions. I do not present a comprehensive analysis of how additional assumptions modify *EQC* derivation.

Repression of *RNR3* transcription in cells treated with high (0.1% v/v) MMS was not anticipated in my examination of experimental data and therefore I do not have true positive expectations for topologies involving the high MMS signal. I found high MMS data yielded inference of *EQC8* or *EQC10*. A DDC-independent edge, from the signal to the trait, is conserved among all topologies in these *EQC*s. Influences on the *RNR3* promoter independent of the DDC may be discerned by examining an alternative trait measurement, for example a constitutive promoter having a single integrated *CRT1* DNA binding element.

I anticipated MMS dose from 0.015% to 0.1% v/v (1.77 to 11.80 mM) to cause either a monotonic increase or a saturation in *RNR3* transcription. Several previous studies support the expectation that transcription is inducible at the highest concentrations considered in the present study (i.e. 0.1 % v/v). Northern blots in [81] showed induction of *RNR3* transcription after 1 h of exposure to 0.1% v/v, relative to untreated cells. Northern blots in [82] showed induction of *RNR3* and *CRT1* gene transcription following 0.1% v/v MMS exposure (30 min to 2 h). Several studies suggest kinases functioning upstream of *RNR3* are catalytically active in cells treated with 0.1% v/v for 2 h. Dun1p extracted from treated cells could be auto-phosphorylated *in vitro*, in contrast to protein extracted from untreated cells [81, 83]. Rad53p kinase extracted from treated cells had a phosphorylation induced electrophoretic shift, in contrast to Rad53p extraction from untreated cells [84].

In contrast, I can also identify previous studies agreeing with our finding that fluorescence is not inducible in 0.1% v/v MMS treated cells. Studies [81] and [83] showed by Western blot endogenous levels of Dun1p to be reduced or undetectable when extracted from cells treated with 0.1% v/v MMS (2 h), in contrast to untreated or hydroxyurea-treated cells. This suggests high MMS may inhibit expression or cause degradation of Dun1p. Studies employing LacZ [85–87], Cypridina luciferase [88] or GFP [89] reporters of the *RNR3* promoter report non-monotonic responses to MMS as we observe, wherein 0.01 to 0.02% v/v MMS treatment results in 15 to 55 fold higher reporter levels than untreated, whereas 0.1% v/v MMS treatment results in 1 to 4 fold induction relative to untreated. The results of study [90] suggest translation may be repressed by high but not low MMS doses. This finding may explain the discrepancy between studies measuring endogenous transcript versus reporter protein level.

3.5 Methodology

3.5.1 Simulated trait values

For each of 35 topologies I consider 100 simulated genetic interaction datasets. The datasets are from [74], obtained by adaptation of the method of [76] by Mads Kærn. Each dataset is defined by topology-independent assumptions of the simulation method, topology-dependent parameters, and topology-independent parameters. Each of 100 simulations per topology differs in the latter.

To simulate genetic interaction data for a topology in absence of noise, each node i ($i \in \{X, Y, \rho\}$) in the topology except the signal is assigned two ODEs:

$$\frac{dRNA_i}{dt} = m_i \cdot f_i - \lambda_i^{RNA} \cdot RNA_i \tag{3.1}$$

$$\frac{dProt_i}{dt} = r_i \cdot x_i - \lambda_i^{Prot} \cdot Prot_i \tag{3.2}$$

They define the rate of change of the concentration of mRNA (RNA_i) and protein ($Prot_i$) abundance, respectively, for node i. The activity of a node corresponds to $Prot_i$. Genetic interaction data represent the steady-state abundances of $Prot_\rho$ under the defined perturbation conditions.

Topology-independent parameters include the maximum transcription m_i or translation r_i rate, and mRNA decay λ_i^{RNA} or protein decay λ_i^{Prot} rate. In a given simulation these parameters are randomized by sampling from a normal distribution. Parameters m_i and r_i are each assigned a value between $ln(2)/22$ and $ln(2)/20$. Parameters λ_i^{RNA} and λ_i^{Prot} are then assigned as m_i and r_i, respectively.

To model deletion of gene i, m_i is set to zero. To model loss of signal, signal activity is set to zero. In absence of perturbation, the signal is a constant.

Topology-dependent terms are encompassed in f_i. The equation for f_i depends on the number of upstream input edges to i. When there is no upstream input, e.g. i is constitutively expressed, f_i is defined by a constant. When i has one upstream input j, $Prot_j$ is modelled as the abundance of an upstream transcription factor (TF) which can bind one cis-regulatory module (CM) in the transcriptional promoter of i, thereby modifying its expression. When i has multiple inputs, each additional input represents an additional TF, which binds an additional CM. TF-CM binding events are modelled to be thermodynamically independent, and thereby independently contribute to the quantity of f_i.

To achieve this, the number of CMs considered in f_i is equal to the number of edges which are input to i in the network, and f_i is the sum of probabilities of transcription for each possible configuration of the promoter. Configuration is a permutation of the combinations of each TF being bound or unbound to its respective CM. Therefore there are 2^n configurations for a promoter with n input edges ($n \leq 3$).

Thus, in the case a node ρ has the maximum of three inputs S,X and Y, f_i is calculated as follows

$$f_i = \sum_{s,x,y} \alpha \cdot \mathrm{fc}_s \cdot \mathrm{fc}_x \cdot \mathrm{fc}_y \cdot CM_s \cdot CM_x \cdot CM_y \tag{3.3}$$

where $s, x, y \in \{0, 1\}$ define the possible promoter configurations wherein each input S, X and Y, respectively, is bound (one) or unbound (zero), where α is the basal transcription level when all CMs are unbound and fc_s is the fold change in transcription relative to α when S is bound ($s = 1$) or unbound ($s = 0$). $CM_{s=1}$ is the probability of transcription when the TF representing input S is bound $CM_{s=1} = P_{CM_s}$ whereas

$CM_{s=0}$ is the probability it is unbound $CM_{s=0} = 1 - P_{CM_s}$. Therefore, the probability of transcription of a given configuration is the product of probabilities of transcription for each CM.

For a given upstream input j, probability of transcription from its corresponding CM of promoter i, $P_{CM_{ji}}$ depends on the protein abundance of j ($Prot_j$) according to the Hill equation as follows

$$P_{CM_{ji}} = \frac{\left(\frac{Prot_j}{K_{ji}}\right)^{nH_{ji}}}{\left(\frac{Prot_j}{K_{ji}}\right)^{nH_{ji}} + 1} \tag{3.4}$$

if node j encodes an activator, or

$$P_{CM_{ji}} = \frac{1}{\left(\frac{Prot_j}{K_{ji}}\right)^{nH_{ji}} + 1} \tag{3.5}$$

if node j encodes a repressor, where nH_{ji} is the Hill coefficient and K_{ji} the dissociation constant. Both nH_{ji} and K_{ji} are topology-independent parameters. In the simulation, nH_{ji} is randomly assigned between 1 and 3 (uniform sampling) and K_{ji} randomly assigned between 0.1 and 0.8 (uniform sampling). $P_{CM_{ji}}$ satisfies the assumption of thermodynamic independence since the probability depends on a single input.

For any simulated topology, the number of repressors and activators is randomized between zero and three (i.e. all nodes excepting ρ) by uniform discrete distribution sampling. This selected number then influences the assignment of constants reflecting basal transcription level (α). If all nodes excepting the trait are repressors, α is randomly assigned between 0.9 and one (uniform distribution), whereas if all activators α is assigned between zero and 0.1, and if mixed repressors and activators α is assigned between 0.45 and 0.55.

Similarly to α, each fold change value in Eq 3.3 (fc_i) depends on whether node i is a repressor or activator, as well the number of total inputs of a promoter. As indicated for Eq 3.3, fc_i is dependent on promoter configuration. When i is bound in the configuration, $\text{fc}_{i=1} = r$ if i is an activator or $\text{fc}_{i=1} = 1/r$ if i is a repressor. The parameter r is randomly assigned (uniform distribution) based on the number of inputs modelled to bind to the same promoter as i. If i is the only input, r is assigned between 8 and 12, whereas if there are two inputs r is between 4 and 6, and if three inputs r is between 8/3 and 4. Following these assignments, each fc value for a given promoter is rescaled by dividing by the maximum product of the assigned α and all $\text{fc}_{i=1}$ terms considered. Independently of whether a node i is an activator or repressor, $\text{fc}_{i=0}$ is set to one and is not considered in the above calculation or rescaling.

To obtain genetic interaction data once all parameters have been assigned, steady-state levels of $Prot_\rho$ are simulated for each set of perturbations. Each simulated genetic interaction dataset qualifies certain constraints, as follows: (1) all individual trait values are within the range of zero to one; (2) perturbation of the signal causes a minimum of two fold change in trait value; and (3) perturbation of each gene causes a minimum of 20% change in trait value. Notably, parameter selection may be biased to allow these constraints since simulations not qualifying these constraints were discarded.

3.5.2 Topology score calculation

The topology score is a sum of all dependency scores. Dependency scores are calculated for every pair of nodes in the topology excluding ρ. For a hypothetical pair of nodes, X and Y, the dependency score is the absolute difference between an experimental trait value when both nodes are absent $T_{\Delta x \Delta y}$ and an expectation value $\epsilon_{\Delta x \Delta y}$. ϵ is calculated according to the relationship between X and Y in a hypothetical topology.

For a dependency of type m, the dependency score is $\delta^m_{x,y} = |\epsilon^m_{\Delta x \Delta y} - T_{\Delta x \Delta y}|$. For a hypothetical topology γ, the topology score is $TS_\gamma = \sum_1^n \delta_n$, where n is the number of unique node pairs.

There are three possible dependency types. A given node pair can be of only one type in a topology. The dependency type dictates the expectation value and therefore modifies the dependency score of a node pair in a topology-dependent manner. For nodes X and Y ($X \neq Y, X, Y \neq \rho$), the dependency type is identified based on topology edges and reachability.

The following procedures describe how to identify each dependency type in a topology and calculate the corresponding expectation value:

1. **Full dependency of X on Y.** To identify in topology: ρ is reachable from X and Y, Y is downstream (reachable) from X. If Y is removed, ρ is not reachable from X, i.e. the influence of gene X on ρ is fully dependent on Y.

   - Expectation value: $\epsilon^f_{\Delta x \Delta y} = T_{\Delta y}$,

   where $T_{\Delta y}$ is the trait value when only node Y is perturbed.

2. **Independence of X and Y.** To identify in topology: ρ is reachable from X and Y, neither X or Y is reachable from the other. If Y is removed, the trait is reachable from X, and vice versa, i.e. X and Y can independently influence ρ.

   - Expectation value: $\epsilon^i_{\Delta x \Delta y} = T_{\Delta x} * T_{\Delta y} / T_{xy}$,

   where $T_{\Delta x}$ and $T_{\Delta y}$ are the trait values when only node X or Y is perturbed, respectively, and T_{xy} is the trait value when neither X nor Y is perturbed.

3. **Partial dependence of X on Y.** To identify in topology: ρ is reachable from X and Y, and Y is downstream (reachable) from or X. If Y is removed,

ρ is reachable from X, i.e. the influence of gene X on the trait is partially dependent on Y.

- Expectation value: $\epsilon^{p}_{\Delta x \Delta y} = [\epsilon^{i}_{\Delta x \Delta y} + \epsilon^{f}_{\Delta x \Delta y}]/2$.

I derived the topology score method by excluding or modifying a number of terms included in a previously described scoring method [60]. In particular, I did not consider prior probabilities of dependency types (useful when subsets of genetic interaction data are missing), nor the Pearson correlation coefficient of two nodes' vectors of genetic interaction terms with remaining nodes of the topology (calculable when topology size is larger than we consider). I modified how one calculates deviations between dependency type expectation value and trait value. While I calculate the absolute difference, [60] calculated the ratio between this difference and a constant reflecting trait value uncertainty. I also modified the calculation of the expectation value for the partial dependency type, as described §3.3.2 of Results. Because of these modifications, the present study is not an analysis of inference errors of the previous method.

3.5.3 Theoretical trait expression calculation

To derive the expression of the theoretical trait ρ I defined each node i's activity ($i \in X, Y, \rho$) as an arbitrary function f_i of the activities of its upstream nodes connected by an edge. In absence of input a node j's activity may be equal to a basal term b_j ($j \in X, Y$) or a term representing the activity of the signal S. The eight node perturbation conditions C indicate which of S,X, and Y is deleted/absent (0) or not (1), and may alter the trait expression. If a node is deleted in C, its activity is set to zero in the theoretical trait expression corresponding to C. Trait expressions

were derived as strings using the software package MATLAB (MATLAB 9.0 2016a, The MathWorks, Inc., Natick, Massachusetts, United States).

3.5.4 Equivalence class definition

Equivalence classes were defined by classifying each topology according to whether it had a unique trait equivalence pattern, therefore forming a new class, or a pattern defined by other topologies of an identified class.

For each topology, the theoretical trait expression equivalence pattern was determined by identifying matching strings among the eight strings representing all trait expressions for eight perturbation conditions for a topology.

3.5.5 Simulated trait values with noise

Trait values with noise are determined from trait values in absence of noise (see §3.5.1). For each simulated trait value in absence of noise, I uniformly sample three values from Gaussian distribution P_ρ with a mean μ, equal to the trait value in absence of noise, and a standard deviation σ, as defined in figures. The corresponding trait value with noise is equal to the mean or median of the values sampled from P_ρ, rounded to four digits to the right of the decimal point.

I analyze 16 non-zero values of σ. These values increase by one quarter of an order of magnitude between $1.0x10^{-4}$ and $7.5x10^{-1}$. When trait values of a given topology are simulated to have different levels of uncertainty, this is achieved by uniform sampling from these 16 values between zero and the maximum value as defined in figures. Including σ equal to zero, there are 17 possible noise conditions considered. Analysis of all conditions of noise for all simulated datasets consists of 4.76 x 10^5 simulated

trait values. Most values reflect a coefficient of variation (CV) less than 0.1 (89%), and nearly all traits have a CV less than 2.6 (>99%).

Adding noise to the data in this manner assumes that σ is analogous to the standard deviation among experimental replicate means.

3.5.6 Numerical trait equivalence identification

Equivalence among simulated trait values is identified by a single threshold (1) or a noise-dependent threshold (2).

1. With a single threshold θ, trait values t_i and t_j are considered equivalent when $|t_i - t_j| \leq \theta$.

2. I consider three variants of noise-dependent thresholds (a to c).

   (a) For the first variant, traits t_i and t_j are considered equivalent when $t_i + z \cdot \bar{\sigma}_i \geq t_j - z \cdot \bar{\sigma}_j$, where $t_i \leq t_j$, t_i is the estimated mean and $\bar{\sigma}_i$ is the estimated standard deviation of values sampled to obtain the value of trait i with noise (see §3.5.5), and z is defined in Figures.

   (b) For the second variant, traits t_i and t_j are considered equivalent when $t_i + \overline{MAD}_i \geq t_j - \overline{MAD}_i$, where $t_i \leq t_j$, t_i is the estimated median and $\overline{MAD}_i$ is the estimated median absolute deviation of values sampled to obtain the value of trait i with noise (see §3.5.5).

   (c) For the third variant, traits t_i and t_j are considered equivalent when $\frac{|t_i - t_j|}{\sqrt{\bar{\sigma}_i^2 + \bar{\sigma}_j^2}} \leq 1$, where t_i and $\bar{\sigma}_i$ are as defined for the first variant.

Equivalence among experimental trait values is identified by a noise-dependent threshold analogous to the first variant applied to simulated trait values, where t_i is

the mean and $\bar{\sigma}_i$ the standard deviation of replicate sample means. A sample mean is the mean cell fluorescence ratio of population of cells in a given replicate culture (see § 3.5.7).

Unless otherwise indicated, I assume all numerical equivalence relations are transitive. For example if I identify two numerical equivalence groups $\{a, b\}$ and $\{a, c\}$, I assume $\{a, b, c\}$ to be the true equivalence group.

3.5.7 Experimental genetic interaction data

Strain generation

Each strain considered was haploid with the construct of Figure 3.13B integrated by replacing the open reading frame of *SML1*. Integration was achieved by transformation of BY4741 with the DNA construct amplified by PCR with 60 bp primers, consisting of 20bp for construct amplification and 40bp homology to one of two loci flanking the *SML1* open reading frame. The DNA construct design and assembly is described in §4.5.1. The yeast transformation protocol is described in detail in [72].

Haploid gene deletion strains were obtained from diploid strains by sporulation and selection. First, a diploid wildtype strain was generated by mating a BY4741 MATa strain with genotype *sml1Δ0::RNR3pr-GFP/URAMX/ACT1pr-BFP*) to a BY4742 MATα strain without the construct and genotype *sml1Δ0::LEU2MX*. Then, additional genes were deleted in the diploid by transforming the diploid with plasmid-amplified *NATMX* or *KANMX* DNA with 40bp flanking sequences homologous to genomic DNA flanking the open reading frame of the target gene to be deleted [15]. Successfully transformed cells were identified by selection, followed by PCR validation. Haploid, single or double gene deletion strains were obtained by sporulating the

transformed diploid strain, followed by tetrad dissection, selection and a mating-type test. Standard yeast protocols for tetrad dissection and mating-type testing are described in [91].

All strains are listed in Table 3.4. I note that for certain genotypes, I did not obtain both haploid mating types. This cannot be explained by genetic linkage since *RAD53*, *MEC1*, *CRT1*, *DUN1* and *SML1* are located on different chromosomes. However, I hypothesize this is due to a combination of the rare finding of a spore of three selection markers and a specific mating type (1/16), as well as the potential for inviability of DDC gene deletion strains. Nonetheless, I did validate that mating type does not influence green or blue fluorescence in a wildtype haploid strain, and I do not expect DDC gene deletions to cause mating-type specific effects on fluorescence.

Table 3.4: Yeast strains

Strain	Genotype
BY4741	*MATa S288C his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 ho*
BY4742	*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*
yD01	BY4741 *sml1Δ0::RNR3pr-GFP/URAMX/ACT1pr-BFP*
yD02	BY4742 *sml1Δ0::RNR3pr-GFP/URAMX/ACT1pr-BFP*
	yD01 *crt1::NATMX*
	yD01 *crt1::NATMX dun1::KANMX*
	yD01 *mec1::KANMX*
	yD01 *mec1::KANMX crt1::NATMX*
	yD01 *rad53::KANMX*
	yD01 *rad53::KANMX crt1::NATMX*
	yD02 *crt1::NATMX dun1::KANMX*
	yD02 *dun1::NATMX*

Cell culture and MMS exposure

Cells were grown overnight (16h) in liquid culture and released prior to 4h exposure to MMS. Specifically, for each replicate experiment, a single colony from an agar plate with selective growth medium was used to inoculate 400 μl synthetic complete (SC) growth medium (recipe in § 4.5.2) with 2% w/v glucose and 0.042 g/l adenine hemisulfate (Sigma, Cat. no. A9126) in a 96 deep well plate. After 16 h growth (shaking 250 RPM, 30 degC), cell density was estimated by 596-604 nm light absorbance of a 1/20 diluted culture (Victor 3V Plate Reader, Perkin Elmer) and cells were then diluted to a density corresponding to an absorbance reading of 0.025 by using an small volume of cells to inoculate 400 μl SC medium in a deep well plate. After 2.5h growth (shaking 250 RPM, 30 degC), cells were diluted 1/2 with SC medium containing methyl methanesulfonate (Sigma, cat. no. 129925) to result in the final MMS concentration in v/v as defined in figures. Note that MMS was added to growth medium 5 min before cell exposure. After 4h growth, cell gene expression was measured by flow cytometry.

Flow cytometric measurements of gene expression

Flow cytometric measurements were acquired with a Cyan ADP 9 Analyzer (Beckman Coulter), equipped with a plate sampler having an automated fluidic system to generate air-gap-separated well samples for delivery to the flow cytometer [92] (HyperCyt® System, IntelliCyt Corp). HyperView® Analysis Software was used to obtain an individual CyAn Summit FCS3.0 file for individual well data [92]. Well data were analyzed using custom MATLAB functions. To obtain cell population gates, event data were analyzed to obtain a defined percentage (50%) of events for each sample that maximized event frequency per area in a 2D plot of Log pulse height forward scatter versus side scatter (488 nm). Within each gated population, cell blue fluorescence protein expression was estimated as the Log pulse height of 425-475 nm filtered emission with 405 nm excitation, and cell green fluorescence protein expression as the Log pulse height of 510-550 nm filtered emission with 488 nm excitation. Log height values having a range of zero to 1023 were rescaled to have a range of zero to one, linearized, and then rescaled to a range of zero to 1023. Distributions in Figure 3.13C consist of ratio of green to blue fluorescence values after gating, linearization and rescaling. Bar plots in Figure 3.13D consist of replicate means of sample means among cells of a green/blue fluorescence ratio distribution as in Figure 3.13C.

4

High variability in DNA damage responses is due to drug-induced disruption of a cell cycle checkpoint

Summary

In the previous chapter I observed that *RNR3* transcription, a marker of DNA damage checkpoint activity, was repressed by high concentrations of the DNA alkylating drug methylmethane sulfonate (MMS). To test the hypothesis that high drug concentration de-activates the checkpoint, I conduct biochemical, cell division and gene expression assays. These experiments advance knowledge about the MMS-dose dependency of cell division arrest, wherein high drug concentrations appear to allow cells to override a pre-mitotic checkpoint.

Contributions

The experiments and analyzes described in this chapter are unpublished. Dr. Mads Kærn, Dr. Kristin Baetz, Mila Tepliakova and I contributed to experimental design. Mila Tepliakova and I generated the strains required. Daniel Jedrysiak contributed to generating DNA constructs for DNA-damage responsive transcriptional reporters. Mila Tepliakova and I conducted all experiments, with exceptions. Immunoprecipitation experiments were designed, conducted and analyzed by Dr. Michael Downey and Christine Nwosu. Dr. Vera Tang conducted fluorescence-activated cell sorting. I analyzed and interpreted all experimental data with the advice of Dr. Mads Kærn. Dr. Mads Kærn developed algorithms for flow cytometry analysis. Dr. Adam Rudner contributed protocols, materials and instruments. Drs. Noel Lianga and Elizabeth Dodd-Moher provided advice for microscopy and tetrad dissection.

Contents

4.1 Background

Activity of eukaryotic DNA damage checkpoint (DDC) kinases is required for multiple cell responses to DNA damage including cell division cycle arrest and DNA repair gene transcription [81]. Activation of DNA repair gene transcription by the DDC is exemplified by genes encoding ribonucleotide reductase (RNR). In *S. cerevisiae*, one branch of the DDC supported by the literature is that wherein Mec1p kinase phosphorylates Rad53p kinase, phospho-Rad53p phosphorylates Dun1p kinase, phospho-Dun1p phosphorylates transcriptional repressor Crt1p, phospho-Crt1p allows activation of transcription of a subset of genes encoding ribonucleotide reductase namely *RNR2*, *RNR3* and *RNR4*. This branch (Figure 4.1A) is often depicted as a linear pathway schematic (for example [93]), although multiple auto-phosphorylation regulatory loops are well-documented [83, 94], as well as protein dimerization, and additional scaffold proteins that can modify phosphorylation signals conditional on the type of DNA damage stress [95, 96].

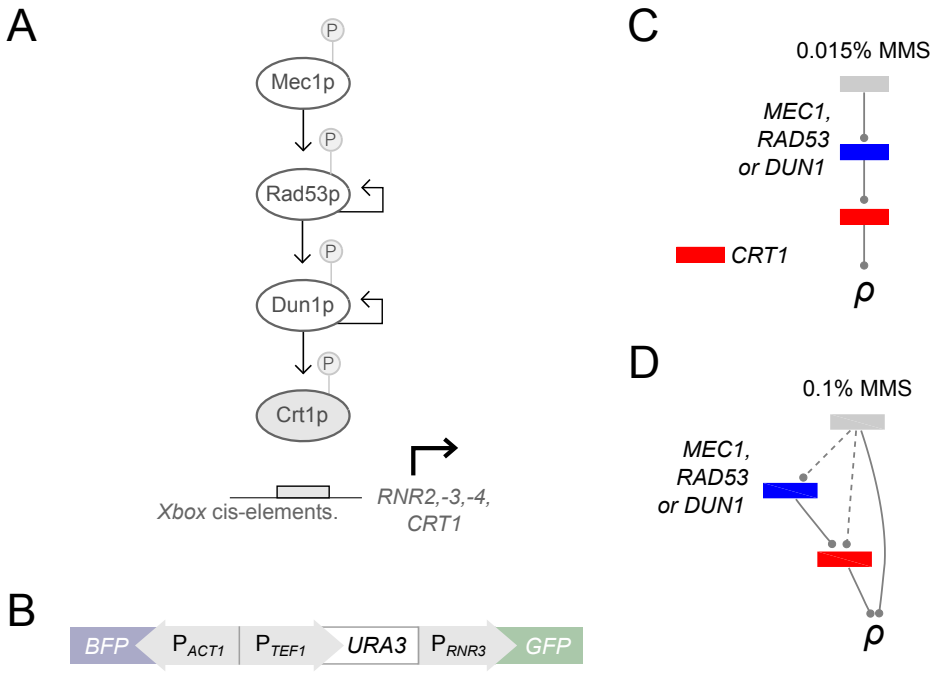

Figure 4.1: High MMS may repress *RNR3* transcription independently of the DDC. **A.** Four-gene model for DNA damage-induced derepression (bold arrow) of *RNR3* transcription as represented in current literature. Derepression requires gene-encoded kinase (white oval) or DNA binding (grey oval) activities, regulated by phosphorylation (P) by upstream kinases (arrow) or auto-phosphorylation (feedback arrow). **B.** A quantitative phenotype to report on network activity is *RNR3* transcription, achieved by integrating the DNA construct shown at the *met15*Δ locus of haploid cells. The construct contains a growth selection marker (*URA3*) and transcriptional control (*ACT1*). DDC gene topologies inferred from combinatorial DDC gene deletion experiments with low (0.015% v/v) MMS (**C**) agree with the model, whereas topologies inferred from experiments with high (0.1% v/v) MMS (**D**) yield the hypothesis that *RNR3* reporter expression (ρ) is DDC-independent. Full details of the experiments and inferences of **C** and **D** are provided in Chapter 3.

To estimate DDC activity, the transcriptional promoter of the *RNR3* gene (*pRNR3*) is often fused to reporter genes facilitating detection of expression by survival assays or fluorescence measurements. For example, expression of the *URA3* gene encoding orotidine-5'-phosphate decarboxylase can be detected by growth in media lacking uracil or absence of growth in media containing 5-fluorouracil (5-FOA) [82, 83]. Alternatively, expression of the LacZ gene encoding $\beta-$galactosidase can be detected by catalysis with the X-gal substrate resulting in colour change, or expression of fluorescent proteins can be detected by various methods of UV- to visible light spectroscopy [82, 83]. These assays show agreement with parallel estimates of endogenous transcript levels, for example by Northern blot [81, 82].

Knowledge about the regulatory diagram of the DDC pathway branch (4.1A) was primarily achieved by the Elledge group, beginning with systematic testing of mutations altering *pRNR3*-reporter expression. Genes *CRT1*, *SSN6* and *TUP1* were identified because their mutations resulted in constitutive RNR transcription (CRT) [97], whereas *DUN1* was identified because its mutation resulted in RNR transcription uninducible by DNA damage (DUN) [83]. These and the remaining interactions depicted in the pathway were deduced by examining DDC gene mutation or over-expression effects on protein phosphorylation, *in vitro* kinase auto-phosphorylation activity or *RNR* gene transcription with and without the presence of DNA damage [81–84]. Cellular DNA damage in these experiments was exemplified by exposure to hydroxyurea (10 to 200 mM HU) or methyl methanesulfonate (0.01 to 0.1 % v/v MMS) for one to four hours.

In the previous chapter (Chapter 3), I used this DDC pathway branch as a test case for topology inference methods. The data for inference consisted of fluorescence reporter expression measurements of a strain set having an $P_{RNR3} - GFP$ reporter and combinations of DDC genes in Figure 4.1A deleted. These measurements were obtained for four conditions of MMS exposure between 0% and 0.1% v/v (4 h). While

data obtained for low MMS (0.015% v/v) allowed inference that agreed with the model in Figure 4.1A, data obtained for higher MMS conditions ($\geq$ 0.06% v/v) did not.

Additionally, I found wildtype cells repressed reporter expression in response to these higher concentrations. This suggests that the DDC is inactive under these conditions. DDC inactivity of cells while exposed to DNA damaging conditions, as enabled by DDC gene mutations [81, 98] or DNA damage adaptation [99], is reported to have severe consequences on genome sequence maintenance and cell viability. These consequences are hypothesized to be primarily due to loss of cell division cycle arrest in a pre-mitotic state (G2/M checkpoint), since co-treatment of mutant cells with drugs that repress continued cycling can prevent cell death [81, 98, 100, 101].

I hypothesized that for these higher MMS conditions, one or all of following assumptions about this branch of the DDC pathway may be false: the model of the pathway in Figure 3.13A is correct, derepression of *RNR3* transcription is positively correlated with activation of the pathway, activation of the pathway is positively correlated with MMS drug concentration. The inference conducted in Chapter 3 allowed testing of the second assumption. For high (0.1% v/v) MMS, I inferred a topology differing from Figure 3.13A by the addition of a feedforward loop. This feedforward loop corresponds to the hypothesis that high MMS causes cells to repress *RNR3* transcription independently of DDC genes.

In the present chapter, I test this hypothesis as well as the alternative assumptions above-described. First, I conduct a detailed characterization of *RNR3* transcription as a function of MMS to examine its correlation with MMS dose. Second, I test if DDC-independent repression of *RNR3* can be explained by cell death or general repression of transcription and translation. Such toxicity has been suggested in previous studies [87, 88, 90]. Third, I test the possibility that transcription is repressed because

high MMS causes DDC inactivity. This possibility could not be determined in the previous chapter since transcription in cells with *DUN1*, *RAD53* or *MEC1* deleted was indistinguishable from wildtype cells (0.1% v/v MMS).

4.2 Objectives

- The first objective is to characterize the MMS dose-dependency of *RNR3* transcription in individual cells.
- The second objective is to test the hypothesis that *RNR3* transcription is repressed due to cell death or general repression of gene expression.
- The third objective is to test the hypothesis that *RNR3* transcription is repressed due DDC inactivation.

4.3 Results

4.3.1 *RNR3* transcription is a non-monotonic function of MMS

To quantify transcription influenced by the DNA damage checkpoint (DDC) signalling pathway in single cells *in vivo*, the following three expression cassettes were assembled in a 5.5kb DNA construct: the transcriptional promoter of the *RNR3* gene driving expression yeast-enhanced green fluorescent protein (P_{RNR3}-*yEGFP*), the promoter of *ACT1* driving expression of yeast-enhanced blue fluorescent protein (P_{ACT1}-*yEBFP*), and the promoter of *TEF1* driving expression of orotidine 5-phosphate decarboxylase (P_{TEF1}-*URA3*), an enzyme required for cell growth in absence of uracil. In yeast cells transformed with the construct, quantifying green and blue fluorescence of cells allows one to distinguish DDC-dependent and -independent transcription, respectively. *ACT1* transcript level is a marker of RNA polymerase II activity [102] and has been used as a normalization coefficient for DNA damage - [103–106] and cell division cycle-regulated [107–110] transcripts.

To characterize DDC-dependent transcriptional activation as a function of methyl methanesulfonate (MMS) exposure, yeast cells carrying the construct were treated with 24 MMS concentrations between zero and high (0.1% v/v or 11.8 mM) for four hours, followed by live cellular fluorescence quantification by flow cytometry. For each cell, I obtained a measure of green fluorescence, blue fluorescence, and laser light deflection correlated with size or morphology. To filter out non-cell events, I restricted the analysis to the 60% of events in each sample having size and morphology values within the highest density region of those parameters. Because mean forward and side scattering properties vary as a function of MMS, I performed gating based on these properties in a sample-dependent manner by estimating the cell population as

60% of events within the densest region of a two-dimension plot of forward and side scatter.

Mean cellular DDC-dependent transcription in response to MMS is non-monotonic (Figure 4.2). This response is reproducible in repeat experiments and for different yeast strain backgrounds. The response is not observed for DDC-independent transcription and the response is maintained when the data are corrected for autofluorescence. The response is approximated by a two-term exponential model. With reference to MMS concentrations, the response is an increasing function of MMS concentration up to a peak of transcription activation at 0.0175 %v/v MMS ($\pm$ 8.98 x 10^{-4}, 95% CI). Above concentrations of 0.0175 %v/v, the response is a decreasing function of MMS. At the highest MMS concentration tested, the response is similar to untreated cells (Figure 4.2).

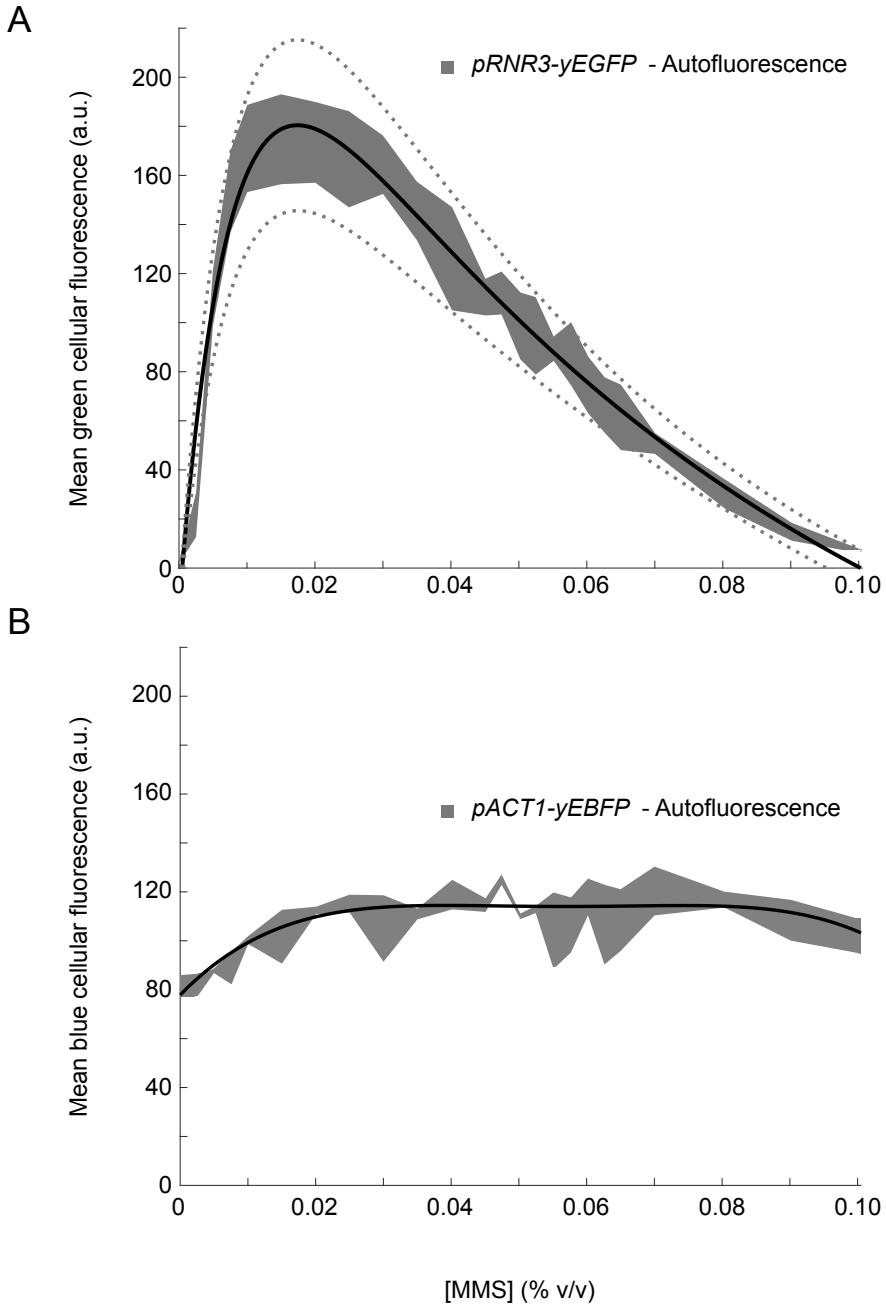

Figure 4.2: High MMS causes non-monotonic *RNR3* transcriptional activation. After exposing cells to MMS for 240 min, we quantified fluorescence by flow cytometry. To estimate mean cellular Gfp levels expressed from the *RNR3* promoter, I linearized data and subtracted mean values of autofluorescent cells from corresponding values of cells carrying the Gfp reporter. The resulting estimate (A) is an increasing function of MMS for concentrations between zero and 0.175 ($\pm$ 8.98 x 10^{-4}, 95% CI), but a decreasing function of MMS when concentrations are above 0.175. Control Bfp levels in the same cells, obtained by the analogous normalization procedures, are shown in (B). Symbols: shaded areas, two standard deviations across three experimental replicate means; black lines, mean fit to replicate data; dotted lines, 95% confidence interval of fit.

The flow cytometric distributions underlying non-monotonic mean cellular *RNR3* transcription level as a function of MMS dose shows that the transition from *RNR3* transcription being an increasing function of MMS to a decreasing function coincides with the appearance of a low Gfp subpopulation. The frequency of cells in this low subpopulation increases as MMS is increased above 0.035% v/v. At the highest MMS concentration tested, the low subpopulation is the only apparent population.

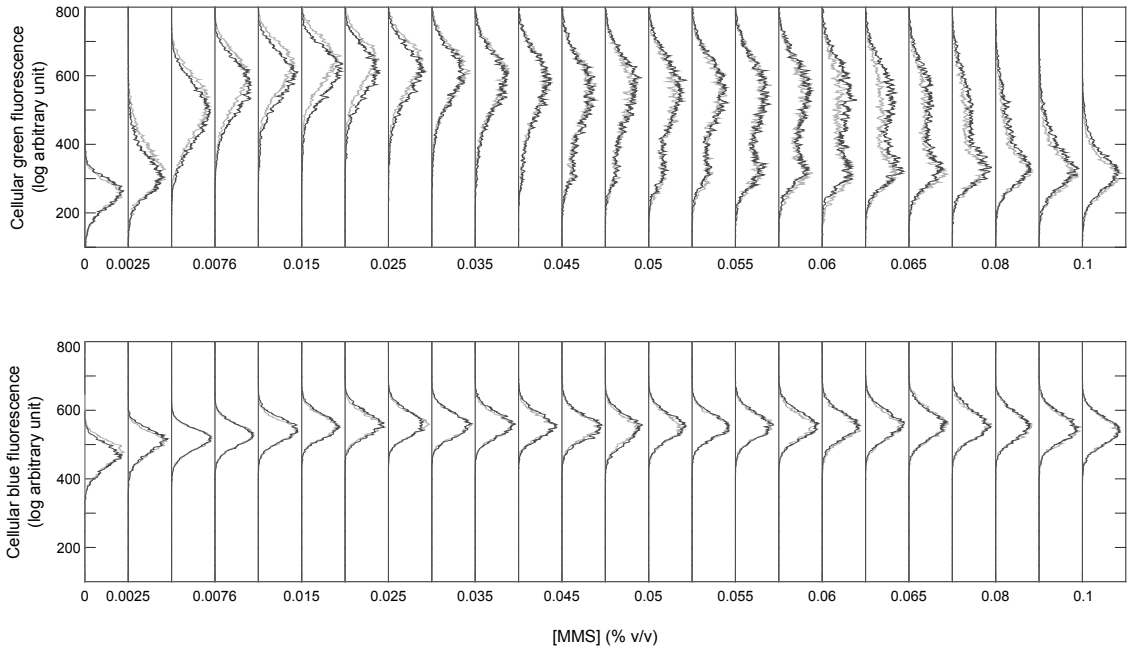

Figure 4.3: High MMS causes emergence of distinct subpopulations. Cells expressing Gfp from the *RNR3* promoter and Bfp from the *ACT1* promoter were exposed to MMS for 240 min. Green and blue fluorescence was quantified at the end of drug treatment by flow cytometry. We find that when histograms of log cell fluorescence are plotted across MMS condition, a low fluorescence subpopulation arises at high MMS concentrations for Gfp emission but not Bfp emission of the same cells, indicating that high MMS causes a change in cellular *RNR3* transcription but not control *ACT1* transcription. The number of cells in the low subpopulation increases as MMS concentrations are increased above 0.035%. Symbols: light and dark gray distinguish replicate experiments obtained with MATa or MATα haploid cells; histogram bin counts are relative to corresponding half maximum of bin count.

4.3.2 Low *RNR3* transcription cannot be explained by cell death

I observe that P_{RNR3} repression during high MMS treatment cannot be explained by cell death or inability to synthesize new protein (Figure 4.4). I find high MMS cells are not permeable to propidium iodide, in contrast with dead cells, and also have *pACT1-BFP* expression levels distinguishable from dead cells (Figure 4.4A). I hypothesized that cells inability to synthesize new protein may be masked by the $P_{ACT1} - BFP$ control reporter due the combination of cells having high constitutive expression prior to MMS treatment and then slowed division rate during MMS treatment, limiting protein dilution by cell division. Therefore, I examined the ability of cells to express *GFP* from a galactose-inducible promoter P_{GAL10} when 2% (w/v) galactose and MMS were added to the cell culture at the same time. I observed that after four hours of galactose and high (0.08% v/v) MMS treatment, $P_{GAL10} - GFP$ expression is inducible to $> 60\%$ of the expression level when treated with galactose and low (0.02% v/v) MMS (Figure 4.4B,C). In contrast, $P_{RNR3} - GFP$ expression is $< 10\%$ as inducible at high (0.08% v/v) versus low (0.02% v/v) MMS (Figure 4.4B). Cell expression capacity is reduced by high MMS, but it does not fully account for the P_{RNR3} repression observed. I note that cells imaged during high MMS exposure in a time-lapse experiment are capable of inducing $P_{RNR3} - GFP$ expression within two hours after MMS removal, confirming that cells with repressed $P_{RNR3} - GFP$ expression are capable of expression (Figure 4.4D).

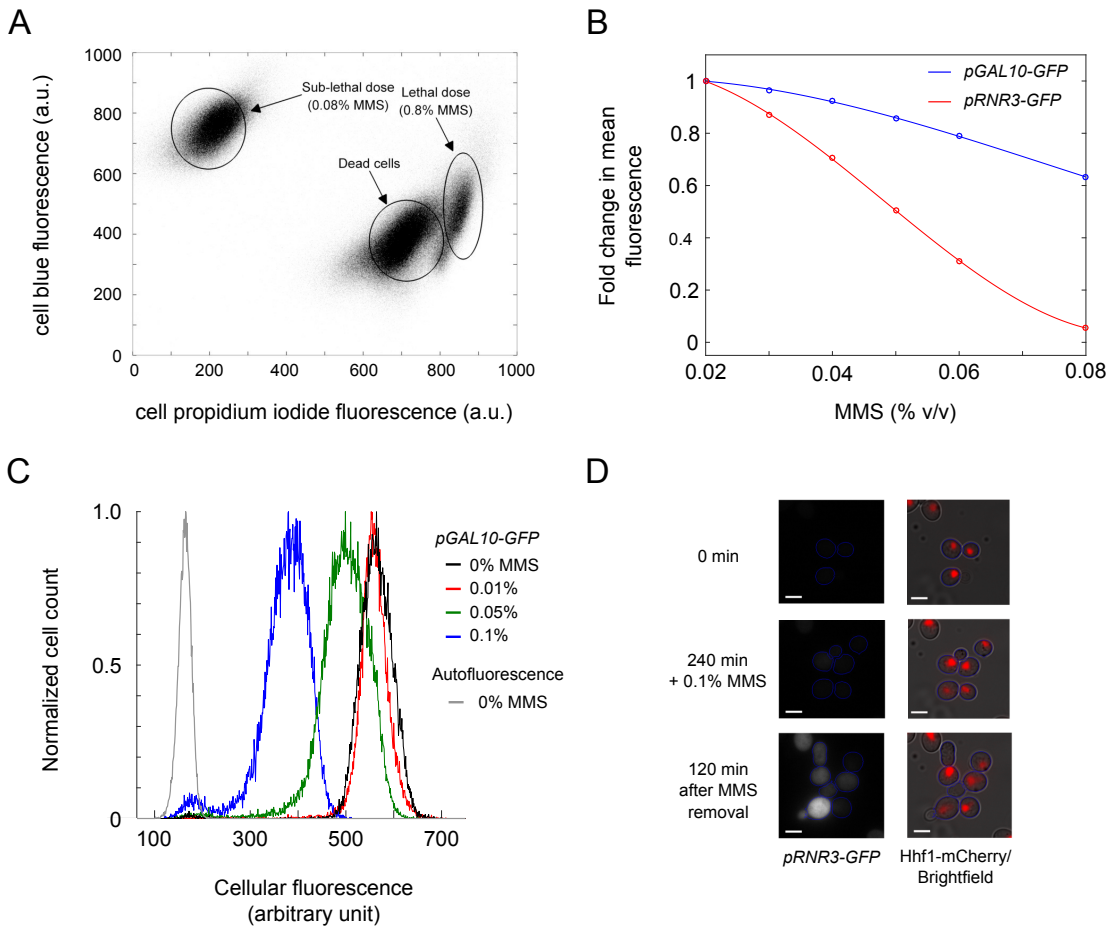

Figure 4.4: *RNR3* repression is not due to low expression capacity or cell death. **A:** Cells treated with high (0.08% v/v) MMS have *pACT1-BFP* expression and propidium iodide permeability distinct from dead cells, killed by heat treatment or 10 fold increase in MMS dose. **B,C:** Alternative promoter driven GFP expression is inducible in cells treated with high (0.08% v/v) MMS. High (0.08% v/v) MMS treated cells have 60% of the capacity of low (0.02%) MMS treated cells for inducing expression from the GAL10 promoter in an experimental design where MMS and 2% galactose are added to cell cultures at the same time (240 min co-treatment). Expression of the same protein from the RNR3 promoter, in contrast, is less than 5% as inducible with high MMS compared to low. Expression data is obtained by flow cytometry and data acquisition, filtering and autofluorescence correction is comparable to that described in Figure 4.2. **D:** Cells treated with high (0.1% v/v) MMS are capable of inducing expression of *pRNR3-GFP* within two hours of MMS removal. Cells were treated with MMS within a microfluidics chamber and imaged every hour for during 4h MMS treatment and for 2h after MMS treatment. Symbols: size bar, 5 μm.

4.3.3 Low *RNR3* transcription cannot be explained by DNA damage adaption or recovery

The alternative hypothesis was P_{RNR3} is actively repressed by Crt1p because the DDC pathway is not active under high MMS conditions. Following DNA damage, I consider three possible cell fates. First, DNA damage repair is not complete and cells are arrested in S or pre-anaphase with an active DDC. Second, the cells have repaired the damaged DNA and de-activated the DDC to resume cell division, a process known as recovery [99]. Third, the cells have not repaired the damage and nonetheless de-activate the DDC to continue division, a process known as adaptation [99]. The DDC would be inactive in the two latter fates, the fates can therefore be distinguished by the continued presence of a DNA damage signal.

To test if a DNA damage signal was still present in high MMS-treated cells despite the absence of DDC transcription, I measured a cellular marker of active DNA repair, the presence of Rad52-Gfp foci. Members of the Rad52 epistasis group are sensitive to the MMS doses we test [111], and Rad52p foci are hypothesized to correspond to active centres of repair of double stranded DNA breaks by homologous recombination [112]. I observed cells treated with high (0.1%) and low (0.02%) MMS for 4h had comparable percentages of cells with foci (Figure 4.5A). This result, together with the very distinct RNR3 expression levels at these two concentrations of MMS (Figure 4.2), suggests that lack of DDC transcription cannot be explained by DNA damage recovery, leaving the possibility of DNA damage adaptation.

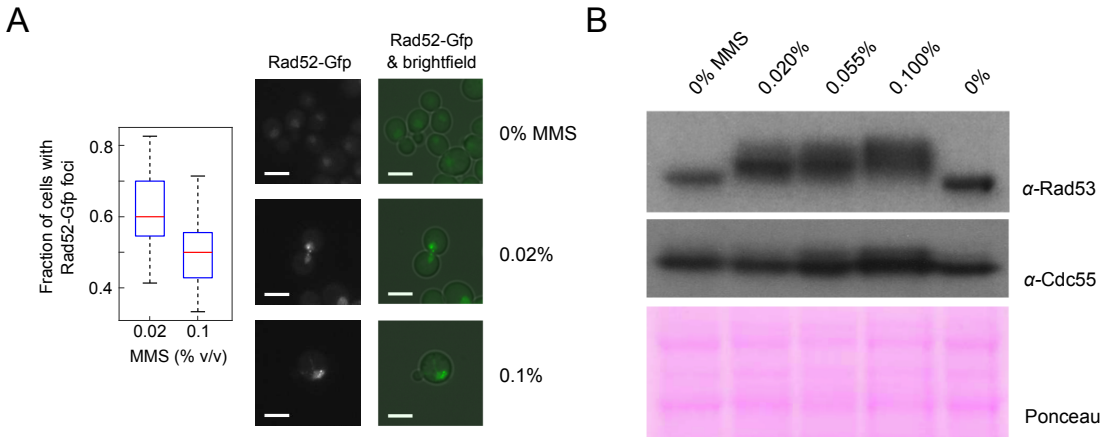

Figure 4.5: *RNR3* repression is not due to DDC adaptation. A: I observe cells treated with either high or low MMS (240 min) to be comparable in their probabilities of possessing one or more Rad52-Gfp foci. Variability among the median (red) represents differences among random samplings of subsets of data, obtained by visual inspection of fluorescence microscopy images. Cells treated with either high or low MMS (240 min) have Rad53 protein with an electrophoretic shift characteristic of DDC activation (standard immunoblotting). Cdc55 is a control for protein abundance and cell number. **Symbols:** size bar in images, 5μm; α, primary antibody.

DNA damage adaptation is characterized by a prolonged ($\sim$ 10h) S or pre-anaphase arrest in response to DNA damage, after which, despite continued presence of the damage signal, cells de-activate the DDC and resume cell divisions. DDC de-activation is characterized by the loss of hyperphosphorylation of DDC effector Rad53p kinase [113, 114], whereas resumed cell division is quantified as the ability of a cell to divide to form a microcolony of more than $\sim$ 20 cells within 24 hours [114].

To test if cells de-activated the DDC during high MMS treatment, electrophoretic mobility was measured of Rad53 protein extracts from cells treated with a range of MMS doses. Decreased Rad53 electrophoretic mobility is characteristic of hyperphosphorylated Rad53 protein [115], dependent on Mec1 and Tel1 [84]. Protein extracted from cells treated with high MMS had a comparable, if not more pronounced, decrease in Rad53 mobility compared to low MMS treated samples. I confirmed that RNR3 transcription of cells in these samples matched the observations in Figure 4.2. I note a previous study also reports high (0.1% v/v) MMS induced shift of Rad53p within 2 h [84], and did show the shift was lost if they pre-treated Rad53 immunoprecipiate with phosphatase [84]. The result suggests that Rad53 is phosphorylated in high MMS treated cells despite low RNR3 transcription, and low RNR3 transcription is not due to DNA damage adaptation. I note additional conditions for adaption were not met, i.e. the presence of a large bud during DNA damage exposure, and the resumed cell division from this arrest. The timescale of my experiments (4 h) also does not agree with the timescale of observed adaptation in previous studies ($>$ 10 h) [114] in absence of additional mutations that accelerate adaptation [113].

4.3.4 Cell cycle phase after MMS exposure is dose dependent

I sought alternative explanations for P_{RNR3} repression and the dependency of induction on cell division phase appears to be often overlooked yet reproducible in numerous studies. Mazumder et al. quantified RNR gene transcripts in single cells to study the variability in DDC activation and found that *RNR3* transcripts were not evenly expressed in a population of low MMS treated cells, but were inducible only in budded (S phase) cells [116]. This observation agrees with two previous studies showing deletion of *SWI4* reduces or prevents induction of RNR3 transcription by genetic [117] or environmental [118] conditions causing DNA damage. *SWI4* encodes a component of the SBF transcription factor required for transcription of G1/S cyclins Cln1p and Cln2p [119]. The *RNR3* promoter sequence also contains overlapping SBF/MBF binding sites.

To formally test if cell division cycle phase progression is altered by MMS dose, I measured DNA content in single cells by propodium iodide staining and flow cytometry. I observe that DNA content distributions vary with MMS concentration between 0 and 0.1% MMS. As MMS is increased above 0.0025% MMS the distribution changes from a bimodal distribution, characteristic of unsynchronized cells, to unimodal with a peak around 2N, anticipated with a pre-anaphase arrest following DNA damage exposure. As MMS concentration is increased further, the peak of the unimodal distribution shifts higher, reaching a maximum at about 0.025% MMS. Concentrations above 0.025% cause the peak to shift back down, with a mode between 1N and 2N at 0.065% MMS, suggesting an S phase arrest. At concentrations above 0.065% MMS, the distribution is increasingly wide, with two visualizable peaks at 0.1% MMS. This bimodal distribution differs from the untreated and unsynchronized population because the untreated distribution has the majority of cells near 2N

whereas the high MMS distribution has the majority of cells slightly above 1N. The overall trend appears to be non-monotonic, wherein increasing MMS dose causes an increase in DNA content to a certain threshold (approximately 0.025% MMS), and concentrations above this threshold cause a decrease in DNA content. The deviation from this overall trend is the presence of two peaks at very high MMS. However, given alternative estimates of cell division phase by Tub1p localization (Figure 4.6B), I hypothesize that high MMS cells are primarily G1 cells, where the low N peak represents unbudded G1, and the higher peak represents two post-mitotic cells with defective or incomplete cytokinesis.

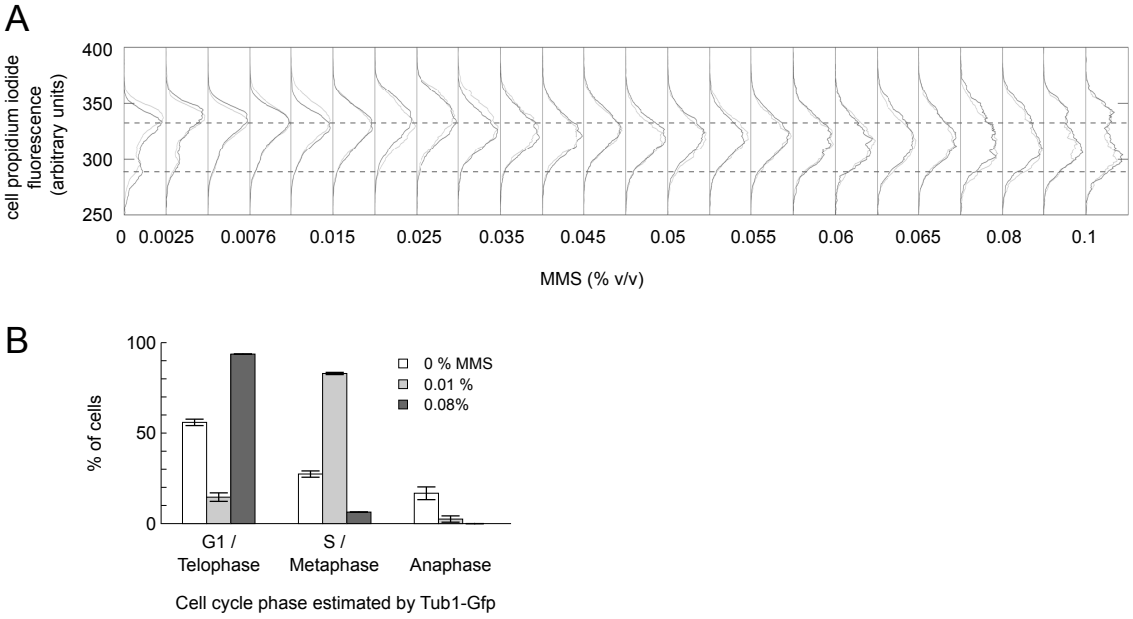

Figure 4.6: High MMS causes changes in cell cycle distribution. I observe genome copy number, as estimated from fluorescence of propidium iodide-stained cells to vary with MMS dose. The distributions suggest high MMS-treated cells do not frequently begin or complete S phase, and low MMS-treated cells do not frequently begin or complete mitosis (A). An alternative indicator of cell division phase (B), microtubules visualized by Tub1-Gfp, indicate that high (0.08% v/v) MMS increases the probability cells will be in G1 or Telophase at the end of MMS exposure (240 min), in contrast to low (0.01% v/v) MMS wherein cells are primarily in S or metaphase at the end of exposure. Symbols: light and dark grey distributions, replicate experimental data with bin counts relative to corresponding half maximum bin count; error bars, standard deviation of means of two replicate experiments estimated by visual inspection of a minimum of 60 cells for each replicate and each condition.

4.3.5 Low *RNR3* transcription coincides with increased propensity of G1 arrest

I hypothesize that cells with low *RNR3* transcription levels are in G1. This hypothesis would predict that in a population of cells treated with intermediate MMS having a bimodal distribution, the low subpopulation represents G1 cells whereas the high subpopulation represents cells in S phase or pre-anaphase. I tested this prediction by examining the bud indices of cells sorted from low versus high MMS subpopulations from a population of cells treated with intermediate MMS (0.055% v/v). In agreement with the hypothesis, cells from the low *RNR3* mode have a median of 10% of cells with a bud, in contrast to cells from the high mode with a median of 100% of cells with a bud (Figure 4.7).

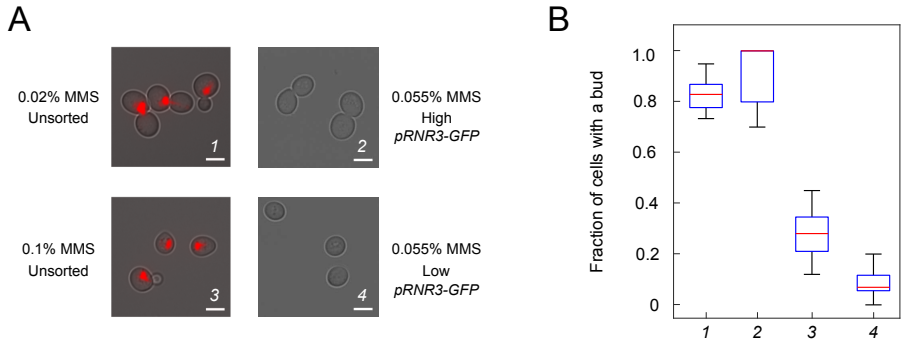

Figure 4.7: Cell budding indices are correlated with *RNR3* transcription. Cells treated with low MMS (0.02% v/v, 240 min) have high probability of a bud (1), whereas cells treated with high MMS (0.1% v/v, 240 min) low probability of a bud (3), suggesting there is a correlation between transcriptional repression and cell cycle phase. If cell cycle phase is a predictor of *RNR3* transcription, there should be significant cell cycle differences among $P_{RNR3} - GFP$ expressing cells sorted by flow cytometry based on Gfp from a bimodal Gfp population exposed to intermediate MMS (0.055%, 240 min). In agreement with this expectation, cells sorted from the high expression mode (2) have bud probability similar to low MMS-treated cells, whereas cells sorted from the repressed expression mode (4) have bud probability comparable to high MMS-treated cells. Symbols: box plot medians and variance, estimated by random sampling of manually annotated bright field microscopy images (minimum of 49 cells annotated per estimated median); red fluorescence, Hhf1-mCherry to label cell nuclei; size bar in images, 5 μm

Because *RNR3* transcription decreases as a function of MMS concentrations above $\sim 0.02\%$ (v/v) (Figure 4.2), the hypothesis also predicts that the number of cells in G1 will increase as a function of these MMS concentrations. To test the prediction, I quantified the number of cells with nuclear localization of Whi5-Gfp following MMS exposure to doses of 0.02 to 0.1% v/v. Nuclear localization of Whi5-Gfp is a marker of a cell in G1. Whi5p has been shown to accumulate in the nuclei of post-mitotic cells, where it represses SBF driven transcription of G1-S cyclins encoded by *CLN1* and *CLN2*. Cln1p, Cln2p or Cln3p in complex with Cdc28p can phosphorylate Whi5p and cause its nuclear export [120]. I observe that the percentage of cells nuclear localized Whi5-Gfp increases reproducibly as a function of increasing MMS. At the highest MMS concentration, over 95% of cells have nuclear-localized Whi5-Gfp. As predicted by the hypothesis, the number of cells in G1 is positively correlated with MMS dose at concentrations above 0.02% MMS.

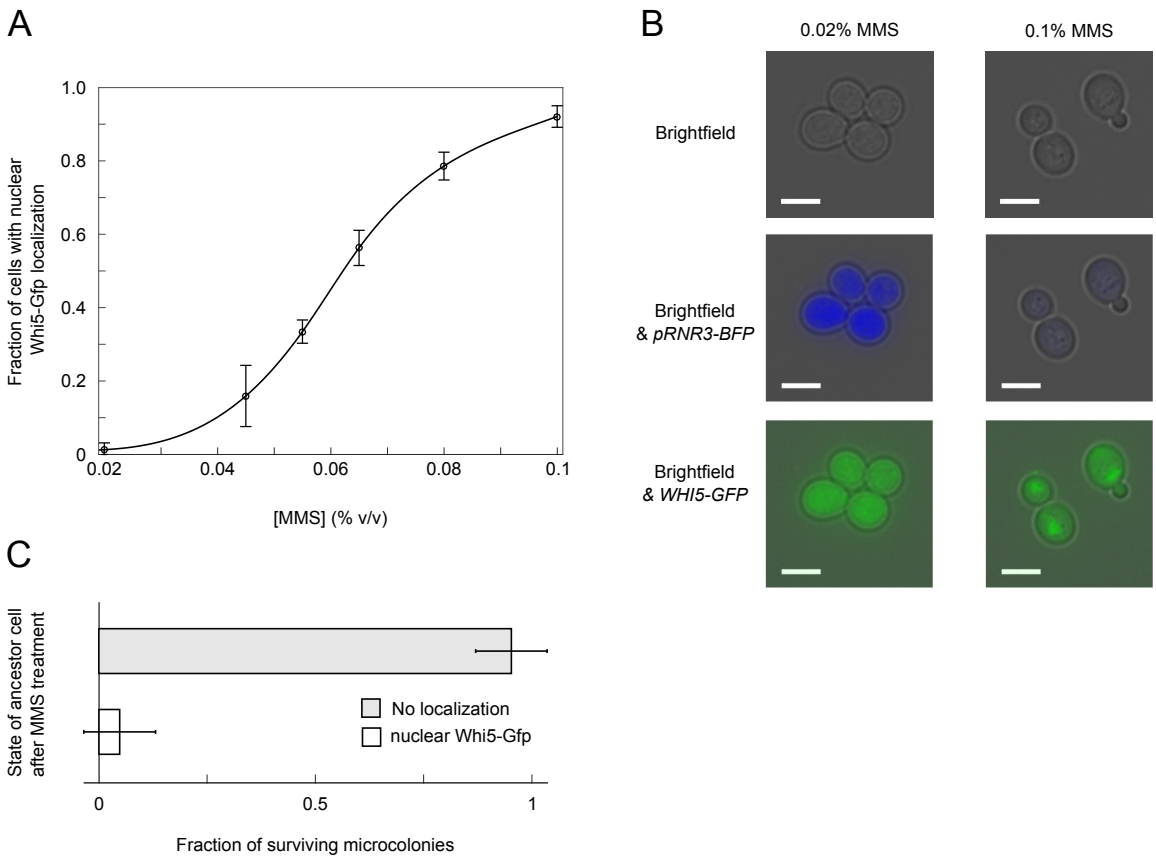

Figure 4.8: High MMS triggers nuclear Whi5 localization. I further characterized cell cycle changes following MMS treatment by quantifying the presence or absence of a nuclear localized fluorescence signal in cells expressing a Whi5-Gfp fusion protein from the *WHI5* locus. Nuclear localization of Whi5-Gfp is typically observed in cells in G1. As a function of increased MMS dose (240 min treatment), I find an increasing fraction of the cell population with Whi5-Gfp localization (**A**). This pattern agrees with the expectation high MMS prevents cells from beginning or completing S phase based on results of Figure 4.6. The corresponding fluorescence microscopy images of cells used for quantification (**B**) indicate that cells with high expression of *pRNR3-BFP* exposed to low MMS (0.02% v/v) lack localization, whereas cells with low expression exposed to high MMS (0.1% v/v) do have localization. **C.** After cells were exposed to intermediate MMS (0.055% v/v, 4h) I examined whether presence or absence of localization correlated with cellular fitness after MMS removal by loading cells into a microfluidics device allowing time-lapse microscopic imaging for 20h in absence of drug. Whereas 32 ±3% had localized Whi5-Gfp at the end of MMS treatment as expected from A, 95±8% of cells that were able to reproduce to form micro-colonies (see text) did not have localized Whi5-Gfp (cells counted: 558, total % cells that form microcolonies: 2.8±0.7%. Symbols: error bars in A, ± standard deviation of the mean of two replicate experiments obtained by visual inspection of a mean of 216 and minimum of 69 cells for each replicate and each MMS condition; error bars in C, ± standard deviation of three replicate experiments; size bar in images, 5 μm.

To analyze if cell division phase influences cell survival, I performed time-lapse microscopy experiments wherein cells having Whi5-Gfp were exposed to intermediate MMS (0.055% v/v) in a microfluidic growth chamber (240 min). These conditions result in a mixed population where 32 $\pm$ 3% of cells have nuclear localized Whi5-Gfp. By continuing to monitor cells for 20 h after MMS removal, I found that nearly all (95 $\pm$ 8%) surviving micro colonies ($>$ 20 cells) did not have localized Whi5-Gfp at the time of MMS removal (Figure 4.8). This suggests that cells without localized Whi5-Gfp have a survival advantage.

The above-described results suggest that MMS concentration determines the proportion of cells in G1 when cells are initially unsynchronized. To investigate how cell division phase of cells prior to MMS treatment has an effect on these proportions, I treated cells with high MMS (0.1% v/v MMS) at multiple time points after cells had been released from G1 arrest by α-factor treatment. In this experiment, I counted the fractions of cells with localized Whi5p-Gfp, both at the time of MMS addition and after MMS exposure (180 min). I observe the starting fraction of cells with localized Whi5-Gfp ($<$ 10% to $>$ 90%) has no detectable influence on the proportion of cells after MMS exposure ($\sim$ 70-80%) (Figure 4.9).

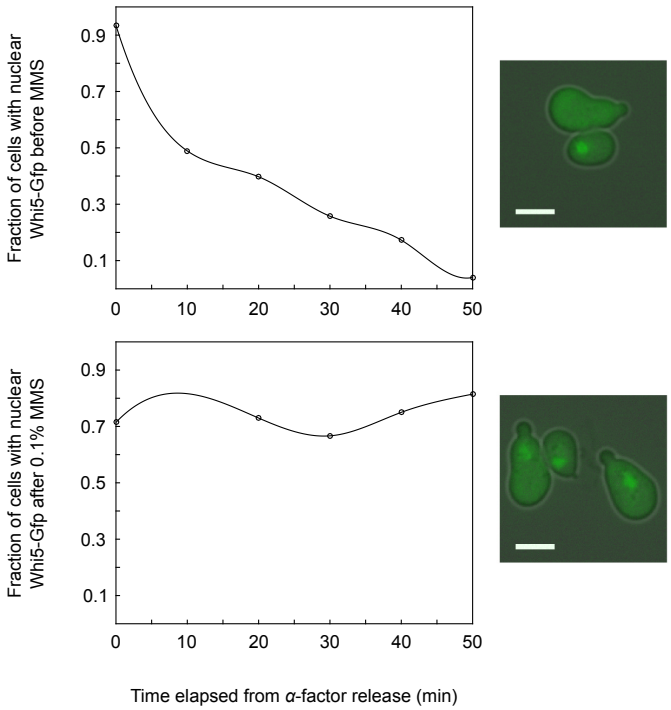

Figure 4.9: High MMS triggers nuclear Whi5 localization in α factor synchronized cells. Cells expressing a Whi5-Gfp fusion protein from the *WHI5* locus were arrested in G1 by α factor treatment, and released. At regular time intervals following release, high MMS was added to aliquoted cultures (final 0.1% v/v MMS), and the number of cells with localized Whi5-Gfp was counted both at the time of adding MMS (A) and after MMS exposure (180 min) (B). I observe the fraction of cells with localized Whi5-Gfp after treatment is not correlated with the fraction of cells before treatment. This result suggests that cell cycle phase before treatment does not influence the probability of high MMS triggering Whi5-Gfp localization. Symbols: lines, spline fits to raw data; open circle, raw data reflecting fraction of cells with localization determined by manual inspection of a minimum of 37 cells per data point with an average of 142 cells per data point.

4.3.6 High MMS forces cells out of canonical G2/M arrest induced by DNA damage

An important additional observation is the significant fraction of high MMS-treated cells possessing both nuclear localized Whi5-Gfp and a bud. This combination of cellular markers of G1 and S (e.g. Figure 4.8B bottom right, Figure 4.9 bottom right), is never observed in parallel characterizations of untreated cells. In dividing cells, loss nuclear Whi5-Gfp localization immediately precedes bud initiation [120]. Presence of a bud and nuclear Whi5p could reflect a number of possible deviations from a normal cell division, for example: cell division progression could be normal with non-functional repression of Whi5p that remains nuclear, or cells could progress normally, losing Whi5p localization prior to budding and then re-gain nuclear Whi5p. The first possibility is less likely given the budding indices and DNA content changes I observe at higher MMS doses. Under the assumption that Whi5-Gfp localization accurately reflects the cyclin-dependent kinase activity of a G1 cell, the second possibility would imply that high MMS concentrations can cause cells to re-enter a G1-like phase.

I performed experiments to determine whether the second option was possible. Specifically, I started with a culture of S phase arrested cells which clearly lacked Whi5 nuclear localization, and then added high MMS to the culture. This experiment allows one to test if cells arrested without localization can re-gain localization. I pre-treated cells with low MMS (0.02% MMS) or hydroxyurea (100mM) for three hours(Figure 4.10A,B), after which both samples had a median of less than 10% of cells with nuclear Whi5 localization (Figure 4.10C). I then split each culture to maintain the arrest or to add high (0.1% v/v) MMS. After three hours of high MMS or control condition, I quantified the percentages of cells with nuclear Whi5-Gfp localization.

To the cultures I had added high MMS, the median percentage of cells with Whi5-Gfp nuclear localization increased to approximately 70% if pre-treated with HU, or approximately 50% if pre-treated with low MMS (Figure 4.10C). In contrast, the controls maintained median percentages of approximately 10% (Figure 4.10C). These results suggest that high MMS can cause cells to gain nuclear Whi5-Gfp localization. Notably, the cells having the localized Whi5-Gfp have either completed mitosis, marked by the presence of two nuclei in adjoint cells, or have gained localization of Whi5-Gfp in the nucleus of the pre-mitotic mother cell having a bud. These experiments suggest that high MMS can force cells out of canonical G2/M arrest induced by DNA damage.

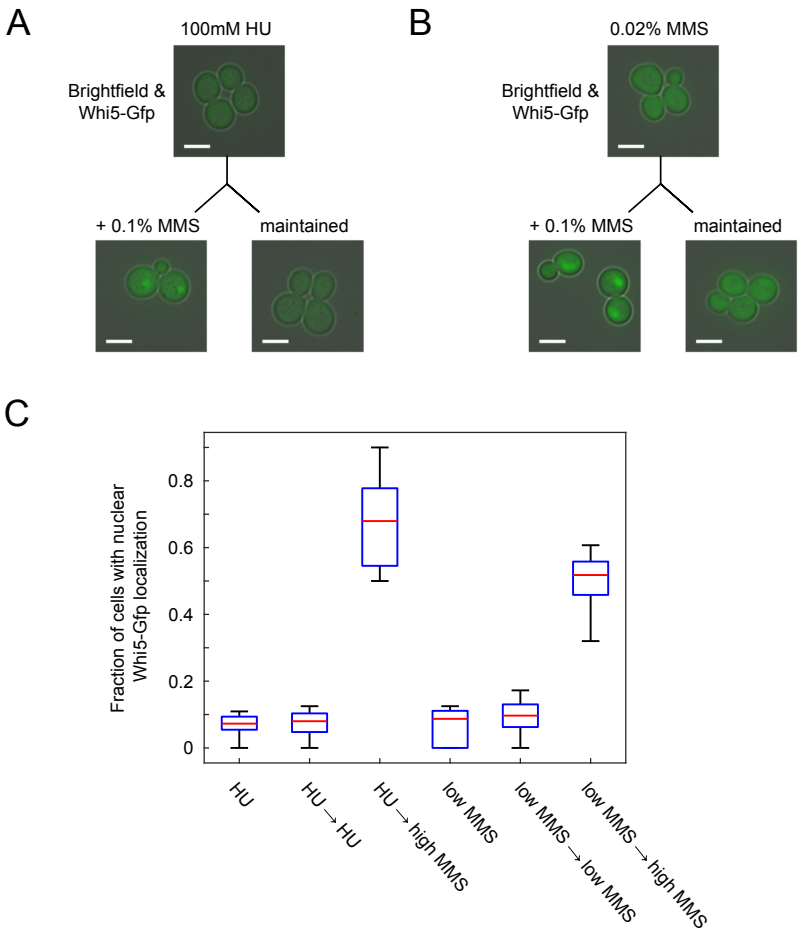

Figure 4.10: High MMS triggers Whi5 nuclear localization in S-phase arrested cells. Cells expressing a Whi5-Gfp fusion protein from the *WHI5* locus were arrested with 100mM hydroxyurea (A) or 0.02% v/v MMS (B) before splitting the cultures into two conditions: one where MMS was added to increase the final concentration by 0.1% v/v (bottom left of A,B), and second where arrest was maintained (bottom right of A,B). I quantified the number of cells with nuclear localized Whi5-Gfp before splitting the culture, and after 180 min in each condition of the split cultures. The data (C) show that high MMS causes cells to have nuclear Whi5 localization even when they were arrested prior to exposure. Cells with localization after high MMS treatment are unbudded, as expected for G1 cells, or have a small or large bud (A,B). Symbols: size bar in images, 5 μm; boxplot medians and variance, obtained by random sampling of manually annotated bright field and green fluorescence microscopy images (minimum of 27 cells and mean of 94 cells counted per condition).

4.4 Discussion

High population variability in a response to DNA damage observed at intermediate MMS dose can be explained by a dependency of *RNR3* promoter activity on cell division phase. I find that cell average *RNR3* transcription level is inversely correlated with the probability of cells having nuclear Whi5p localization, a marker of G1 cells. I find when cellular *RNR3* transcription is bimodally distributed in a population, cells in the low expression subpopulation have budding indices characteristic of G1 cells, and cells in the high subpopulation have indices reflecting S- or meta-phase.

Although low *RNR3* transcription has been linked to G1 in previous studies, the observations of the present chapter support inconsistent causes. Specifically, in an unsynchronized population of cells with variable *RNR3* transcription under low MMS conditions, Mazumder et al. [116] found cells with low levels of transcript were in G1 and also lacked two characteristics of DDC pathway activity: phoshorylated H2A and cytoplasmic Rnr2p [121]. In contrast, at the highest MMS concentration I test (0.1% v/v MMS), cells with low *RNR3* transcription do have characteristics of an active DDC pathway, namely electrophoretic shift in Rad53p characteristic of hyperphosphorylation. The discrepancy cannot be explained by use of lysate assay versus the single cell assay described by Mazumder et al. [116] since about 95% of high MMS treated cells in the present study have a Whi5-Gfp G1 marker. Such a high percentage of the population should produce a visible change in the electrophoretic shift of protein from lysed cells if *RNR3* repression was indeed due to reduced activity of Mec1p.

These inconsistencies are challenging to resolve owing to the numerous alternative pathways by which the DDC may be activated in a cell-cycle dependent manner [122], and the numerous phases of the cell division cycle wherein the DDC may cause arrest (or slowed progression). For example, while Mazumder et al. [116] suggest low *RNR3*

transcription is owing to the inability of cells to activate the DDC in G1, alternative studies suggest that delay in transition from G1 to S phase in response to high MMS treatment [123], UV- [124] or γ-irradiation [124, 125] requires DDC activity.

The inconsistencies therefore may be due to differences in cellular responses to low versus high MMS. In the literature, low MMS (0.015-0.033% v/v) is reported to cause S phase delay and activation of the G2/M checkpoint regardless of whether MMS is added to cultures of unsynchronized cells [101] or following release from G1 arrest by α-factor treatment [126]. In the study wherein cells were initially unsynchronized [101], continued low MMS (0.015% v/v) exposure causes cells to become synchronized G2/M within 3 h, and maintain this arrest for at least 4 h. G2/M arrest was measured by cells predominantly ($> 90\%$) having a large bud and a duplicated genome (2N DNA content). These studies did not examine the effects of higher MMS concentrations. However, in another study by Sidorova et al. [123], α-factor arrested cells co-treated with high (0.1-0.2% v/v) MMS, and then released after 30 min MMS exposure, are reported to have a 15-30 min delay to begin S phase relative to untreated cells. The delay was reported to be reduced but not completely absent in cells having Rad53 mutated (Rad53-11). Delay in transition was measured by cells acquiring the following phenotypes later than untreated cells: α-factor resistance; G1/S cyclin transcription (CLN1); bud initiation; and increased DNA content. Additionally, delay was accompanied by Rad53p-dependent Swi6p phosphorylation, and could be partly alleviated by a mutation of Swi4p that allows SBF transcription in absence of Swi6p [123]. The data of Sidorova et al. [123] therefore support the possibility that *RNR3* transcription is repressed in G1 cells despite having an active DDC, as I observe.

Sidorova et al. [123] did not report the effect of longer (> 30 min) exposure to high MMS, nor the influence of the initial cell division phase on high MMS-induced G1 delay. Similarly, [126] and [101] did not report the effect of higher MMS concentrations.

The characterization in the present chapter of dose-dependency of cell division phase on MMS with prolonged exposure (240 min) therefore links these previous studies by showing that drug dose can alter the propensity of unsynchronized cells to be in a given phase after treatment. To test if this propensity was indeed determined by drug concentration independent of the cell division phase preceding drug exposure, I added high (0.1% v/v) MMS to cells at various time points after release from α-factor. In this experiment, I did not detect a dependency on starting cell division phase.

There are a number of indications that suggest the pattern of cell division phase changes I observe as a function of MMS dose is different from previous studies. This was first suggested by the observation that cells treated with high MMS can possess both G1 and non-G1 markers. Although nearly all cells treated with high MMS have Whi5-Gfp nuclear localization, the DNA content distribution of these cells was bimodal (Figure 4.6) and a number of high MMS-treated cells with a Whi5-Gfp nuclear localization also possess a bud. Additionally, high MMS-treated cells had Rad52-Gfp foci, hypothesized to require S phase cyclin-Cdk activity [127]. These data suggest that cells only have a subset of indicators of a G1 state, and I therefore refer to the state as G1-like. I note that the discrepancy between bud index and alternative markers of cell division progression has been reported before. For example, bud index has over-estimated progression compared to DNA content under conditions of where G1 synchronized cells are UV irradiated and released [81] or in S phase cells with artificially induced expression of S phase inhibitors [128]. Similarly, DNA content distributions acquired by flow cytometry have reported biases towards higher N [129] relative to alternative markers of cell division phase.

To further explore the discrepancy among different cell cycle phase markers, I examined the response to high MMS of synchronized cells arrested in S phase or pre-anaphase and predominantly lacking nuclear localization of Whi5-Gfp. I found that within three hours large percentages of cells entered into a G1-like state, characterized by presence

of nuclear Whi5-Gfp localization. Some cells appear to have completed mitosis prior to gaining this G1 marker, whereas others regain the marker in absence of mitosis. Because high MMS treatment is not accompanied by a loss of Rad53p electrophorectic shift, these experiments suggest that the G2/M checkpoint characteristic of cells exposed to low MMS or HU exposure may be overridden by a mechanism distinct from DNA damage adaptation [114].

The causes and consequences of apparent entry into a G1-like state without the completion of mitosis are beyond the scope of this study. I note however, that such transitions are possible in a number of organisms. G1-to-S-to-G1 cell cycling, i.e. DNA synthesis in the absence of mitosis, is termed endoreduplication or endocycling and is observed naturally in various multi-cellular eukaryotic systems, often limited to a cell type or developmental stage (reviewed in [130]). Endocycling has also been studied in unicellular eukaryotes, but primary examples required mutation and/or exposure to severe stress. *S. pombe* deleted for the M cyclin Cdc13 undergoes endocycling, as do cells with Cdc13 mutations when treated with severe stress [131]. *S. cerevisiae* has multiple M cyclins, but has also been shown to endocycle. In these experiments, cells were pretreated with nocodazole, causing pre-anaphase arrest, followed by transient, galactose induced expression of Sic1, an inhibitor of Clb-Cdk (S and M cyclin-Cdk) complexes [132]. Yeast endocycling is thus achieved by inhibiting mitosis and then allowing a transient reduction of S- and M-Cdk activity. These synthetic experiments are analogous to the inhibition of M-Cdk activity combined with oscillations in S-Cdk activity that coincide with endocyling in multicellular organisms. I note that in these examples cells are hypothesized to complete S phase before entering G1, whereas in the experiments of the present study cell bud sizes suggest cell re-entry to G1 from early S phase, or potentially re-entry following budding in absence of DNA synthesis altogether. A second note is that the use of severe stress, namely nitrogen deprivation and/or heat shock, to induce endocycles in *S. pombe* mutants raises the question

as to whether high MMS may cause cells to activate a comparable stress response that parallels DNA damage, for example, by S_N2-type reactions between MMS and non-DNA cellular nucleophiles [133].

4.5 Methodology

4.5.1 DNA construct assembly

The DNA construct shown in Figure 4.11 was assembled by a modification of the 3A Assembly method described in [134]. The method involves sequential restriction digests and ligations in *E. coli*, as well as homologous recombination in *S. cerevisiae* by lithium acetate method for yeast transformation adapted from [91]. The final assembly was of three DNA parts having 5' and 3' ends with restriction sites following the 3A assembly standard. These parts are $P_{AgTEF} - NATMX - T_{AgTEF}$ (as described in [135]), $P_{RNR3} - GFP - T_{AgTEF}$ and $P_{ACT1} - BFP - T_{ADH1}$. The two last parts were first each assembled by two iterations of modified 3A assembly from the corresponding three subparts of transcriptional promoter, open reading frame and transcriptional terminator. Descriptions of each DNA part are provided below (§4.5.1). The complete DNA construct in plasmid pSB1T3 ([134]) was PCR amplified and used to transform BY4741 to replace genomic DNA of the *ADE4* open reading frame (60bp primers). Following transformation, PCR validation and cell fluorescence validation, cells carrying the construct were transformed with a DNA fragment encoding $URA3 - T_{PGK1}$ (also assembled by modified 3A assembly), having 40 bp flanking DNA regions homologous to P_{RNR3} and P_{AgTEF}, allowing replacement of $NATMX - T_{AgTEF}$ by $URA3 - T_{PGK1}$. The resulting DNA construct as integrated in *S. cerevisiae* is depicted in Figure 4.11 and was sequenced from DNA obtained by PCR amplification of the construct from genomic DNA extracted after transformations and validations. These sequencing results revealed a point mutation in *GFP* and *URA3*, and a 5bp deletion in T_{PGK1}. To avoid complications with deletion of a gene encoding an adenine biosynthesis enzyme encoded by *ADE4*, the complete construct was amplified and re-integrated at the neutral *met15*Δ*0* locus in MATa (BY4741) or MATα (Y8205) haploid *S. cerevisiae*. Nonetheless, I found

that locus did not change Gfp expression when measured as a function of MMS concentration. Yeast strains BY4741 and Y8205 are kind gifts from Drs. Kristin Baetz and Charlie Boone, respectively.

Figure 4.11: A DNA construct containing $P_{RNR3} - GFP$ and $P_{ACT1} - BFP$ reporters allows quantification in single cells of DDC-dependent and -independent transcription. DNA parts used to assemble this construct are described in §4.5.1. Symbols: white, open reading frame; dark gray, promoter; light gray, terminator; arrow direction of promoter, direction of promoted transcription.

Sources of DNA

All DNA parts were PCR amplified from yeast genomic DNA extracted from BY4741, with the following exceptions. Blue fluorescent protein (BFP) is mTagBFP (Evrogen Joint Stock Company) with amino acid sequence described in [136]. mTagBFP is a monomeric blue fluorescent protein obtained by mutagenisis from from TagRFP. The yEBFP DNA sequence used in the present study has been optimized for yeast codon usage (sequence provided below). Green fluorescent protein is yeGFP3, optimized for yeast codon usage as described in [71]. The *GFP* DNA sequence in the present study deviates from yeGFP3 as the 3' end has a restriction site inserted upstream from the yeGFP3 stop codon. The DNA sequence for mCherry is as described by [137] and was a kind gift of Dr. Michael Knop. The $P_{AgTEF} - NATMX - T_{AgTEF}$ DNA sequence is as described in [135] and was a kind gift of Dr. Kristin Baetz. The *URA3* DNA sequence corresponds to the open reading frame of *YEL021W* of S288C.

For DNA parts PCR amplified from yeast genomic DNA extracted from BY4741, the DNA sequences used are as follows, wherein +1 represents the first base of the start codon of a gene. P_{RNR3} is DNA encoding the promoter (-616 to -1 bp) of $RNR3$. T_{PGK1} is DNA encoding the terminator (+1249 to +1442 bp) of $PGK1$. P_{ACT1} is DNA encoding the promoter (-479 to -1 bp) of $ACT1$. T_{ADH1} is DNA encoding the terminator (+922 to +1213 bp) of $ADH1$.

The DNA sequence for mTagBFP from [136], optimized for yeast codon usage, is as follows:

ATGTCCGAATTGATCAAGGAAAACATGCACATGAAATTGTA
TATGGAAGGTACTGTCGACAACCACCACTTCAAATGCACCT
CCGAAGGTGAAGGTAAACCTTATGAAGGTACACAAACCATG
AGAATCAAAGTCGTCGAAGGTGGTCCATTGCCATTTGCTTT

CGACATTTTGGCCACATCTTTCTTGTATGGTTCCAAAACTT
TCATCAATCACACCCAAGGTATTCCAGACTTCTTCAAACAA
TCTTTCCCTGAAGGTTTCACTTGGGAAAGAGTCACCACCTA
TGAAGATGGTGGTGTCTTGACTGCTACTCAAGACACATCCT
TACAAGACGGTTGCTTGATCTATAACGTCAAGATTAGAGGT
GTCAACTTCACATCAAACGGTCCTGTCATGCAAAAAAAGAC
ATTGGGTTGGGAAGCTTTCACCGAAACTTTGTATCCTGCCG
ACGGTGGTTTAGAAGGTAGAAACGACATGGCCTTAAAATTG
GTCGGTGGTAGTCACTTGATTGCCAACATCAAAACAACCTA
TAGATCCAAAAAACCTGCCAAAAACTTGAAAATGCCTGGTG
TCTATTATGTCGACTATAGATTGGAAAGAATTAAGGAAGCC
AACAACGAAACTTATGTCGAACAACACGAAGTTGCTGTCGC
CAGATATTGTGACTTGCCTTCAAAATTGGGTCACAAATTGA
ACTAA.

Strains BY4741 WHI5::WHI5-GFP and BY4741 TUB1::TUB1-GFP are from a yeast collection of ORF-GFP tagged strains as described in [38] and are kind gifts of Dr. Kristin Baetz.

Table 4.1: Yeast strains

Genotype
BY4741 met15Δ0::pACT1-BFP/pTEF1-URA3/pRNR3-GFP
BY4741 met15Δ0::pACT1-BFP/pTEF1-URA3/pRNR3-GFP HHF1::HHF1-mCherry/pTEF1-KANMX
Y8205 met15Δ0
Y8205 met15Δ0::pACT1-BFP/pTEF1-URA3/pRNR3-GFP
BY4741 WHI5::WHI5-GFP/pTEF1-KANMX met15Δ0::pRNR3-BFP/pTEF1-URA3
BY4741 ade2::LEU2MX
BY4741 ade2::pGAL10-GFP/LEU2MX

4.5.2 Cell culture and drug exposure

Unless otherwise stated, all MMS exposure experiments were performed as follows. A colony from an agar plate with selective growth medium was used to inoculate 5 ml of synthetic complete (SC) growth medium. After 16 h growth (30°C, 250 rpm shaking), I pelleted (900 g, 2 min) and washed cells with fresh SC and estimated cell density by 600 nm light absorbance of a 1/20 diluted culture in a standard cuvette (Ultrospec 2100 pro, Biochrom, Ltd.). I diluted all samples to the equivalent of 0.125 absorbance units and allowed continued growth with fresh growth medium for 2.5 h (30°C, 250 rpm shaking). After 2.5 h growth, I reduced cell density by half by adding growth medium containing the 2x of indicated concentration of MMS (Sigma, cat. no. 129925). Unless otherwise indicated, I generated MMS stocks by diluting MMS in SC, less than five minutes before we added MMS to cells. For hydroxyurea (HU) experiments, I generated HU stocks by adding HU powder (Sigma, cat. no. H8627) to SC less than five minutes before I added HU to cells. For flow cytometric experiments all cultures were in 15 ml glass or plastic tubes with the exception of four hour MMS exposure, prior to which I transferred cultures to a polypropylene 96 well 1ml assay block (Corning, cat. no. 3958). For all other experiments, I cultured cells in 15 ml glass or polypropylene tubes.

Synthetic complete growth medium

Growth medium containing double distilled H_2O, 6.7 g/l yeast nitrogen base without amino acids (Wisent Inc., cat. no. 800-152) and 2.0 g/l amino acid mix was autoclaved prior to adding filter sterilized glucose solution in double distilled H_2O (final concentration 2% w/v) and filter sterilized adenine hemisulfate salt solution in double distilled H_2O (final concentration 0.042 g/l). The corresponding final concentrations of amino acids and nucleobases are: 20 mg/l of adenine hemisulfate

salt, 180 mg/l of L-Leucine, and 90 mg/l of each of L-arginine, L-aspartic acid, L-histidine monohydrochloride monohydrate, L-isoleucine, L-lysine, L-methionine, L-phenylalanine, L-threonine, L-tryptophan, L-tyrosine, uracil, L-valine, DL-serine, glycine sodium salt hydrate, L-glutamic acid monosodium salt hydrate, L-alanine, L-glutamine, L-asparagine monohydrate, L-proline and L-cysteine hydrochloride monohydrate. Excepting yeast nitrogen base and double distilled H_2O, media ingredients are from Sigma-Aldrich Co. LLC. having $\geqslant 98\%$ purity.

4.5.3 Flow cytometric measurements of gene expression

I acquired flow cytometric measurements using a Cyan ADP 9 Analyzer (Beckman Coulter), equipped with a plate sampler having an automated fluidic system to generate air-gap-separated well samples for delivery to the flow cytometer [92] (HyperCyt® System, IntelliCyt Corp). I used HyperView® Analysis Software to obtain an individual CyAn Summit FCS3.0 file for individual well data [92]. I analyzed well data using custom MATLAB functions developed by Mads Kærn, as described in §3.5.7. To obtain cell population gates, I analyzed event data to obtain a defined percentage of events for each sample that maximized event frequency per area in a 2D plot of Log pulse height forward scatter versus side scatter (488 nm). For data shown in Figure 4.2 I applied a 60% gate. For data shown in Figure 4.4B,C I applied a 50% gate. Within each gated population, I estimated cell blue fluorescence protein expression as the Log pulse height of 425-475 nm filtered emission with 405 nm excitation, and cell green fluorescence protein expression as the Log pulse height of 510-550 nm filtered emission with 488 nm excitation.

4.5.4 Yeast cell protein extraction and immunoblotting

At the end of 4 h MMS treatment or control condition, I measured cell culture absorbance of 600 nm light (Ultrospec 2100 pro, Biochrom, Ltd.) to estimate cell density and diluted cultures to obtain equal numbers of cells per sample, equivalent to six absorbance units. Thereafter, cells were pelleted, washed with ice-chilled sterile H20, transferred to screw-cap tubes, pelleted to remove supernatant, and then flash-frozen in liquid nitrogen and stored at -80C. The following steps of the protocol were conducted by Christine Nwosu and Michael Downey. They lysed thawed cells and extracted protein using a trichloroacetic acid (TCA) protein precipitation reaction as described in [138]. Briefly, they resuspended cells in 300 μl 20% TCA and 100 μl glass beads and lysed cells in suspension using a cell disrupter (Bio Spec Products Inc.). They transferred supernatant to a fresh tube and added a second supernatant obtained by washing the beads with 300 μl 5% TCA. They centrifuged the combined supernatant at 17000 g for 4 min, removed resulting supernatant and resuspended the protein precipitate in 3xSDS-PAGE loading budder containing 1M DTT and 1M Tris-HCl. They boiled the suspension for 5 min, and centrifuged to clarify the sample (17000 g, 4 min). They loaded 10-20 μl of each sample in a lane on an 8% SDS-PAGE gel with 37.5:1 acrylamide:bisacrylamide (BioRad). They transferred the gels to a PVDF membrane and soaked the membrane for 1 h in blocking solution (tris-buffered saline containing 5% milk and 0.1% Tween-20). For Rad53p immunoblotting, they soaked the membrane overnight at 4°C in blocking solution containing 1:2000 goat polyclonal IgG epitope mapped to the C-terminus of *S. cerevisiae* Rad53p (Santa Cruz Biotechnology, cat. no. sc-6749), and then for 1 h in blocking solution containing 1:10000 donkey anti-goat IgG conjugated to horseradish peroxidase (HRP) (Santa Cruz Biotechnology, cat. no. sc-2020). For Cdc55p immunoblotting, they soaked the membrane overnight at 4°C in blocking solution containing 1:10000 rabbit anti-Cdc55 (gift from Adam Rudner), and then for

1 h in blocking solution containing 1:10000 goat anti-rabbit IgG conjugated to HRP (Bio-Rad Laboratories, Inc., cat. no. 170-6515). They estimated HRP activity from chemiluminescence using HRP substrate (EMD Millipore, cat. no. WBKLS0500) and autoradiographic film (UltiDent Scientific, cat. no. 39-20810).

4.5.5 Flow cytometric measurements of DNA content

To obtain DNA content distributions as a function of MMS dose Mila Tepliakova and I performed 4h MMS exposure experiments with strain N05 as described above (§4.5.2). At the end of MMS treatment we transferred a select volume of sample to a 1.5 ml snap cap centrifuge tube and followed Adam Rudner's DNA staining protocol with modification for small volume samples. Briefly, to process a comparable number of cells per sample, we transferred either 24 ul, 125 ul or 200 ul of sample to a 1.5 ml tube, depending on MMS concentration of sample. Once transferred, we pelleted cells and washed with 50 mM NaCitrate buffer (pH 7.4), followed by 1 h incubation at 50°C in 0.25 mg/ml RNase solution in buffer, and 1 h incubation at 60°C in with proteinase K solution in buffer added to final concentration of 0.25 mg/ml proteinase K (Sigma, cat. no. P4850) and 3 mM $CaCl_2$. Following RNA and protein degradation, we added propidium iodide (Sigma, cat. no. P4170) solution in buffer to a final concentration of 0.2 μg/ml and incubated in the dark (30 min, room temperature). We then sonicated cells and estimated cell distributions of propidium iodide by flow cytometry. We acquired flow cytometric measurements as described for gene expression, with the following modifications. Cell propidium iodide fluorescence was estimated from Log pulse height of 603-623 nm filtered emission with 488 nm excitation, from within a cell population gate corresponding to 70% of events per sample.

4.5.6 Fluorescence-activated cell sorting

A MoFlo Astrios Cell Sorter (Beckman Coulter Inc.) was used to sort cells representative of each mode from a cell population with bimodal $P_{RNR3} - GFP$ expression. Sorting was conducted by Vera A. Tang, Ph.D. of the University of Ottawa Flow Cytometry and Virometry Core Facility. To sort cells, Vera and I designed combinations of logical gates incorporating all five gates set based on morphology or green fluorescence (G1-G5). A cell population gate (G1) was defined by Pulse Area of 488 nm forward and side scatter signals. Wide gates corresponding to High (G3) and low (G2) Gfp subpopulation gates were each set based on Log Pulse Width of green fluorescence. Narrow gates corresponding to modes of high (G5) and low (G4) Gfp subpopulations were set based on Log Pulse Height of green fluorescence. Green fluorescence was estimated from 487-529 nm emission signals with 488 nm excitation. Sorted low Gfp cells were defined as being within gates G1, G2, G4, and not G3 and G5. Sorted high Gfp cells were defined as being within gates G1, G3 and G5, and not G2 and G4. Accuracy of the sort was verified by re-measuring the above parameters of sorted cells. While the sorted subpopulations were slightly wider than G4 and G5, respectively, the two sorted subpopulations nonetheless had the expected modes and were non-overlapping in Pulse height of green fluorescence.

4.5.7 Fluorescence microscopy

Before microscopy I increased the density of the culture to optimize number of cells per field of view. Briefly, I pelleted cells by centrifugation (900 g for 2 min), and siphoned off most supernatant to resuspend cells in 10-50 μl. I then pipetted 10 μl of the cell suspension onto a glass slide and acquired brightfield and fluorescence images. To acquire images I used an Eclipse Ti inverted microscope (Nikon) equipped with an

Intensilight C-HGFI Mercury Lamp (Nikon), a 60x/1.40 Plan Apo VC Oil Objective (Nikon), a Cool Snap HQ2 CCD camera (Photometrics), NIS Elements acquisition software (Version 3.22, Nikon), and the following excitation/emission light filter cubes: red fluorescence, 510-560 nm excitation and 575-645 nm emission (Chroma Technology Corp., cat. no. 41002); green fluorescence, 460-500 nm excitation and 510-560 nm emission (Chroma Technology Corp., cat. no. 41001); blue fluorescence, 378-402 nm excitation and 435-485 nm emission (Chroma Technology Corp., cat. no. 96340). Unless otherwise stated, lamp light intensity setting was ND 2 for non time-lapse measurements. Time-lapse microscopy experiments were conducted with the above-described microscope setup and a CellASIC® ONIX Microfluidic system (EMD Millipore) with CellASIC® ONIX Y04C Microfluidic plate (EMD Millipore) designed for haploid yeast. Rate of chamber perfusion was set to 2 psi for the duration of all experiments. Synthetic complete growth medium with and without MMS for these experiments was as described above (§4.5.2). Prior to loading cells, I primed chambers of the microfluidics plate by perfusion with SM for a minimum of 15 minutes at 8 psi.

5

Conclusion

In Chapter 2, I developed a gene topology model that reflects the classical rules of epistasis analysis. The model incorporates a number of modifications needed to improve applicability of classical rules to quantitative phenotypes. I described how to develop a method to infer topologies by comparing model-derived expected trait patterns to experimental genetic interaction data. These advancements are important because of extensive application of the classical rules in biological research despite the disagreement of data with classical assumptions.

Advancements to knowledge

1. Deduce assumptions needed for the classical rules to work
2. Formulate new rules to cover additional cases
3. Develop a method enabling automatic pathway inference

The author's contribution to these advancements was

- Generate experimental data
- Contribute to method development and testing
- For the original manuscript which describes this work [61], contribute to generation of figures and writing of manuscript

In Chapter 3, I generalized the rules of epistasis analysis to allow inference of any acyclic topology from quantitative genetic interaction experiments. This generalization increases knowledge about how discernible alternative topology models are for such experiments, both theoretically and in practice. This advance is important because these data are extensively gathered in biological and clinical research without thoroughly tested methodologies.

In addition, my analysis of experimental genetic interaction data identifies a gap in knowledge about transcription regulated by the DNA damage checkpoint (DDC) pathway, which is inexplicably inactive at the conditions tested.

Advancements to knowledge

1. Adapt the Battle scheme to enable discrimination among two-gene signal-responsive pathways
2. Demonstrate the deficiencies of the adapted scoring scheme
3. Create a theoretical framework for equivalence testing
4. Demonstrate predicted limits of topology inference from combinatorial perturbation data
5. Demonstrate the practical limitations of theoretical predictions

The author's contribution to these advancements was

- Analyze algorithm of Battle et al. [60]
- Contribute to code generation and benchmarking
- Document the limitations of the theoretical analysis and its implications for the interpretation of experimental data
- Generate experimental data and perform all analyses of experimental data
- Document apparent DDC pathway inactivity at high concentrations of methyl methanesulfonate (MMS)

To consider the implications of the contributions of Chapter 2 and 3, an important question is "has a universal method to genetic interaction data analysis been found and can one be found"? In the Introduction I provided examples of conflicting rules to interpret genetic interaction data from previous studies wherein the underlying assumptions were ambiguous. These conflicts and ambiguities set the focus of the thesis, that is to increase knowledge about the models which underly rules for interpreting these data.

In Chapter 2, the topology model presented allows one to find that conflicting rules of interpreting gene order of two previous studies, [43] and [55], can arise from this

model and set of assumptions about gene topology functions. The appearance of a conflict was therefore due to the different phenotype calculations performed to make an interpretation in these two previous studies. These results suggest that a universal set of assumptions may underly different methods of analysis.

In contrast, in Chapter 3, the model presented differs in assumptions from the model in Chapter 2, and the derived rules from the two models for interpreting gene order are in disagreement. By further study, I have found that the rule of Chapter 2 can be derived from the model of Chapter 3, if the assumptions are modified. This finding suggests that the assumptions are not universal, and that selection of the wrong assumption in some cases can result in the wrong conclusion about gene order. The finding therefore reveals another problem, that of determining when assumptions of a model are justified in application. For example, if the nature of gene activity is unknown, is it best to assume gene activity is an arbitrary or Boolean function of the activities of its upstream genes, or to make no interpretation at all?

The universality of genetic interaction data analysis also relates to the question of whether one method can be applied to any dataset of any organism. In Chapters 2 and 3, the methods were applied to datasets obtained by experiments with the unicellular eukaryote *S. cerevisiae*. In theory, the models used to derive these methods should be applicable to any model system wherein gene knockout can be achieved by gene perturbation. Applicability in this case is limited to perturbation technology. There is, however, a problem in applicability that is beyond the limits of perturbation technology: the assumption that a gene topology is the same for a given set of genes in every cell of a multicellular organism. This assumption may be violated often, either by heterogeneity in gene sequences among cells of an organism by acquired mutation, or by cell-to-cell differences in gene expression. Further knowledge about the consequences of these violations on gene topology inference, or how they might be circumvented would be required to address this problem.

In Chapter 4, I conduct biochemical, cell division and gene expression assays to discern possible causes of DDC network inactivity hypothesized in Chapter 3. These experiments are evidence that cells have an active DDC network and, despite this, are in states of cycle progression and transcription distinct from those typically reported under the conditions I test. This advance is important because it identifies a gap in knowledge of the regulation of cell division progression following DNA damage stress, which is a determinant of survival in these conditions.

Advancements to knowledge

1. Document non-monotonicity in *RNR3* transcriptional marker in response to MMS dose
2. Document correlation between cell-cycle phase and this non-monotonic response
3. Demonstrate G2 checkpoint "override"
4. Show phenotypic consequences of cell-cycle phase on cell survival following MMS treatment

The author's contribution to these advancements was

- Design, conduct and analyze all experiments
- Interpret the experiments in the context of current literature

Considering the implications of the contributions of Chapter 4, an important question is "does knowledge of a drug-induced override of a canonical yeast cell division cycle checkpoint influence other theoretical and clinical investigations"?

In Chapter 4, I provide experimental data supporting that methyl methanesulfonate (MMS) causes cells to override a DNA damage checkpoint. Addressing whether this finding has implications for analogous drug treatment of human cells, for example, would require more information. Such information would include identification or prediction of cellular nucleophiles that are alkylated in *S. cerevisiae* following

MMS treatment, and the differences in MMS reactants in human versus yeast cells, for example. Notably, Busulfan (1,4-butanediol dimethanesulfonate) is a clinically relevant antineoplastic drug [139] having two methanesulfonate moieties. Each moitety is modelled to react, as MMS [133], with cellular nucleophiles by S_N2 reaction [139]. Busulfan therefore may be more relevant to study in establishing the implications of the present work in a clinical context.

A second part of the question relates to the paradigm of the eukaryotic cell division cycle. Theoretically, a drug-induced override of a checkpoint, allowing S or G2/M cells to re-enter a G1-like state, conflicts with the paradigm that cell division cycle transitions between phases are "irreversible" [140]. Such irreversible transitions explain how generations of cells have conserved genome and centrosome number. Because abnormalities in genome copy and centrosome number are common in cancerous or transformed cells [141], improved understanding of such disease states may require increasing knowledge about the theoretical and experimental conditions that allow transitions to be reversible.

References

[1] Simon KG Forsberg, Joshua S Bloom, Meru J Sadhu, Leonid Kruglyak, and Örjan Carlborg. "Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast". In: *Nature Genetics* 49.4 (2017), pp. 497–503.

[2] Albert-Laszlo Barabasi and Zoltan N Oltvai. "Network biology: understanding the cell's functional organization". In: *Nature reviews. Genetics* 5.2 (2004), p. 101.

[3] Clyde A Hutchison, Ray-Yuan Chuang, Vladimir N Noskov, Nacyra Assad-Garcia, Thomas J Deerinck, Mark H Ellisman, John Gill, Krishna Kannan, Bogumil J Karas, Li Ma, James F. Pelletier, Zhi-Qing Qi, R. Alexander Richter, Elizabeth A. Strychalski, Lijie Sun, Yo Suzuki, Billyana Tsvetanova, Kim S. Wise, Hamilton O. Smith, John I. Glass, Chuck Merryman, Daniel G. Gibson, and J. Craig Venter. "Design and synthesis of a minimal bacterial genome". In: *Science* 351.6280 (2016), aad6253.

[4] Amy Hin Yan Tong, Guillaume Lesage, Gary D Bader, Huiming Ding, Hong Xu, Xiaofeng Xin, James Young, Gabriel F Berriz, Renee L Brost, Michael Chang, YiQun Chen, Xin Cheng, Gordon Chua, Helena Friesen, Debra S. Goldberg, Jennifer Haynes, Christine Humphries, Grace He, Shamiza Hussein, Lizhu Ke, Nevan J Krogan, Zhijian Li, Joshua N. Levinson, Hong Lu, Patrice Ménard, Christella Munyana, Ainslie B. Parsons, Owen Ryan, Raffi Tonikian, Tania Roberts, Anne-Marie Sdicu, Jesse Shapiro, Bilal Sheikh, Bernhard Suter, Sharyl L. Wong,

Lan V. Zhang, Hongwei Zhu, Christopher G. Burd, Sean Munro, Chris Sander, Jasper Rine, Jack F Greenblatt, Matthias Peter, Anthony Bretscher, Graham Bell, Frederick P. Roth, Grant W. Brown, Brenda J. Andrews, Howard Bussey, and Charles Boone. "Global mapping of the yeast genetic interaction network". In: *Science* 303.5659 (2004), pp. 808–813.

[5] Gibran Hemani, Sara Knott, and Chris Haley. "An evolutionary perspective on epistasis and the missing heritability". In: *PLoS genetics* 9.2 (2013), e1003295.

[6] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. "The mystery of missing heritability: Genetic interactions create phantom heritability". In: *Proceedings of the National Academy of Sciences* 109.4 (2012), pp. 1193–1198.

[7] Rong Chen, Lisong Shi, Jörg Hakenberg, Brian Naughton, Pamela Sklar, Jianguo Zhang, Hanlin Zhou, Lifeng Tian, Om Prakash, Mathieu Lemire, Patrick Sleiman, Wei-yi Cheng, Wanting Chen, Hardik Shah, Yulan Shen, Menachem Fromer, Larsson Omberg, Matthew A Deardorff, Elaine Zackai, Jason R Bobe, Elissa Levin, Thomas J Hudson, Leif Groop, Jun Wang, Hakon Hakonarson, Anne Wojcicki, George A Diaz, Lisa Edelmann, Eric E Schadt, and Stephen H Friend. "Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases". In: *Nature Biotechnology* 34.5 (2016), pp. 531–538.

[8] Yunhua Liu, Xinna Zhang, Cecil Han, Guohui Wan, Xingxu Huang, Cristina Ivan, Dahai Jiang, Cristian Rodriguez-Aguayo, Gabriel Lopez-Berestein, Pulivarthi H. Rao, Dipen M. Maru, Andreas Pahl, Xiaoming He, Anil K. Sood, Lee M. Ellis, Jan Anderl, and Xiongbin Lu. "TP53 loss creates therapeutic vulnerability in colorectal cancer". In: *Nature* 520.7549 (2015), p. 697.

[9] Kirk J McManus, Irene J Barrett, Yasaman Nouhi, and Philip Hieter. "Specific synthetic lethal killing of RAD54B-deficient human colorectal cancer cells by FEN1

silencing". In: *Proceedings of the National Academy of Sciences* 106.9 (2009), pp. 3276–3281.

[10] Philip E Hartman and John R Roth. "Mechanisms of suppression". In: *Advances in genetics* 17 (1973), pp. 1–105.

[11] Calvin B. Bridges. "The origin of variation". In: *The American Naturalist* 56 (1922), pp. 51–63.

[12] Theodore Dobzhansky. "Genetics of natural populations. XIII. Recombination and variability in populations of Drosophila pseudoobscura". In: *Genetics* 31.3 (1946), p. 269.

[13] Sebastian Nijman. "Synthetic lethality: general principles, utility and detection using genetic screens in human cells". In: *FEBS letters* 585.1 (2011), pp. 1–6.

[14] Jolanda van Leeuwen, Carles Pons, Joseph C Mellor, Takafumi N Yamaguchi, Helena Friesen, John Koschwanez, Mojca Mattiazzi Ušaj, Maria Pechlaner, Mehmet Takar, Matej Ušaj, Benjamin VanderSluis, Kerry Andrusiak, Pritpal Bansal, Anastasia Baryshnikova, Claire E. Boone, Jessica Cao, Atina Cote, Marinella Gebbia, Gene Horecka, Ira Horecka, Elena Kuzmin, Nicole Legro, Wendy Liang, Natascha van Lieshout, Margaret McNee, Bryan-Joseph San Luis, Fatemeh Shaeri, Ermira Shuteriqi, Song Sun, Lu Yang, Ji-Young Youn, Michael Yuen, Michael Costanzo, Anne-Claude Gingras, Patrick Aloy, Chris Oostenbrink, Andrew Murray, Todd R. Graham, Chad L. Myers, Brenda J. Andrews, Frederick P. Roth, and Charles Boone. "Exploring genetic suppression interactions on a global scale". In: *Science* 354.6312 (2016), aag0839.

[15] A Baudin, O Ozier-Kalogeropoulos, A Denouel, F Lacroute, and C Cullin. "A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae." In: *Nucleic acids research* 21.14 (1993), p. 3329.

[16] Achim Wach, Arndt Brachat, Christina Alberti-Segui, Corinne Rebischung, and Peter Philippsen. “Heterologous HIS3 marker and GFP reporter modules for PCR-targeting in Saccharomyces cerevisiae”. In: *Yeast* 13.11 (1997), pp. 1065–1075.

[17] Elizabeth A. Winzeler, Daniel D. Shoemaker, Anna Astromoff, Hong Liang, Keith Anderson, Bruno Andre, Rhonda Bangham, Rocio Benito, Jef D. Boeke, Howard Bussey, Angela M. Chu, Carla Connelly, Karen Davis, Fred Dietrich, Sally Whelen Dow, Mohamed El Bakkoury, Françoise Foury, Stephen H. Friend, Erik Gentalen, Guri Giaever, Johannes H. Hegemann, Ted Jones, Michael Laub, Hong Liao, Nicole Liebundguth, David J. Lockhart, Anca Lucau-Danila, Marc Lussier, Nasiha M’Rabet, Patrice Menard, Michael Mittmann, Chai Pai, Corinne Rebischung, Jose L. Revuelta, Linda Riles, Christopher J. Roberts, Petra Ross-MacDonald, Bart Scherens, Michael Snyder, Sharon Sookhai-Mahadeo, Reginald K. Storms, Steeve Véronneau, Marleen Voet, Guido Volckaert, Teresa R. Ward, Robert Wysocki, Grace S. Yen, Kexin Yu, Katja Zimmermann, Peter Philippsen, Mark Johnston, and Ronald W. Davis. “Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis”. In: *Science* 285.5429 (1999), pp. 901–906.

[18] Amy Hin Yan Tong and Charles Boone. “16 High-Throughput Strain Construction and Systematic Synthetic Lethal Screening in”. In: *Methods in Microbiology* 36 (2007), pp. 369–707.

[19] Michael Costanzo, Benjamin VanderSluis, Elizabeth N. Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, Matej Usaj, Julia Hanchard, Susan D. Lee, Vicent Pelechano, Erin B. Styles, Maximilian Billmann, Jolanda van Leeuwen, Nydia van Dyk, Zhen-Yuan Lin, Elena Kuzmin, Justin Nelson, Jeff S. Piotrowski, Tharan Srikumar, Sondra Bahr, Yiqun Chen, Raamesh Deshpande, Christoph F. Kurat, Sheena C. Li, Zhijian Li, Mojca Mattiazzi Usaj, Hiroki Okada, Natasha Pascoe, Bryan-Joseph San Luis,

Sara Sharifpoor, Emira Shuteriqi, Scott W. Simpkins, Jamie Snider, Harsha Garadi Suresh, Yizhao Tan, Hongwei Zhu, Noel Malod-Dognin, Vuk Janjic, Natasa Przulj, Olga G. Troyanskaya, Igor Stagljar, Tian Xia, Yoshikazu Ohya, Anne-Claude Gingras, Brian Raught, Michael Boutros, Lars M. Steinmetz, Claire L. Moore, Adam P. Rosebrock, Amy A. Caudy, Chad L. Myers, Brenda Andrews, and Charles Boone. "A global genetic interaction network maps a wiring diagram of cellular function". In: *Science* 353.6306 (2016), aaf1420.

[20] Assen Roguev, Marianna Wiren, Jonathan S Weissman, and Nevan J Krogan. "High-throughput genetic interaction mapping in the fission yeast Schizosaccharomyces pombe". In: *Nature methods* 4.10 (2007), p. 861.

[21] Assen Roguev, Colm J Ryan, Edgar Hartsuiker, and Nevan J Krogan. "High-throughput quantitative genetic interaction mapping in the fission yeast Schizosaccharomyces pombe". In: *Cold Spring Harbor Protocols* (2017).

[22] Gareth Butland, Mohan Babu, J Javier Dıaz-Mejıa, Fedyshyn Bohdana, Sadhna Phanse, Barbara Gold, Wenhong Yang, Joyce Li, Alla G Gagarinova, Oxana Pogoutse, Hirotada Mori, Barry L Wanner, Henry Lo, Jas Wasniewski, Constantine Christopoulos, Mehrab Ali, Mehrab Mehrab Venn, Anahita Safavi-Naini, Natalie Sourour, Simone Caron, Ja-Yeon Choi, Ludovic Laigle, Anaies Nazarians-Armavil, Avnish Deshpande, Sarah Joe, Kirill A Datsenko, Natsuko Yamamoto, Brenda J Andrews, Charles Boone, Huiming Ding, Bilal Sheikh, Gabriel Moreno-Hagelsieb, Jack F Greenblatt, and Andrew Emili. "eSGA: E. coli synthetic genetic array analysis". In: *Nature methods* 5.9 (2008), pp. 789–795.

[23] Ben Lehner, Catriona Crombie, Julia Tischler, Angelo Fortunato, and Andrew G Fraser. "Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways". In: *Nature genetics* 38.8 (2006), p. 896.

[24] Thomas Horn, Thomas Sandmann, Bernd Fischer, Elin Axelsson, Wolfgang Huber, and Michael Boutros. “Mapping of signaling networks through synthetic genetic interaction analysis by RNAi”. In: *Nature methods* 8.4 (2011), pp. 341–346.

[25] Assen Roguev, Dale Talbot, Gian Luca Negri, Michael Shales, Gerard Cagney, Sourav Bandyopadhyay, Barbara Panning, and Nevan J Krogan. “Quantitative genetic-interaction mapping in mammalian cells”. In: *Nature methods* 10.5 (2013), pp. 432–437.

[26] Christina Laufer, Bernd Fischer, Maximilian Billmann, Wolfgang Huber, and Michael Boutros. “Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping”. In: *Nature methods* 10.5 (2013), pp. 427–431.

[27] Ophir Shalem, Neville E. Sanjana, Ella Hartenian, Xi Shi, David A. Scott, Tarjei Mikkelson, Dirk Heckl, Benjamin L. Ebert, David E. Root, John G. Doench, and Feng Zhang. “Genome-scale CRISPR-Cas9 knockout screening in human cells”. In: *Science* 343.6166 (2014), pp. 84–87.

[28] Bastiaan Evers, Katarzyna Jastrzebski, Jeroen PM Heijmans, Wipawadee Grernrum, Roderick L Beijersbergen, and Rene Bernards. “CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes”. In: *Nature Biotechnology* (2016).

[29] Jeffry D Sander and J Keith Joung. “CRISPR-Cas systems for editing, regulating and targeting genomes”. In: *Nature Biotechnology* 32.4 (2014), pp. 347–355.

[30] Kyuho Han, Edwin E Jeng, Gaelen T Hess, David W Morgens, Amy Li, and Michael C Bassik. “Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions”. In: *Nature Biotechnology* 35.5 (2017), pp. 463–474.

[31] John Paul Shen, Dongxin Zhao, Roman Sasik, Jens Luebeck, Amanda Birmingham, Ana Bojorquez-Gomez, Katherine Licon, Kristin Klepper, Daniel Pekin, Alex N Beckett, Kyle Salinas Sanchez, Alex Thomas, Chih-Chung Kuo, Dan Du,

Assen Roguev, Nathan E Lewis, Aaron N Chang, Jason F Kreisberg, Nevan Krogan, Lei Qi, Trey Ideker, and Prashant Mali. “Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions”. In: *Nature methods* 14.6 (2017), pp. 573–576.

[32] Shengdar Q Tsai, Zongli Zheng, Nhu T Nguyen, Matthew Liebers, Ved V Topkar, Vishal Thapar, Nicolas Wyvekens, Cyd Khayter, A John Iafrate, Long P Le, Martin J Aryee, and J Keith Joung. “GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases”. In: *Nature Biotechnology* 33.2 (2015), pp. 187–197.

[33] Tim Wang, Jenny J Wei, David M Sabatini, and Eric S Lander. “Genetic screens in human cells using the CRISPR-Cas9 system”. In: *Science* 343.6166 (2014), pp. 80–84.

[34] Benjamin P Kleinstiver, Vikram Pattanayak, Michelle S Prew, Shengdar Q Tsai, Nhu T Nguyen, Zongli Zheng, and J Keith Joung. “High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects”. In: *Nature* 529.7587 (2016), pp. 490–495.

[35] Dan Du, Assen Roguev, David E Gordon, Meng Chen, Si-Han Chen, Michael Shales, John Paul Shen, Trey Ideker, Prashant Mali, Lei S Qi, and Nevan J Krogan. “Genetic interaction mapping in mammalian cells using CRISPR interference”. In: *Nature Methods* (2017).

[36] Martin C. Jonikas, Sean R. Collins, Vladimir Denic, Eugene Oh, Erin M. Quan, Volker Schmid, Jimena Weibezahn, Blanche Schwappach, Peter Walter, Jonathan S. Weissman, and Maya Schuldiner. “Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum”. In: *Science* 323.5922 (2009), pp. 1693–1697.

[37] Jincai Li, Shu Wang, William J. VanDusen, Loren D. Schultz, Hugh A. George, Wayne K. Herber, Hee Jeong Chae, William E. Bentley, and Govind Rao. “Green

fluorescent protein in Saccharomyces cerevisiae: real-time studies of the GAL1 promoter". In: *Biotechnology and bioengineering* 70.2 (2000), pp. 187–196.

[38] Won-Ki Huh, James V. Falvo, Luke C. Gerke, Adam S. Carroll, Russell W. Howson, Jonathan S. Weissman, and Erin K. O'Shea. "Global analysis of protein localization in budding yeast". In: *Nature* 425.6959 (2003), p. 686.

[39] Bernd Fischer, Thomas Sandmann, Thomas Horn, Maximilian Billmann, Varun Chaudhary, Wolfgang Huber, and Michael Boutros. "A map of directional genetic interactions in a metazoan cell". In: *Elife* 4 (2015), e05464.

[40] Evan S Snitkin and Daniel Segrè. "Epistatic interaction maps relative to multiple metabolic phenotypes". In: *PLoS genetics* 7.2 (2011), e1001294.

[41] Cory Batenchuk, Lioudmila Tepliakova, and Mads Kærn. "Identification of response-modulated genetic interactions by sensitivity-based epistatic analysis". In: *BMC genomics* 11.1 (2010), p. 1.

[42] Ramamurthy Mani, Robert P St Onge, John L Hartman, Guri Giaever, and Frederick P Roth. "Defining genetic interaction". In: *Proceedings of the National Academy of Sciences* 105.9 (2008), pp. 3461–3466.

[43] Robert P St Onge, Ramamurthy Mani, Julia Oh, Michael Proctor, Eula Fung, Ronald W Davis, Corey Nislow, Frederick P Roth, and Guri Giaever. "Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions". In: *Nature genetics* 39.2 (2007), pp. 199–206.

[44] Sourav Bandyopadhyay, Monika Mehta, Dwight Kuo, Min-Kyung Sung, Ryan Chuang, Eric J. Jaehnig, Bernd Bodenmiller, Katherine Licon, Wilbert Copeland, Michael Shales, Dorothea Fiedler, Janusz Dutkowski, Aude Guénolé, Haico van Attikum, Kevan M. Shokat, Richard D. Kolodner, Won-Ki Huh, Ruedi Aebersold, Michael-Christopher Keogh, Nevan J. Krogan, and Trey Ideker. "Rewiring of genetic networks in response to DNA damage". In: *Science* 330.6009 (2010), pp. 1385–1389.

[45] Amy Hin Yan Tong, Marie Evangelista, Ainslie B. Parsons, Hong Xu, Gary D. Bader, Nicholas Pagé, Mark Robinson, Sasan Raghibizadeh, Christopher W. V. Hogue, Howard Bussey, Brenda J. Andrews, Mike Tyers, and Charles Boone. "Systematic genetic analysis with ordered arrays of yeast deletion mutants". In: *Science* 294.5550 (2001), pp. 2364–2368.

[46] John L Hartman, Barbara Garvik, and Lee Hartwell. "Principles for the buffering of genetic variation". In: *Science* 291.5506 (2001), pp. 1001–1004.

[47] Chad L Myers, Daniel R Barrett, Matthew A Hibbs, Curtis Huttenhower, and Olga G Troyanskaya. "Finding function: evaluation methods for functional genomic data". In: *BMC genomics* 7.1 (2006), p. 187.

[48] Sean R Collins, Assen Roguev, and Nevan J Krogan. "Quantitative genetic interaction mapping using the E-MAP approach". In: *Methods in enzymology* 470 (2010), pp. 205–231.

[49] Anastasia Baryshnikova, Michael Costanzo, Yungil Kim, Huiming Ding, Judice Koh, Kiana Toufighi, Ji-Young Youn, Jiongwen Ou, Bryan-Joseph San Luis, Sunayan Bandyopadhyay, M Hibbs, D Hess, Anne-Claude Gingras, Gary D Bader, Olga G Troyanskaya, Grant W Brown, Brenda Andrews, Charles Boone, and Chad L Myers. "Quantitative analysis of fitness and genetic interactions in yeast on a genome scale". In: *Nature methods* 7.12 (2010), pp. 1017–1024.

[50] Juan I Fuxman Bass, Alos Diallo, Justin Nelson, Juan M Soto, Chad L Myers, and Albertha JM Walhout. "Using networks to measure similarity between genes: association index selection". In: *Nature methods* 10.12 (2013), pp. 1169–1176.

[51] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D. Spear, Carolyn S. Sevier, Huiming Ding, Judice L.Y. Koh, Kiana Toufighi, Sara Mostafavi, Jeany Prinz, Robert P. St. Onge, Benjamin VanderSluis, Taras Makhnevych, Franco J. Vizeacoumar, Solmaz Alizadeh, Sondra Bahr, Renee L. Brost, Yiqun Chen, Murat Cokol, Raamesh Deshpande, Zhijian Li,

Zhen-Yuan Lin, Wendy Liang, Michaela Marback, Jadine Paw, Bryan-Joseph San Luis, Ermira Shuteriqi, Amy Hin Yan Tong, Nydia van Dyk, Iain M. Wallace, Joseph A. Whitney, Matthew T. Weirauch, Guoqing Zhong, Hongwei Zhu, Walid A. Houry, Michael Brudno, Sasan Ragibizadeh, Balázs Papp, Csaba Pál, Frederick P. Roth, Guri Giaever, Corey Nislow, Olga G. Troyanskaya, Howard Bussey, Gary D. Bader, Anne-Claude Gingras, Quaid D. Morris, Philip M. Kim, Chris A. Kaiser, Chad L. Myers, Brenda J. Andrews, and Charles Boone. "The genetic landscape of a cell". In: *Science* 327.5964 (2010), pp. 425–431.

[52] Maya Schuldiner, Sean R Collins, Natalie J Thompson, Vladimir Denic, Arunashree Bhamidipati, Thanuja Punna, Jan Ihmels, Brenda Andrews, Charles Boone, Jack F Greenblatt, Jonathan S Weissman, and Nevan J Krogan. "Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile". In: *Cell* 123.3 (2005), pp. 507–519.

[53] Michael Costanzo, Anastasia Baryshnikova, Chad L Myers, Brenda Andrews, and Charles Boone. "Charting the genetic interaction map of a cell". In: *Current opinion in biotechnology* 22.1 (2011), pp. 66–74.

[54] Frederick P Roth, Howard D Lipshitz, and Brenda J Andrews. "Q & A: epistasis". In: *Journal of Biology* 8.4 (2009), p. 35.

[55] Leon Avery and Steven Wasserman. "Ordering gene function: the interpretation of epistasis in regulatory hierarchies". In: *Trends in genetics* 8.9 (1992), pp. 312–316.

[56] William Bateson. "Facts limiting the theory of heredity". In: *Science* (1907), pp. 649–660.

[57] Becky L Drees, Vesteinn Thorsson, Gregory W Carter, Alexander W Rives, Marisa Z Raymond, Iliana Avila-Campillo, Paul Shannon, and Timothy Galitski. "Derivation of genetic interaction networks from quantitative phenotype data". In: *Genome biology* 6.4 (2005), R38.

[58] David L Aylor and Zhao-Bang Zeng. “From classical genetics to quantitative genetics to systems biology: modeling epistasis”. In: *PLoS Genet* 4.3 (2008), e1000029.

[59] Nancy Van Driessche, Janez Demsar, Ezgi O Booth, Paul Hill, Peter Juvan, Blaz Zupan, Adam Kuspa, and Gad Shaulsky. “Epistasis analysis with global transcriptional phenotypes”. In: *Nature genetics* 37.5 (2005), p. 471.

[60] Alexis Battle, Martin C Jonikas, Peter Walter, Jonathan S Weissman, and Daphne Koller. “Automated identification of pathways from quantitative genetic interaction data”. In: *Molecular systems biology* 6.1 (2010), p. 379.

[61] Hilary Phenix, Katy Morin, Cory Batenchuk, Jacob Parker, Vida Abedi, Liu Yang, Lioudmila Tepliakova, Theodore J Perkins, and Mads Kærn. “Quantitative epistasis analysis and pathway inference from genetic interaction data”. In: *PLoS Comput Biol* 7.5 (2011), e1002048.

[62] Glaucia Mendes Souza, Aline Maria da Silva, and Adam Kuspa. “Starvation promotes Dictyostelium development by relieving PufA inhibition of PKA translation through the YakA kinase pathway”. In: *Development* 126.14 (1999), pp. 3263–3274.

[63] David J Timson. “Galactose metabolism in Saccharomyces cerevisiae”. In: *Dynamic Biochemistry, Process Biotechnology and Molecular Biology* 1.1 (2007), pp. 63–73.

[64] Siddhartha Majumdar, Jhuma Ghatak, Sucheta Mukherji, Hiranmoy Bhattacharjee, and Amar Bhaduri. “UDPgalactose 4-epimerase from Saccharomyces cerevisiae. A bifunctional enzyme with aldose 1-epimerase activity.” In: *The FEBS Journal* 271.4 (2004), pp. 753–759.

[65] Raluca Apostu and Michael C Mackey. “Mathematical model of GAL regulon dynamics in Saccharomyces cerevisiae”. In: *Journal of theoretical biology* 293 (2012), pp. 219–235.

[66] Wenjin Zheng, H Eric Xu, and Stephen Albert Johnston. “The cysteine-peptidase bleomycin hydrolase is a member of the galactose regulon in yeast”. In: *Journal of Biological Chemistry* 272.48 (1997), pp. 30350–30355.

[67] Murat Acar, Attila Becskei, and Alexander van Oudenaarden. “Enhancement of cellular memory by reducing stochastic transitions”. In: *Nature* 435.7039 (2005), p. 228.

[68] Carl Song, Hilary Phenix, Vida Abedi, Matthew Scott, Brian P Ingalls, Mads Kærn, and Theodore J Perkins. “Estimating the stochastic bifurcation structure of cellular networks”. In: *PLoS computational biology* 6.3 (2010), e1000699.

[69] Karen Lai, Matthew J Robertson, and David V Schaffer. “The sonic hedgehog signaling system as a bistable genetic switch”. In: *Biophysical Journal* 86.5 (2004), pp. 2748–2757.

[70] C Baker Brachmann, Adrian Davies, Gregory J Cost, Emerita Caputo, Joachim Li, Philip Hieter, and Jef D Boeke. “Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications”. In: *Yeast* 14 (1998), pp. 115–132.

[71] Brendan P Cormack, Gwyneth Bertram, Mark Egerton, Neil AR Gow, Stanley Falkow, and Alistair JP Brown. “Yeast-enhanced green fluorescent protein (yEGFP): a reporter of gene expression in Candida albicans”. In: *Microbiology* 143.2 (1997), pp. 303–311.

[72] Afnan Azizi, Wilson Lam, Hilary Phenix, Lioudmila Tepliakova, Ian J. Roney, Daniel Jedrysiak, Alex Power, Vaibhav Gupta, Nada Elnour, Martin Hanzel, Alexandra C. Tzahristos, Shihab Sarwar, and Mads Kærn. “No training required: experimental tests support homology-based DNA assembly as a best practice in synthetic biology”. In: *Journal of biological engineering* 9.1 (2015), p. 8.

[73] Mark Johnston, Jeffrey S Flick, and Terry Pexton. “Multiple mechanisms provide rapid and stringent glucose repression of GAL gene expression in Saccharomyces cerevisiae.” In: *Molecular and cellular biology* 14.6 (1994), pp. 3834–3841.

[74] Hilary Phenix, Theodore Perkins, and Mads Kærn. “Identifiability and inference of pathway motifs by epistasis analysis”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 23.2 (2013), p. 025103.

[75] Eugenio Azpeitia, Mariana Benıtez, Pablo Padilla-Longoria, Carlos Espinosa-Soto, and Elena R Alvarez-Buylla. “Dynamic network-based epistasis analysis: Boolean examples”. In: *frontiers in Plant Science* 2 (2011), p. 92.

[76] Daniel Marbach, Robert J Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. “Revealing strengths and weaknesses of methods for gene network inference”. In: *Proceedings of the National Academy of Sciences* 107.14 (2010), pp. 6286–6291.

[77] Thomas Schaffter, Daniel Marbach, and Dario Floreano. “GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods”. In: *Bioinformatics* 27.16 (2011), pp. 2263–2270.

[78] Sara Hooshangi, Stephan Thiberge, and Ron Weiss. “Ultrasensitivity and noise propagation in a synthetic transcriptional cascade”. In: *Proceedings of the National Academy of Sciences* 102.10 (Feb. 2005), pp. 3581–3586. URL: http://dx.doi.org/10.1073/pnas.0408507102.

[79] Gilda Piaggio, Diana R. Elbourne, Douglas G. Altman, Stuart J. Pocock, and Stephen J. W. Evans. “Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement”. In: *Jama* 295.10 (2006), pp. 1152–1160.

[80] Manesh R. Patel, Kenneth W. Mahaffey, Jyotsna Garg, Guohua Pan, Daniel E. Singer, Werner Hacke, Günter Breithardt, Jonathan L. Halperin, Graeme J. Hankey, Jonathan P. Piccini, Richard C. Becker, Christopher C. Nessel,

John F. Paolini, Scott D. Berkowitz, Keith A.A. Fox, Robert M. Califf, and ROCKET AF Steering Committee. “Rivaroxaban versus warfarin in nonvalvular atrial fibrillation”. In: *New England Journal of Medicine* 365.10 (2011), pp. 883–891.

[81] James B Allen, Zheng Zhou, Wolfram Siede, Errol C Friedberg, and Stephen J Elledge. “The SAD1/RAD53 protein kinase controls multiple checkpoints and DNA damage-induced transcription in yeast.” In: *Genes & Development* 8.20 (1994), pp. 2401–2415.

[82] Mingxia Huang, Zheng Zhou, and Stephen J Elledge. “The DNA replication and damage checkpoint pathways induce transcription by inhibition of the Crt1 repressor”. In: *Cell* 94.5 (Sept. 1998), pp. 595–605. URL: http://dx.doi.org/10.1016/S0092-8674(00)81601-3.

[83] Zheng Zhou and Stephen J. Elledge. “DUN1 encodes a protein kinase that controls the DNA damage response in yeast”. In: *Cell* 75.6 (Dec. 1993), pp. 1119–1127. URL: http://dx.doi.org/10.1016/0092-8674(93)90321-G.

[84] Yolanda Sanchez, Brian A. Desany, William J. Jones, Qinghua Liu, Bin Wang, and Stephen J. Elledge. “Regulation of RAD53 by the ATM-like kinases MEC1 and TEL1 in yeast cell cycle checkpoint pathways”. In: *Science* 271.5247 (1996), p. 357.

[85] Xuming Jia, Yu Zhu, and Wei Xiao. “A stable and sensitive genotoxic testing system based on DNA damage induced gene expression in Saccharomyces cerevisiae”. In: *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* 519.1 (2002), pp. 83–92.

[86] Xuming Jia and Wei Xiao. “Compromised DNA repair enhances sensitivity of the yeast RNR3-lacZ genotoxicity testing system”. In: *Toxicological Sciences* 75.1 (2003), pp. 82–88.

[87] Kohei Ichikawa and Toshihiko Eki. “A Novel Yeast-Based Reporter Assay System for the Sensitive Detection of Genotoxic Agents Mediated by a DNA

Damage–Inducible LexA-GAL4 Protein". In: *Journal of biochemistry* 139.1 (2006), pp. 105–112.

[88] Yukari Ochi, Harumi Sugawara, Mio Iwami, Megumi Tanaka, and Toshihiko Eki. "Sensitive detection of chemical-induced genotoxicity by the Cypridina secretory luciferase reporter assay, using DNA repair-deficient strains of Saccharomyces cerevisiae". In: *Yeast* 28.4 (2011), pp. 265–278.

[89] Ting Wei, Chao Zhang, Xin Xu, Michelle Hanna, Xiaohua Zhang, Yan Wang, Heping Dai, and Wei Xiao. "Construction and evaluation of two biosensors based on yeast transcriptional response to genotoxic chemicals". In: *Biosensors and Bioelectronics* 44 (2013), pp. 138–145.

[90] Aprotim Mazumder, Laia Quiros Pesudo, Siobhan McRee, Mark Bathe, and Leona D Samson. "Genome-wide single-cell-level screen for protein abundance and localization changes in response to DNA damage in S. cerevisiae". In: *Nucleic acids research* (2013), gkt715.

[91] D Burke, D Dawson, and T Stearns. *Methods in yeast genetics: a Cold Spring Harbor Laboratory course manual*. 2000th ed. CSHL Press, 2000.

[92] IntelliCyt Corporation. *HyperCyt® System User's Guide –Version 3.2*. Tech. rep. 2009.

[93] Dana Branzei and Marco Foiani. "The checkpoint response to replication stress". In: *DNA Repair* 8.9 (Sept. 2009), pp. 1038–1046. URL: http://dx.doi.org/10.1016/j.dnarep.2009.04.014.

[94] Christopher S Gilbert, Catherine M Green, and Noel F Lowndes. "Budding yeast Rad9 is an ATP-dependent Rad53 activating machine". In: *Molecular Cell* 8.1 (July 2001), pp. 129–136. URL: http://dx.doi.org/10.1016/S1097-2765(01)00267-2.

[95] Zhaoxia Sun, James Hsiao, David S Fay, and David F Stern. "Rad53 FHA domain associated with phosphorylated Rad9 in the DNA damage checkpoint". In: *Science* 281.5374 (1998), pp. 272–274.

[96] Annette A Alcasabas, Alexander J Osborn, Jeff Bachant, Fenghua Hu, Petra JH Werler, Kristine Bousset, Kanji Furuya, John FX Diffley, Antony M Carr, and Stephen J Elledge. "Mrc1 transduces signals of DNA replication stress to activate Rad53". In: *Nature Cell Biology* 3.11 (2001), pp. 958–965.

[97] Zheng Zhou and Stephen J Elledge. "Isolation of crt mutants constitutive for transcription of the DNA damage inducible gene RNR3 in Saccharomyces cerevisiae." In: *Genetics* 131.4 (1992), pp. 851–866.

[98] Ted A Weinert and Leland H Hartwell. "The RAD9 gene controls the cell cycle response to DNA damage in Saccharomyces cerevisiae". In: *Science* 241.4863 (1988), pp. 317–322.

[99] David P Toczyski, David J Galgoczy, and Leland H Hartwell. "CDC5 and CKII control adaptation to the yeast DNA damage checkpoint". In: *Cell* 90.6 (1997), pp. 1097–1106.

[100] Jose Antonio Tercero and John FX Diffley. "Regulation of DNA replication fork progression through damaged DNA by Mec1/Rad53 checkpoint". In: *Nature* 412.6846 (2001), p. 553.

[101] Amanda G Paulovich and Leland H Hartwell. "A checkpoint regulates the rate of progression through S phase in S. cerevisiae in response to DNA damage". In: *Cell* 82.5 (1995), pp. 841–847.

[102] Hongfang Qiu, E Park, L Prakash, and Satya Prakash. "The Saccharomyces cerevisiae DNA repair gene RAD25 is required for transcription by RNA polymerase II." In: *Genes & Development* 7.11 (1993), pp. 2161–2171.

[103] Brian A Desany, Annette A Alcasabas, Jeffrey B Bachant, and Stephen J Elledge. "Recovery from DNA replicational stress is the essential function of the S-phase checkpoint pathway". In: *Genes & Development* 12.18 (1998), pp. 2956–2970.

[104] Vladimir I Bashkirov, Elena V Bashkirova, Edwin Haghnazari, and Wolf-Dietrich Heyer. "Direct kinase-to-kinase signaling mediated by the FHA

phosphoprotein recognition domain of the Dun1 DNA damage checkpoint kinase". In: *Molecular and Cellular Biology* 23.4 (2003), pp. 1441–1452.

[105] Madhu Dyavaiah, John P Rooney, Sridar V Chittur, Qishan Lin, and Thomas J Begley. "Autophagy-dependent regulation of the DNA damage response protein ribonucleotide reductase 1". In: *Molecular Cancer Research* 9.4 (2011), pp. 462–475.

[106] Ulrike Begley, Madhu Dyavaiah, Ashish Patil, John P Rooney, Dan DiRenzo, Christine M Young, Douglas S Conklin, Richard S Zitomer, and Thomas J Begley. "Trm9-catalyzed tRNA modifications link translation to the DNA damage response". In: *Molecular cell* 28.5 (2007), pp. 860–870.

[107] Henning Althoefer, Alexander Schleiffer, Katja Wassmann, Alfred Nordheim, and Gustav Ammerer. "Mcm1 is required to coordinate G2-specific transcription in Saccharomyces cerevisiae." In: *Molecular and Cellular biology* 15.11 (1995), pp. 5917–5928.

[108] Tatjana Trcek, Daniel R Larson, Alberto Moldón, Charles C Query, and Robert H Singer. "Single-molecule mRNA decay measurements reveal promoter-regulated mRNA stability in yeast". In: *Cell* 147.7 (2011), pp. 1484–1497.

[109] Xiaorong Li and Mingjie Cai. "Recovery of the Yeast Cell Cycle from Heat Shock-induced G1 Arrest Involves a Positive Regulation of G1 Cyclin Expression by the S Phase Cyclin Clb5". In: *Journal of Biological Chemistry* 274.34 (1999), pp. 24220–24231.

[110] Yuen Ho, Michael Costanzo, Lynda Moore, Ryuji Kobayashi, and Brenda J Andrews. "Regulation of Transcription at the Saccharomyces cerevisiae Start Transition by Stb1, a Swi6-Binding Protein". In: *Molecular and cellular biology* 19.8 (1999), pp. 5267–5278.

[111] Jennifer Summers McKinney, Sunaina Sethi, Jennifer DeMars Tripp, Thuy N Nguyen, Brian A Sanderson, James W Westmoreland, Michael A Resnick,

and L Kevin Lewis. “A multistep genomic screen identifies new genes required for repair of DNA double-strand breaks in Saccharomyces cerevisiae”. In: *BMC genomics* 14 (2013), p. 251.

[112] Michael Lisby, Jacqueline H Barlow, Rebecca C Burgess, and Rodney Rothstein. “Choreography of the DNA damage response: spatiotemporal relationships among checkpoint and repair proteins”. In: *Cell* 118.6 (2004), pp. 699–713.

[113] Genevieve M Vidanes, Frédéric D Sweeney, Sarah Galicia, Stephanie Cheung, John P Doyle, Daniel Durocher, and David P Toczyski. “CDC5 inhibits the hyperphosphorylation of the checkpoint kinase Rad53, leading to checkpoint adaptation”. In: *PLoS biology* 8.1 (2010), e1000286.

[114] Hery Ratsima, Diego Serrano, Mirela Pascariu, and Damien D’Amours. “Centrosome-dependent bypass of the DNA damage checkpoint by the polo kinase Cdc5”. In: *Cell reports* 14.6 (2016), pp. 1422–1434.

[115] Federica Marini, Tiziana Nardo, Michele Giannattasio, Mario Minuzzo, Miria Stefanini, Paolo Plevani, and Marco Muzi Falconi. “DNA nucleotide excision repair-dependent signaling to checkpoint activation”. In: *Proceedings of the National Academy of Sciences* 103.46 (2006), pp. 17325–17330.

[116] Aprotim Mazumder, Katja Tummler, Mark Bathe, and Leona D Samson. “Single-cell analysis of ribonucleotide reductase transcriptional and translational response to DNA damage”. In: *Molecular and cellular biology* 33.3 (2013), pp. 635–642.

[117] Marta B Davidson, Yuki Katou, Andrea Keszthelyi, Tina L Sing, Tian Xia, Jiongwen Ou, Jessica A Vaisica, Neroshan Thevakumaran, Lisette Marjavaara, Chad L Myers, Andrei Chabes, Katsuhiko Shirahige, and Grant W Brown. “Endogenous DNA replication stress results in expansion of dNTP pools and a mutator phenotype”. In: *The EMBO journal* 31.4 (2012), pp. 895–907.

[118] Yuen Ho, Stephen Mason, Ryuji Kobayashi, Merl Hoekstra, and Brenda Andrews. "Role of the casein kinase I isoform, Hrr25, and the cell cycle-regulatory transcription factor, SBF, in the transcriptional response to DNA damage in Saccharomyces cerevisiae". In: *Proceedings of the National Academy of Sciences* 94.2 (1997), pp. 581–586.

[119] Joseph Ogas, Brenda J Andrews, and Ira Herskowitz. "Transcriptional activation of CLN1, CLN2, and a putative new G1 cyclin (HCS26) by SWI4, a positive regulator of G1-specific transcription". In: *Cell* 66.5 (1991), pp. 1015–1026.

[120] Michael Costanzo, Joy L Nishikawa, Xiaojing Tang, Jonathan S Millman, Oliver Schub, Kevin Breitkreuz, Danielle Dewar, Ivan Rupes, Brenda Andrews, and Mike Tyers. "CDK activity antagonizes Whi5, an inhibitor of G1/S transcription in yeast". In: *Cell* 117.7 (2004), pp. 899–913.

[121] Ruojin Yao, Zhen Zhang, Xiuxiang An, Brigid Bucci, Deborah L Perlstein, JoAnne Stubbe, and Mingxia Huang. "Subcellular localization of yeast ribonucleotide reductase regulated by the DNA replication and damage checkpoint pathways". In: *Proceedings of the National Academy of Sciences* 100.11 (2003), pp. 6628–6633.

[122] Grzegorz Ira, Achille Pellicioli, Alitukiriza Balijja, Xuan Wang, Simona Fiorani, Walter Carotenuto, Giordano Liberi, Debra Bressan, Lihong Wan, Nancy M Hollingsworth, James E. Haber, and Marco Foiani. "DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1". In: *Nature* 431.7011 (2004), p. 1011.

[123] Julia M Sidorova and Linda L Breeden. "Rad53-dependent phosphorylation of Swi6 and down-regulation of CLN1 and CLN2 transcription occur in response to DNA damage in Saccharomyces cerevisiae". In: *Genes & Development* 11.22 (1997), pp. 3032–3045.

[124] Wolfram Siede, Andrew S Friedberg, and Errol C Friedberg. "RAD9-dependent G1 arrest defines a second checkpoint for damaged DNA in the cell cycle of Saccharomyces cerevisiae". In: *Proceedings of the National Academy of Sciences* 90.17 (1993), pp. 7985–7989.

[125] Jonathan N Fitz Gerald, Jacqueline M Benjamin, and Stephen J Kron. "Robust G1 checkpoint arrest in budding yeast: dependence on DNA damage signaling and repair". In: *Journal of cell science* 115.8 (2002), pp. 1749–1757.

[126] José Antonio Tercero, Maria Pia Longhese, and John FX Diffley. "A central role for DNA replication forks in checkpoint activation and response". In: *Molecular cell* 11.5 (2003), pp. 1323–1336.

[127] Jacqueline H Barlow and Rodney Rothstein. "Rad52 recruitment is DNA replication independent and regulated by Cdc28 and the Mec1 kinase". In: *The EMBO Journal* 28.8 (2009), pp. 1121–1130.

[128] Etienne Schwob, Thomas Böhm, Michael D Mendenhall, and Kim Nasmyth. "The B-type cyclin kinase inhibitor $p40^{SIC1}$ controls the G1 to S transition in S. cerevisiae". In: *Cell* 79.2 (1994), pp. 233–244.

[129] Meredith EK Calvert, Joanne A Lannigan, and Lucy F Pemberton. "Optimization of yeast cell cycle analysis and morphological characterization by multispectral imaging flow cytometry". In: *Cytometry Part A* 73.9 (2008), pp. 825–833.

[130] Bruce A Edgar, Norman Zielke, and Crisanto Gutierrez. "Endocycles: a recurrent evolutionary innovation for post-mitotic cell growth". In: *Nature Reviews Molecular Cell Biology* 15.3 (2014), pp. 197–210.

[131] Jacqueline Hayles, Daniel Fisher, Alison Woollard, and Paul Nurse. "Temporal order of S phase and mitosis in fission yeast is determined by the state of the $p34^{cdc2}$-mitotic B cyclin complex". In: *Cell* 78.5 (1994), pp. 813–822.

[132] Christian Dahmann, John FX Diffley, and Kim A Nasmyth. "S-phase-promoting cyclin-dependent kinases prevent re-replication by inhibiting the transition of

replication origins to a pre-replicative state". In: *Current Biology* 5.11 (1995), pp. 1257–1269.

[133] Lidia C Boffa and Claudia Bolognesi. "Methylating agents: their target amino acids in nuclear proteins". In: *Carcinogenesis* 6.9 (1985), pp. 1399–1401.

[134] Reshma Shetty, Meagan Lizarazo, Randy Rettberg, and Thomas F Knight. "Assembly of BioBrick standard biological parts using three antibiotic assembly". In: *Methods Enzymol* 498 (2011), pp. 311–326.

[135] Alan L Goldstein and John H McCusker. "Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae". In: *Yeast* 15.14 (1999), pp. 1541–1553.

[136] Oksana M Subach, Illia S Gundorov, Masami Yoshimura, Fedor V Subach, Jinghang Zhang, David Grüenwald, Ekaterina A Souslova, Dmitriy M Chudakov, and Vladislav V Verkhusha. "Conversion of red fluorescent protein into a bright blue probe". In: *Cell Chemical Biology* 15.10 (2008), pp. 1116–1124.

[137] Nathan C Shaner, Robert E Campbell, Paul A Steinbach, Ben NG Giepmans, Amy E Palmer, and Roger Y Tsien. "Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein". In: *Nature biotechnology* 22.12 (2004), pp. 1567–1572.

[138] Anthony Röossl, Amanda Bentley-DeSousa, Yi-Chieh Tseng, Christine Nwosu, and Michael Downey. "Nicotinamide Suppresses the DNA Damage Sensitivity of Saccharomyces cerevisiae Independently of Sirtuin Deacetylases". In: *Genetics* 204.2 (2016), pp. 569–579.

[139] Ariane Galaup and Angelo Paci. "Pharmacology of dimethanesulfonate alkylating agents: busulfan and treosulfan". In: *Expert opinion on drug metabolism & toxicology* 9.3 (2013), pp. 333–347.

[140] David Owen Morgan. *The cell cycle: principles of control*. New Science Press, 2007.

[141] Zuzana Storchova and Christian Kuffer. “The consequences of tetraploidy and aneuploidy”. In: *Journal of cell science* 121.23 (2008), pp. 3859–3866.