MODELING AND PREDICTION OF GENETIC REDUNDANCY IN ARABIDOPSIS THALIANA AND SACCHAROMYCES CEREVISIAE

By

Siobhan Anne Cusack

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Cell and Molecular Biology—Doctor of Philosophy

2020

ABSTRACT

MODELING AND PREDICTION OF GENETIC REDUNDANCY IN ARABIDOPSIS THALIANA AND SACCHAROMYCES CEREVISIAE

By

Siobhan Anne Cusack

Genetic redundancy is a phenomenon where more than one gene encodes products that perform the same function. This frequently manifests experimentally as a single gene knockout mutant which does not demonstrate a phenotypic change compared to the wild type due to the presence of a paralogous gene performing the same function; a phenotype is only observed when one or more paralogs are knocked out in combination. This presents a challenge in a fundamental goal of genetics, linking genotypes to phenotypes, especially because it is difficult to determine *a priori* which gene pairs are redundant. Furthermore, while some factors that are associated with redundant genes have been identified, little is known about factors contributing to long-term maintenance of genetic redundancy. Here, we applied a machine learning approach to predict redundancy among benchmark redundant and nonredundant gene pairs in the model plant *Arabidopsis thaliana*. Predictions were validated using well-characterized redundant and nonredundant gene pairs. Additionally, we leveraged the availability of fitness and multi-omics data in the budding yeast *Saccharomyces cerevisiae* to build machine learning models for predicting genetic redundancy and related phenotypic outcomes (single and double mutant fitness) among paralogs, and to identify features important in generating these predictions. Collectively, our models of genetic redundancy provide quantitative assessments of how well existing data allow predictions of fitness and genetic redundancy, shed light on characteristics that may contribute to long-term maintenance of paralogs that are seemingly functionally

redundant, and will ultimately allow for more targeted generation of phenotypically informative mutants, advancing functional genomic studies.

ACKNOWLEDGEMENTS

First and foremost, I would like to give my heartfelt thanks to Shin-Han Shiu for taking a chance on me and accepting me into his lab. Without his guidance and mentorship I would not have succeeded in graduate school. I would also like to thank my guidance committee members: Rob Last and Kathy Osteryoung, for their advice and support, and Yair Shachar-Hill for invaluable belief in and advocacy for me throughout my graduate career. The Cell and Molecular Biology program has been a wonderful academic home and source of support, and for that I thank Sue Conrad, Peggy Petroff, and Alaina Burghardt. The Shiu Lab members I've had the pleasure of working with—Melissa Lehti-Shiu, Peipei Wang, Fanrui Meng, Paityn Donaldson, Sarah Horan, Thilanka Ranaweera and Serena Lotreck— made each day a delight to come to work and contributed to the friendly, supportive atmosphere of the Shiu Lab which I will miss very much. Thanks are especially due to Christina Azodi and Bethany Moore, for their willingness to lend a hand as I got the hang of Python and machine learning, and for their friendship, which made the more challenging aspects of grad student life much easier. I will always cherish our Starbucks trips. Finally, I'd like to thank Ronan, for teaching me how to function on less sleep than I ever thought possible, and Geoff, for the truly relentless encouragement and for sticking with me through all the ups and downs. Here's to the next 80 years together.

TABLE OF CONTENTS

LIST OF FIGURES

KEY TO ABBREVIATIONS

AUC-ROC	Area under the curve – receiver operating characteristic
AU-PRC	Area under the precision recall curve
DM	Double mutant
EN	Elastic net
GB	Gradient boosting
GO	Gene ontology
MAPK	Mitogen-activated protein kinase
ML	Machine learning
MYA	Million years ago
NR	Nonredundant
PCC	Pearson's correlation coefficient
PTM	Post-translational modification
RD	Redundancy definition
RF	Random Forest
SD	Standard deviation
SGD	Saccharomyces Genome Database
SM	Single mutant
SMF	Single mutant fitness
SVM	Support vector machines
SVR	Support vector regression
TAIR	The Arabidopsis Information Resource

WGD	Whole genome duplication
WT	Wild type

CHAPTER 1: INTRODUCTION: A HISTORY AND FUTURE OF GENETIC REDUNDANCY

1.1 How has the definition and understanding of genetic redundancy changed over time?

Conceptually, genetic redundancy was originally placed in the context of the origins of cellular life (Gabriel 1960), wherein the shift from free-floating organic molecules to protocells, capable of metabolizing, growing, and dividing into functional daughter protocells, would have been haphazard. Protocells would not have had carefully controlled systems to ensure even distribution of genetic material during division. It was therefore posited that polyploidy was likely a necessary pre-condition for cellular life, as it would increase the chances that a full set of genes would be passed on to each new protocell (**Figure 1.1**). As a side effect, this would result in what was termed "redundant genetic material" in protocells that ended up with multiple copies of each gene. It was further hypothesized that the development of increased fidelity in vertical gene transmission would allow duplicate gene copies to diverge and for some to begin to develop

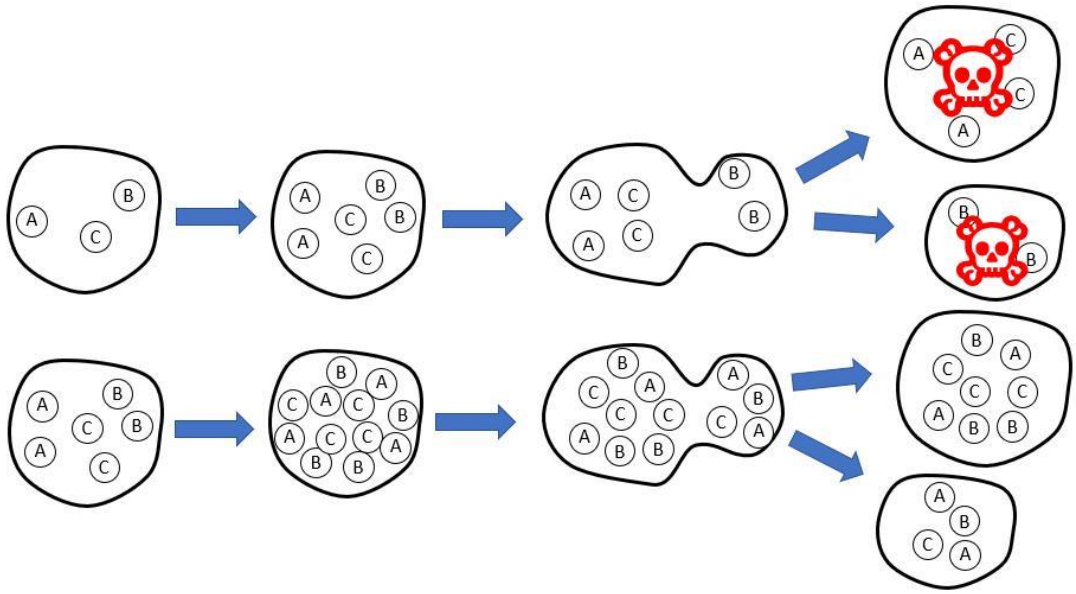

Figure 1.1: Polyploidy as an ancient strategy for survival after protocell division.
Illustration of polyploidy as an ancient strategy for survival after protocell division. The top row represents a haploid protocell, which does not produce viable daughter protocells after unequal distribution of genetic material in division. The bottom row represents a diploid protocell, which produces viable daughter protocells even after unequal division.

new functions.

From these hypothetical conceptions of genetic redundancy, there has been a considerable shift in how it is defined. Early research on genetic redundancy was primarily in the model organism *Saccharomyces cerevisiae* (baker's yeast), as it was a common model system used for genetics and biochemistry due to its genetic tractability and ease of growth (Botstein and Fink 1988). Some studies characterizing biochemical activity found that there were multiple structurally similar enzymes performing the same functions (e.g., Hunter and Markert 1957, Lacroute et al. 1965). In this context of biochemical activity, genetic redundancy referred to unlinked genes encoding enzymes capable of catalyzing the same reaction (Mortimer 1969). A similar definition placed genetic redundancy in terms of functional robustness in biochemical pathways due to interconnected metabolic networks (Weintraub 1993). In these definitions, there was no requirement for the genes themselves to be related, and thus could include non-paralogous genes derived from convergent evolution (Pickett and Meeks-Wagner 1995).

While many early studies on genetic redundancy focused on biochemical activity, some distinguished between genetic redundancy (caused by paralogous genes with identical protein products) and functional redundancy (caused by genes with different protein products that catalyze the same reaction) (Khan and Haynes 1972). This more genetics-based conception of genetic redundancy was later expanded to also include partially redundant paralogs that had evolved new functions since the duplication event from which they arose (Pickett and Meeks-Wagner 1995). While there are still nuanced disagreements about how "true" genetic redundancy presents experimentally (discussed below), recent studies primarily use this last definition, defining genetic redundancy as paralogous genes that maintain some of the same functionality (e.g., Kempin et al. 1995).

1.2 How is genetic redundancy relevant in genetics studies and how is it addressed?

Experimentally, genetic redundancy is commonly observed as a single gene knockout (referred to here as a single mutant) that shows no abnormal phenotype compared with a wild-type organism, while a double or higher-order mutant of that same gene does have an abnormal phenotype (**Figure 1.2**). Studies of organisms with genomes that have been mutated to the point of essential saturation in single mutant knockouts suggest that genetic redundancy is extremely common; a genome-wide single mutant knockout study in yeast found that only 15% of mutants had decreased growth under normal growth conditions (Giaever et al. 2002), while in *Arabidopsis thaliana* single mutants with a loss of function phenotype have been described for about 10% of the genes in the genome (Lloyd and Meinke 2012). Because the paradigm of reverse genetics (and to some extent, forward genetics) relies on observable mutant phenotypes

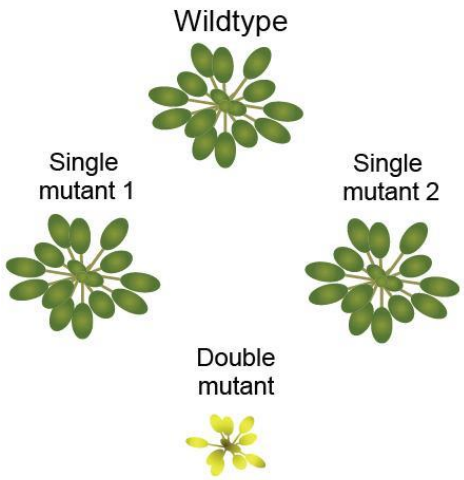

Figure 1.2: Single/double mutant phenotypes indicating genetic redundancy.
Simple example of how genetic redundancy may manifest in the context of a single-double mutant knockout study. This example depicts mutants of *Arabidopsis thaliana*. Neither single mutant has an abnormal phenotype compared to the wildtype, while the double mutant is noticeably smaller and chlorotic. The lack of abnormal phenotype in either single mutant indicates that the genes can compensate for one another, meaning they may have redundant functions.

to elucidate gene functions, this poses a fundamental problem in the pursuit of linking genotypes to phenotypes. Efforts have been undertaken in both systems to overcome this problem by generating large sets of double and higher-order mutants. The synthetic genetic array method, a way of systematically generating all possible combinations of double mutants (Tong 2001), has made this a relatively straightforward proposition in yeast, and allowed for exhaustive observation of single and double mutant phenotypes (Costanzo et al. 2016). In plants, it is simply not feasible with current technologies to generate all possible combinations of double mutants to exhaustively identify phenotypes among gene pairs that may be redundant. Nonetheless, large sets of double mutants have been generated in Arabidopsis to enable the study of genetic redundancy (e.g., Bolle et al. 2013, Su and Krysan 2016), which has generated valuable phenotypic information about single mutants and corresponding double mutants in paralogous gene pairs in plants.

As mentioned above, there is not universal agreement about the severity of mutant phenotypes necessary for two genes to be considered redundant. In the most striking examples of genetic redundancy, single mutants show no phenotype while the corresponding double mutant is lethal. Some well-studied examples of this in Arabidopsis include the mitogen activated protein kinase (MAPK) double mutant *mpk3/mpk6* (Wang et al. 2007) and the plasma membrane H^+ pump double mutant *aha1/aha2* (Haruta et al. 2010) which are lethal, while the respective single mutants show no phenotype. It is also possible for the double mutant to have a deleterious but nonlethal phenotype; for example, the MAPK double mutant *mpk9/mpk12,* which has altered response to abscisic acid (Jammes et al. 2009), and the MAPK kinase kinase mutant *anp1/anp3,* which has reduced growth (Krysan et al. 2002). These two situations, where the single mutants have no phenotype, are considered by some to be the only true type of genetic redundancy (e.g.,

Brookfield 1997). However, other definitions also include single mutants with phenotypes, as long as the double mutant phenotype is more severe. Examples of this include MAPK double mutants such as *mpk4/mpk11* (Kosetsu et al. 2010), *mpk4/mpk5*, and *mpk4/mpk13* (Su and Krysan 2016), where each single mutant has a dwarf phenotype and the double mutants have more severe dwarf phenotypes. Interestingly, the severity of the dwarf phenotype varies between the double mutants. This may be a result of unequal genetic redundancy, where a gene product A can compensate for gene product B more effectively than B can compensate for A. For example, two tryptophan synthase β subunit genes (TSB1 and TSB2) are expressed at very different levels, with TSB1 being expressed at levels approximately 15 times higher than TSB2 (Pruitt and Last 1993). As a result, a knockout for TSB1 shows a normal phenotype due to compensatory function of TSB2 under low light conditions, because the plants are growing relatively slowly. However, there is a severe chlorotic phenotype of this same mutant under high light conditions, because when plants are growing quickly the lower expression level of TSB2 is not sufficient to compensate for TSB1 (Last et al. 1991). Genes such as MPK4, which appears to show different degrees of redundancy with various members of the gene family, and TSB2, which has conditional compensation for its homolog, suggests that there are complicated evolutionary mechanisms acting on duplicate genes, which will be discussed more below.

1.3 How can genetic redundancy be maintained over evolutionary time?

The question of how genetic redundancy can be stably maintained is a puzzling one. A commonly held idea is that there is an evolutionary advantage to maintaining multiple copies of a gene; if one copy was compromised, the presence of the other would provide a phenotypic buffering effect (Zhang 2012). However, there are several issues with this hypothesis. One is that

genes with critical “housekeeping” functions are more likely to be singletons (DeSmet et al. 2013), which would limit the utility of genetic redundancy as an evolutionary failsafe. A second issue is that this hypothesis rests on the capacity of an organism to select for traits that are beneficial to potential future circumstances, which is contrary to our current understanding of selective pressure. Additionally, there is an energetic cost to maintain every gene in the genome (Lynch and Marinov 2015), and therefore an evolutionary force towards elimination of unnecessary material to minimize energetic losses. In fact, the outcome most commonly seen after gene duplication events in plants is pseudogenization (Panchy et al. 2016). Thus, while the idea of an organism retaining multiple copies of a gene makes some intuitive sense, genetic redundancy in fact represents an evolutionary paradox (Nowak et al. 1997).

In spite of this, the literature is replete with examples of genetic redundancy throughout the tree of life, from the relatively small genomes of many bacteria (Guckes et al. 2019) and the model animal *Caenorhabditis elegans* (Tischler et al. 2006) to the larger genomes of the model animal *Mus musculus* (Peters et al. 1999), monocots (Yao et al. 2008, Li et al. 2019) and dicots such as Arabidopsis (discussed above). It has been hypothesized that genetic redundancy is a temporary and unstable state which only occurs during the few million years after a duplication event, before there has been enough time for one copy to become pseudogenized. However, while this may be true in some cases, redundant genes in *S. cerevisiae* and *C. elegans* have been shown to originate from duplication events that happened over 600 million years ago (Vavouri et al. 2008), and Arabidopsis has genetically redundant paralogs from duplication events that happened over 50 million years ago (Blanc and Wolfe 2004; Maere et al. 2005), far longer than it would be expected to take for one of the genes to become pseudogenized or develop a new function.

Several other mechanisms have been proposed to explain how long-term stable maintenance is possible. In the case of paralogs with both overlapping and more recently evolved separate functions, both genes may be retained due to independent positive selection (Pickett and Meeks-Wagner 1995). Multiple copies of a gene may be retained if their presence provides a selective advantage through increased levels of the same gene product. For example, some bacteria with a relatively high copy number of rRNA genes can take advantage of nutrient-rich environments by growing more quickly than bacteria with fewer rRNA genes (Klappenbach et al. 2000); a higher copy number allows for comparatively rapid generation of ribosomes through parallel transcription and subsequently for quicker protein synthesis. A selective advantage may also be conferred by maintenance of redundant paralogs through the action of several similar gene products that act as functionally redundant checkpoints, for example in cell cycle progression, to ensure that important cellular processes are carried out properly (Thomas 1993). Mathematically modelling genetic redundancy showed that stable maintenance over time can occur if the efficacy of each duplicate at performing a given function (measured by the fitness of single mutants) is inversely correlated with the mutation rate in each gene (Nowak et al. 1997). Large-scale duplication events can result in retention of duplicate genes that encode subunits of multimeric protein complexes, as proper functioning of the complex is dependent on the stoichiometry of the subunits, which would be disrupted by the loss of one duplicate (Birchler and Veitia 2010). Finally, there is recent evidence of coordinated differential expression among some paralogs, such that knocking out one gene triggers compensatory upregulation of another (El-Brolosy et al. 2019; Ma et al. 2019), which would be expected to promote retention of both. Thus, there are many explanations as to how redundancy could be stably maintained even over

millions of years, but fewer answers as to which mechanisms generally allow for prolonged retention of genetic redundancy.

1.4 What questions about genetic redundancy remain and how can they be addressed?

Genetic redundancy is likely overestimated in the literature. Identification of redundant genes typically requires the discovery of a single mutant with no phenotype; however, there are many reasons a phenotype in a single mutant may not be readily apparent. For example, phenotypes may be present but conditional, subtle, or tissue-specific (discussed by Bouché and Bouchez 2001; Bolle et al. 2013). A more accurate assessment of the prevalence of genetic redundancy will require studies specifically designed to take these aspects into account. Furthermore, while genetic redundancy tends to be discussed as a binary trait (gene pairs are classified as redundant or nonredundant), or sometimes as a categorical trait (gene pairs are fully redundant, partially redundant, or nonredundant; e.g., Khan and Haynes 1972), phenotypic data from gene families suggest that the degree of redundancy exists over a range (as mentioned above with MPK4) which could therefore be calculated as a continuous trait based on fine-scale phenotypic data. As this is likely to be more biologically relevant than a classification scheme, a model of genetic redundancy should be built that allows for degrees of genetic redundancy on a continuum rather than a binary or categorical state.

Some of the factors associated with long-term maintenance of genetic redundancy in model organisms have been identified. For example, transcription factors are enriched among redundant gene pairs in Arabidopsis, suggesting that function of the gene product influences retention (Blanc and Wolfe 2004). We also know that the mechanism(s) of duplication play a role; for example, stress related genes derived from tandem duplication are retained at a higher

rate in Arabidopsis than those derived from other duplication mechanisms (Hanada et al. 2008). In yeast, redundant paralogs have a lower nonsynonymous substitution rate than nonredundant paralogs (Li et al. 2010). However, there are no obvious patterns that can be reliably used to determine *a priori* which gene pairs are likely to be redundant and therefore should be targeted in double mutant studies to generate phenotypically informative mutants.

To begin addressing the challenges presented above, the research in this dissertation focuses on two major points:

1) Building a binary (classification) model of genetic redundancy in Arabidopsis to predict redundancy among paralogous gene pairs using several different definitions of genetic redundancy, to ultimately aid in future targeted mutant generation.
2) Building a continuous model of genetic redundancy in yeast to identify features associated with redundancy, to begin answering evolutionary questions about its maintenance.

Together, these approaches contribute to understanding of the biological basis of genetic redundancy and provide a foundation for future modeling as more fine-scale data become available.

CHAPTER 2: GENOME-WIDE PREDICTIONS OF GENETIC REDUNDANCY IN *ARABIDOPSIS THALIANA* USING MACHINE LEARNING

2.1 Abstract

Genetic redundancy refers to a situation where an individual with a loss-of-function mutation in one gene (single mutant) does not show an apparent phenotype until one or more paralogs are also knocked out (double/higher-order mutant). Previous studies have identified some characteristics common among redundant gene pairs, but a predictive model of genetic redundancy incorporating a wide variety of features has not yet been established. In addition, the relative importance of these characteristics for genetic redundancy remains unclear. Here, we establish machine learning models for predicting whether a gene pair is likely redundant or not in the model plant *Arabidopsis thaliana*. Benchmark gene pairs were classified based on six feature categories: functional annotations, evolutionary conservation including duplication patterns and mechanisms, epigenetic marks, protein properties including post-translational modifications, gene expression, and gene network properties. The definition of redundancy, data transformations, feature subsets, and machine learning algorithms used affected model performance significantly. Among the most important features in predicting gene pairs as redundant were having a paralog(s) from recent duplication events, annotation as a transcription factor, downregulation during stress conditions, and having similar expression patterns under stress conditions. Predictions were then tested using phenotype data withheld from model building and validated using well-characterized, redundant and nonredundant gene pairs. This genetic redundancy model sheds light on characteristics that may contribute to long-term maintenance of paralogs that are seemingly functionally redundant, and will ultimately allow for more targeted generation of functionally informative double mutants, advancing functional genomic studies.

2.2 Introduction

Genetic redundancy, which refers to multiple genes that perform the same function, has been defined in many ways since the mid-1900s (Gabriel 1960). An early study of genetic redundancy in *Saccharomyces cerevisiae* discussed it in the context of unlinked genes encoding enzymes catalyzing the same reaction (Mortimer 1969). A later study took a broader view of genetic redundancy, with the degree of redundancy ranging from "complete redundancy" among genes with housekeeping functions to "partial overlap of function" among genes with primarily regulatory functions (Pickett and Meeks-Wagner 1995). In studies from a number of model organisms, multiple examples of what is considered genetic redundancy have been given, including: genes derived from convergent evolution encoding enzymes that perform the same function (Pickett and Meeks-Wagner 1995); biochemical pathways that are redundant due to interconnected metabolic networks (Weintraub 1993); and genes from the same family (paralogs) that maintain some of the same functionality (Kempin et al. 1995). Discussions of genetic redundancy in recent literature mostly encompass this last definition, where a duplication event results in multiple copies of a gene that retain overlapping functions (e.g., Chen et al. 2010, Bolle et al. 2013, Rutter et al. 2017). Practically, genetic redundancy is commonly observed as a single gene knockout mutant that shows no phenotype or a mild phenotype compared with a wild-type organism, with a double or higher-order mutant showing a more severe phenotype.

After a gene is duplicated, selection may be relaxed on each copy, allowing accumulation of mutations in one copy, which can lead to pseudogenization (Brookfield 1992); thus, the presence of genetically redundant paralogs long after the duplication event would seem to be an evolutionary paradox (Nowak et al. 1997). In spite of this, the literature is replete with examples of genetic redundancy, and many redundant genes in species such as *S. cerevisiae* and

Caenorhabditis elegans originated from duplication events that happened over 600 million years ago (Vavouri et al. 2008). At least two mechanisms may explain how this is possible. Redundant copies can be retained for a long time due to the slow pace of genetic drift in large populations. Based on a few key assumptions, it is estimated that a mutation deleterious to the function of a duplicate copy could take 0.75 to 5 million years to be fixed in *Arabidopsis thaliana* (Panchy et al. 2016). However, this cannot account for the apparent redundancy among paralogs from the most recent whole genome duplication (WGD) in the Arabidopsis lineage ~50 million years ago (Bowers et al. 2003). Another possibility is that genetic redundancy is selected for due to its ability to buffer the effect of a deleterious mutation in one paralog (Zhang 2012). The issue is that such a mechanism requires selection based on future needs, which is counter to our understanding of evolution. A mathematical model has been used to demonstrate that redundancy can be stably maintained over time (Nowak et al. 1997). However, the model requirement for perfect equivalency in gene functions and in mutations between paralogs seems unrealistic. Due to the challenges in assessing functions of paralogs, the extent of genetic redundancy and the factors contributing to it remain largely unclear.

Plants are an excellent resource for studying the fate of duplicated genes due to the relatively high rate of WGD events. While pseudogenization (loss of gene function) is the most common fate of duplicated genes in plants (Panchy et al. 2016), some duplicates are retained. By identifying and comparing characteristics of paralogous gene pairs and singleton genes, studies have revealed, for example, a lower synonymous substitution rate among retained (i.e., not pseudogenized) paralogs (Jiang et al. 2013), suggesting that these gene pairs are relatively recent duplicates or that there is selective pressure to retain the ancestral (or a more recently evolved) function. Retention bias is also seen for some gene functions. For example, paralogous

transcription factor and signaling genes are retained at a higher rate than DNA repair genes (Blanc and Wolfe 2004). Retention rates of paralogs also vary by duplication mechanism—retained tandem duplicates are more frequently involved in stress responses (Hanada et al. 2008), and genes involved in signaling processes are preferentially retained when derived from WGD rather than smaller duplication events (Maere et al. 2005). While these studies reveal some characteristics of genes that are retained after duplication, they do not directly address whether these retained paralogs maintain redundant functions. A landmark study in Arabidopsis addressed this question using machine learning to integrate 43 gene features related to sequence similarity and gene expression, and predicted that ~50% genes in the Arabidopsis genome have at least one redundant paralog (Chen et al. 2010). In this study, a gene whose single mutant showed no abnormal phenotype (or a mild phenotype) and its closest match in the genome based on sequence similarity were defined as a redundant pair. The most important features for predicting redundancy included differences in isoelectric point, molecular weight, and predicted protein domains between genes in a pair. While this pioneer study provided insights into the prevalence of genetic redundancy, redundancy was defined in only one way without considering the phenotypes of the corresponding double mutants. Also, in the decade since that study substantially more functional genomic data have become available; inclusion of these data in addition to sequence similarity and gene expression may improve the accuracy of redundancy predictions.

While the definition of redundancy presented above is prevalent, observation of unequal genetic redundancy, where the single mutant for one paralog shows a much more severe phenotype than the other and the double mutant has a still more severe phenotype (Briggs et al. 2006), promotes the idea that redundancy is more accurately conceptualized as a continuum.

However, the time-consuming nature of precise phenotyping required to quantify redundancy in this manner means that such data are available for relatively few paralogs, and discussions of genetic redundancy frequently exclude single mutants with severe phenotypes. Here we build upon previous work by modeling genetic redundancy using multiple definitions of redundancy by including single mutants in multiple phenotypic categories, and incorporating over 4,000 gene features from six categories, including functional annotations, evolutionary properties, protein sequence properties, gene expression patterns, epigenetic modifications, and network properties. We compared several machine learning algorithms and feature selection methods to identify which of the features have the most predictive power with respect to redundancy. Independent of the model, we additionally performed statistical analysis to identify features common among redundant gene pairs using nonredundant gene pairs as a contrast. To estimate the prevalence of genetic redundancy throughout the genome, we used two of the best-performing genetic redundancy definitions to predict whether ~18,000 gene pairs in the Arabidopsis genome are genetically redundant. Finally, to assess the accuracy of our model, we validated predictions using a "holdout" testing dataset and a handful of experimentally well-characterized gene pairs.

2.3 Results and Discussion

2.3.1 Definitions of genetic redundancy

The designation of a gene pair as genetically redundant requires phenotype data for double mutants and the corresponding single mutants. To define a set of benchmark redundant and nonredundant gene pairs, we used phenotype data for 2,400 single and 347 higher-order Arabidopsis mutants (including 271 double mutants) from a previous study (Lloyd and Meinke 2012) in which mutants were classified as having no phenotype, a less severe phenotype (i.e.,

conditional, cellular/biochemical, or morphological), or a severe phenotype (i.e., lethal, indicating the gene is essential) based on comparison with wild-type individuals. We assigned these categories phenotype class numbers: 0 (no phenotype), 1 (conditional), 2 (cellular/biochemical), 3 (morphological), and 4 (lethal) (**Figure 2.1A**) and applied this same phenotype classification to 29 additional gene pairs (Bolle et al. 2013), resulting in a final benchmark set of 300 single and double mutant trios (two single mutants and one corresponding double mutant). Note that our data are from experiments generally not designed to assess genetic redundancy and typically conducted in one or a limited number of conditions and environments. Thus, it more straightforward to identify an abnormal phenotype (indicative of nonredundancy) than to prove the absolute absence of an abnormal phenotype (indicative of redundancy).

Using the benchmark phenotype data and the core idea for defining genetic redundancy based on phenotype severity of single mutants only in comparison to the corresponding double mutant, we established nine redundancy definitions (RD1 to 9) (**Figure 2.1B**). The most inclusive definition of genetic redundancy was RD9. Under this definition, a gene pair was considered redundant if the phenotypes of both single mutants were less severe (i.e., assigned a lower phenotype class number) compared with that of the double mutant; 190 benchmark gene trios met this definition. The other RDs were subsets of RD9. Using these more stringent definitions, only mutants of a particular phenotype class were included in the benchmark dataset; for example, when using RD4 only single mutants of class 0 (no phenotype) and double mutants of class 4 (lethal) were considered. The use of multiple definitions offered insulation against errors due to the inherent challenges of classifying phenotypes into specific categories (e.g., some morphological phenotypes are much more severe than others; under specific conditions, conditional lethal is effectively the same as lethal). For example, while RD4 excluded

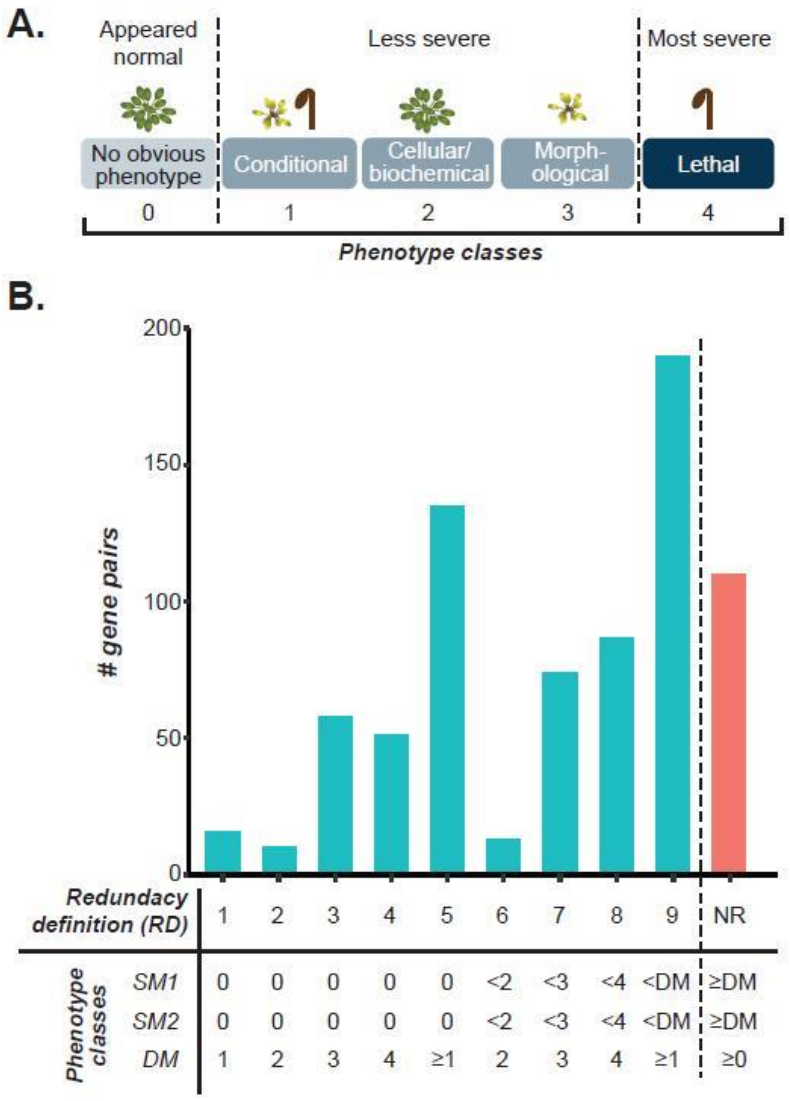

Figure 2.1: Arabidopsis phenotype categories and redundancy definitions.
(*A*) Phenotype classes from Lloyd and Meinke 2012. (*B*) Definitions of redundancy and nonredundancy (NR) based on phenotype classes of both single mutants (SM1 and SM2) and the double mutant (DM) for each gene pair. The number of gene pairs assigned to each definition is shown. The comparison signs (≥ and <) are in reference to phenotype category numbers; for example, a phenotype class of "<4" could be 0, 1, 2, or 3. RD5 is RD1-4 combined and RD9 is RD1-8 combined.

conditional lethal double mutants, as these belonged to phenotype class 1, both types of mutants were included in RD5 and RD9. While we acknowledge that this classification of phenotype severity has caveats, in the absence of quantitative phenotype data on a large scale, quantitative categories together with our multiple definitions of redundancy allow us to better utilize the dataset and begin addressing redundancy more as a continuum than as a binary problem.

To define nonredundant gene pairs, a single definition was used: two genes were considered nonredundant if the double mutant was in the same phenotype class as either single mutant or in a class with a lower number; that is, at least one single mutant had an equal or more severe phenotype than the double mutant (**Figure 2.1B**). The nonredundant set contained 110 gene trios. The nearly 2:1 ratio of redundant to nonredundant gene pairs may reflect a bias in the reporting of double mutant phenotypes; a positive result supporting the presence of a more severe phenotype in the double mutant would tend to be reported, with negative results less likely to appear in the literature. Because comparably fewer gene pairs for which the double mutant has no abnormal phenotype have been reported, nonredundant gene pairs very likely are less common in our dataset than they are in nature. Double mutants with much more dramatic phenotypes compared with the single mutants were also overrepresented in our dataset (**Figure 2.S1**), likely for similar reasons. As a result, some definitions that included only double mutants with mild or no phenotypes had too few gene pairs (RDs 1, 2, and 6, which had 16, 10, and 13 gene pairs, respectively) to generate robust models and were therefore excluded from further analyses.

2.3.2 Optimal parameters for prediction of genetic redundancy with machine learning

Machine learning allows integration of multiple data types to build a statistical model that

can predict a specific outcome. In our case, we were interested in establishing a machine learning model that could predict whether a gene pair was redundant or not using six broad categories of data: functional annotations, evolutionary properties, protein properties, gene expression patterns, epigenetic modifications, and network properties (**Table S1**). The general approach we took is illustrated in **Figure 2.2A**. Here the input for the model consisted of benchmark gene pairs (instances), classified as redundant or nonredundant (labels) according to our nine definitions, and information about the genes and gene pairs from the six categories of data (referred to as features). To alleviate the possibility of overfitting our model due to the large number of features examined (~4,000) compared with the number of instances (161-300 depending on the definition), we used 90% of the benchmark gene pairs (training set) to train a model for predicting if a gene pair is redundant or not in a 10-fold cross-validation scheme. Performance was measured using the Area Under the Curve-Receiver Operating Characteristic (AUC-ROC); higher scores indicate a higher true positive rate (proportion of all redundant gene pairs correctly predicted) over the range of false positive rates (proportion of gene pairs incorrectly predicted as redundant). Performance was additionally measured using the Area Under the Precision Recall Curve (AU-PRC); higher scores here indicate greater precision (proportion of gene pair predictions that are correct) over the range of true positive rates ("recall"). Because we used a binary classification scheme (redundant or not) for machine learning, a model classifying gene pairs at random would have a score of 0.5 for both the AUC-ROC and AU-PRC measures, while a perfect model would have a score of 1. Comparing three commonly used machine learning algorithms, Support Vector Machine (SVM), Random Forest, and Gradient Boosting, we found SVM was on average the best-performing algorithm when

using RD9 (**Figure 2.S2A-B**; ANOVA, p-value < 2 x 10^{-16}, and Tukey's Honestly Significant Difference test, p-values < 0.008).

We next explored how the number of features examined and feature value transformation affected model performance. While models using multiple features generally perform better than those based on single features, the presence of uninformative features can decrease model performance. We tested two methods, Random Forest and Elastic Net, for selecting the most informative features. We looked at the effect of transformation because transforming feature values (e.g., taking the square of values) can amplify small differences, allowing subtle patterns to be more readily identified. We transformed the features four different ways (log, square, reciprocal, and binned) and compared model performance using the original, non-transformed features; the best transformation for each feature (as determined by feature importance scores from the trained models); or multiple transformations of the same original feature. We tested 24 feature combinations (see **Table S2**) by asking how well the model based on each feature combination performed in predicting the RD9/nonredundant benchmark genes in cross-validation. We found that using 200 features selected with Random Forest, using the best transformations of each, led to the best performing model (AUC-ROC = 0.74, **Figure 2.S2C** and AU-PRC = 0.72, **Figure 2.S2D**), with a 15% and 18% improvement in performance over a model using all of the untransformed features (AUC-ROC = 0.64, **Figure 2.S2E** and AU-PRC = 0.61, **Figure** 2.**S2F**). The selected features included many that were different representations of the same, raw feature. For example, several features related to total synonymous substitution rate (*Ks*), namely maximum *Ks*, minimum *Ks*, average *Ks*, difference in *Ks*, and total (sum) *Ks* for genes in a pair (see **Methods**) were all among the features selected for RD9, demonstrating that

representing a characteristic such as *Ks* in a variety of ways provides distinct and useful information for building the model.

2.3.3 Comparison of models built with different redundancy definitions

We anticipated that the training sets established using some RDs would result in more accurate predictions than others. Therefore, we next identified the RD that resulted in the best predictions of redundancy using the optimal algorithm (SVM) and input feature set that we identified (200 features selected with Random Forest, using only the best transformation of each feature). When comparing how well each model performed on the cross-validation sets, the model built for RD4 (referred to as the RD4 model) had the best performance (AUC-ROC = 0.84, **Figure 2.2B**; AU-PRC = 0.82, **Figure 2.2C**; light blue lines). This RD had the highest contrast between single and double mutant phenotype classes (0—no apparent phenotype—and 4—lethal, respectively). A likely reason for the better performance of the RD4 model is that it was easier to build a model to distinguish between redundant and nonredundant gene pairs when the phenotype differences were the most extreme. The second-best models were the ones with the largest training sample sizes, RD5 and RD9 (yellow and green lines, respectively, **Figure 2.2B-C**). Thus, it appears that phenotype class contrast and sample size were the most important factors influencing model performance. We therefore focused on models built with the highest phenotype class contrast (RD4) and the largest sample sizes (RD5 and RD9) for further model building.

While the RD4 model performed the best in cross-validation, the majority of redundant gene pairs in the Arabidopsis genome do not have such a high phenotype class contrast. We

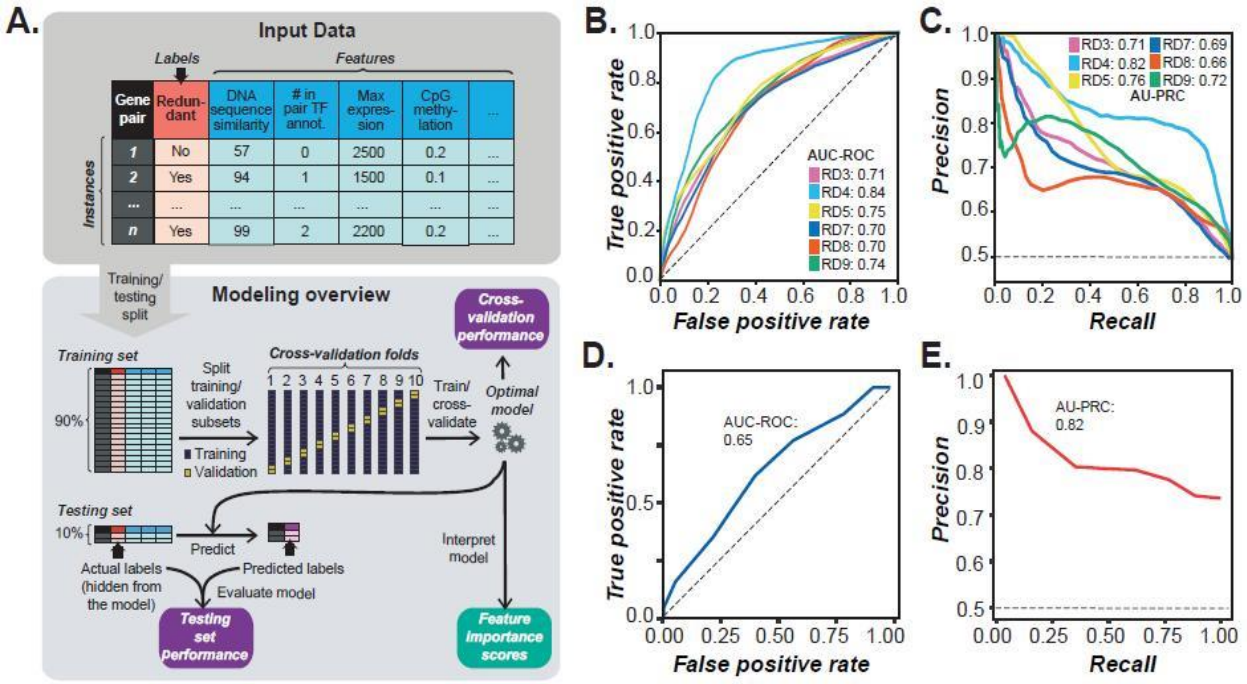

Figure 2.2: Machine learning pipeline workflow and performance of models built with different redundancy definitions.
(*A*) Machine learning pipeline workflow. Input data consisted of instances (gene pairs) with labels (redundant or nonredundant) and values of features (characteristics of gene pairs). Example features, as shown in the table, include DNA sequence similarity, the number of genes in a pair annotated as having transcription factor (TF) activity, maximum gene expression level, and the average level of CpG methylation among genes in the pair. The full input data are provided in Supplemental Data. Instances were first split into training and testing sets. The training set was further split into a training subset (90%) and validation subset (10%) in a 10-fold cross validation scheme. The optimal model after tuning the model parameters was used to provide performance metrics based on cross-validation, predict labels in the training set for model evaluation purposes, and to obtain feature importance scores. (*B*-*C*) Cross-validation performance of models built using six of nine RDs based on (*B*) Area Under the Curve - Receiver Operating Characteristic (AUC-ROC) and (*C*) Area Under the Precision-Recall Curve (AU-PRC). RD1, 2, and 6 were not included due to small training data sizes. A model classifying gene pairs perfectly would have AUC-ROC and AU-PRC scores of 1.0; black dotted lines represent the performance of a model classifying at random, in which AUC-ROC and AU-PRC scores would be 0.5 given that we used balanced data (i.e., equal number of redundant and nonredundant instances). (*D*) AUC-ROC and (*E*) AU-PRC for a model trained using RD4 gene pairs and half of the nonredundant pairs (randomly selected) then applied to RD9 gene pairs (excluding RD4) and nonredundant pairs that did not overlap with those used in training the RD4 model.

therefore tested whether the RD4 model would prove useful in predicting redundancy between gene pairs when there were less extreme phenotype differences between the single and double mutants. The RD4 model was applied to a test set composed of RD9 gene pairs (after removing RD4 pairs) and a random subset of half the nonredundant gene pairs. While the AUC-ROC was only 0.62 (**Figure 2.2D**), the high AU-PRC score (0.82, **Figure 2.2E**) indicated that, as expected from applying a model built with a more conservative definition of redundancy, this model errs on the side of having a higher number of false negatives rather than false positives. Similarly, the RD5 model was applied to a test set composed of RD9 gene pairs (after removing RD5 pairs) and a random subset of half the nonredundant gene pairs. The performance of this model was significantly worse (AUC-ROC = 0.57, **Figure 2.S2G**; AU-PRC = 0.59, **Figure 2.S2H**). Thus, the best-performing models for predicting redundancy among gene pairs with all types of phenotype contrasts were those trained on RD4 and RD9; therefore, these two models were used in the following analyses.

2.3.4 Important evolutionary features in predicting redundant and nonredundant gene pairs

Because the identification of features that are distinct between redundant and nonredundant gene pairs can provide insights about the biological underpinnings of redundancy, we next assessed whether the distribution of values for each feature among the six feature categories was significantly different between redundant and nonredundant gene pairs based on the RD4 and RD9 definitions (see **Materials and Methods**). For RD4 and RD9, evolutionary properties had the highest percentage of features statistically associated with redundancy (55% and 53% respectively, multiple testing-adjusted *p*-value (*q)* <0.05; **Figure 2.3A-B**), and these

features tended to be the most significantly correlated with redundancy (median *q*-value of significant features = 0.0003 and 0.004 respectively; **Figure 2.S3A-B**). Overall, a shared set of 159 features were significantly associated with redundancy in both RD4 and RD9, and there was a correlation between -log(*q*-values) for each feature in RD4 and RD9 (Spearman's rank ρ = 0.75, $p < 2.2 \times 10^{-16}$; **Figure 2.3C**). This suggested that some features may be significantly associated with redundancy regardless of definition. However, among the top 200 features selected for building the RD4 and RD9 models, we found that only 33% and 25%, respectively, were significantly associated with redundancy when considered individually (**Figure 2.S3C-D**), highlighting the utility of considering features jointly using machine learning.

We next looked into individual features that distinguished redundant gene pairs defined using RD4 and RD9 from nonredundant gene pairs using feature importance scores from the trained models (**Table S3)**. In this case, an importance score represents the degree to which an individual feature contributes to the separation of redundant from nonredundant gene pairs by the algorithm, with features with a higher importance score having a larger contribution. In total, 51 features were shared between the two models (**Table S3**) with well correlated importance ranks (PCC = 0.63, **Figure 2.S4A**), suggesting that a core set of features are important for predicting redundancy using multiple definitions. However, a shared set of 51 features leaves ~75% of the 200 features selected for each model as unique, highlighting the significant effect of redundancy definition on the models and the types of important features recovered.

The relative importance of the six feature categories ranked from best to worst based on median importance ranks for features in those categories in RD4/RD9-based models was as follows: functional annotations (32/17), evolutionary properties (63.5/81.5), network properties (123/81.5), gene expression patterns (110.5/101.5), epigenetic modifications (108/140), and

protein properties (139/133.5). Note that the importance ranks do not mirror the findings in **Figure 2.3A-B**, indicating that, for example, while the distributions of >50% of evolutionary property-based features significantly differed between redundant and nonredundant pairs, these features were not as important in predicting redundancy as functional annotation features. The most important feature in both the RD4 and the RD9 models, as determined by feature importance scores, was whether the gene pairs were duplicates from the α-WGD event (for the importance scores of the top 20 features, see **Figure 2.S4B-C**), with α-WGD-derived gene pairs more likely to be redundant (**Figure 2.3D**). The α event is the most recent WGD event in the Arabidopsis lineage, and despite it having likely occurred ~50 million years ago, the importance of this feature suggests that gene pairs derived from this event have not diverged in sequence and function sufficiently to appear nonredundant.

Two other evolutionary property features that were important for both definitions were reciprocal best match (rank=7 and 15 for RD4 and RD9, respectively, **Figure 2.S4B-C**) and a lethality score-derived feature (discussed below). Reciprocal best matches are paralogous gene pairs that do not have additional retained paralogs generated since their divergence; gene pairs that were reciprocal best matches were more likely to be redundant. As a pair of genes without more recent duplicates are themselves likely to be the product of a relatively recent duplication event (**Figure 2.S4D**), they are expected to have had less time to diverge in sequence and function, explaining their enrichment among redundant gene pairs. Consistent with this, *Ka* and *Ks*-related features ranked as high as 30 and 32 in the RD4 and RD9 models, respectively. Nonetheless, contrary to our expectations, these evolutionary rate-related features were not the most informative. Instead, other characteristics confounded with rates of evolution, such as

mechanism/mode of duplication and, as discussed in the following sections, gene functions and expression profiles, played more important roles in the model.

The reciprocal difference in lethality score was an important feature in both models (rank=2 and 9 for RD4 and 9, respectively, **Figure 2.S4B-C)**. Lethality score is the likelihood that mutation in a gene will lead to a lethal phenotype in Arabidopsis (Lloyd et al. 2015). Thus, we would expect that each gene in a redundant pair would have a low lethality score, and therefore a relatively small difference in lethality score for the gene pair. In contrast to our expectation, we found that redundant gene pairs generally had a smaller difference in reciprocal lethality scores (which equates to a larger difference in raw lethality score) compared with nonredundant gene pairs, although the difference was not significant (Wilcoxon test, q-value < 0.11). This unexpected result was likely an artifact of a bias in our data—lethality scores were predicted by Lloyd et al. (2015) for genes without known single mutant phenotypes, but 92% of the genes included in our benchmark dataset have known (nonlethal) phenotypes. In the absence of a predicted lethality score, we used a score of 0 for known nonlethal mutants, which likely artificially lowered the average lethality scores in our benchmark set. Nonetheless, the lethality scores still provided useful information, as indicated by the high importance ranking in both RD4 and RD9.

2.3.5 Important gene expression, functional, and network characteristics

Features related to gene expression made up the largest portion of features selected for RD4 and RD9 model building, with a total of 126 gene expression features selected for one or both models. The predicted directionality of four features varied between the two definitions, meaning that for a given feature, redundant gene pairs according to one RD had higher values

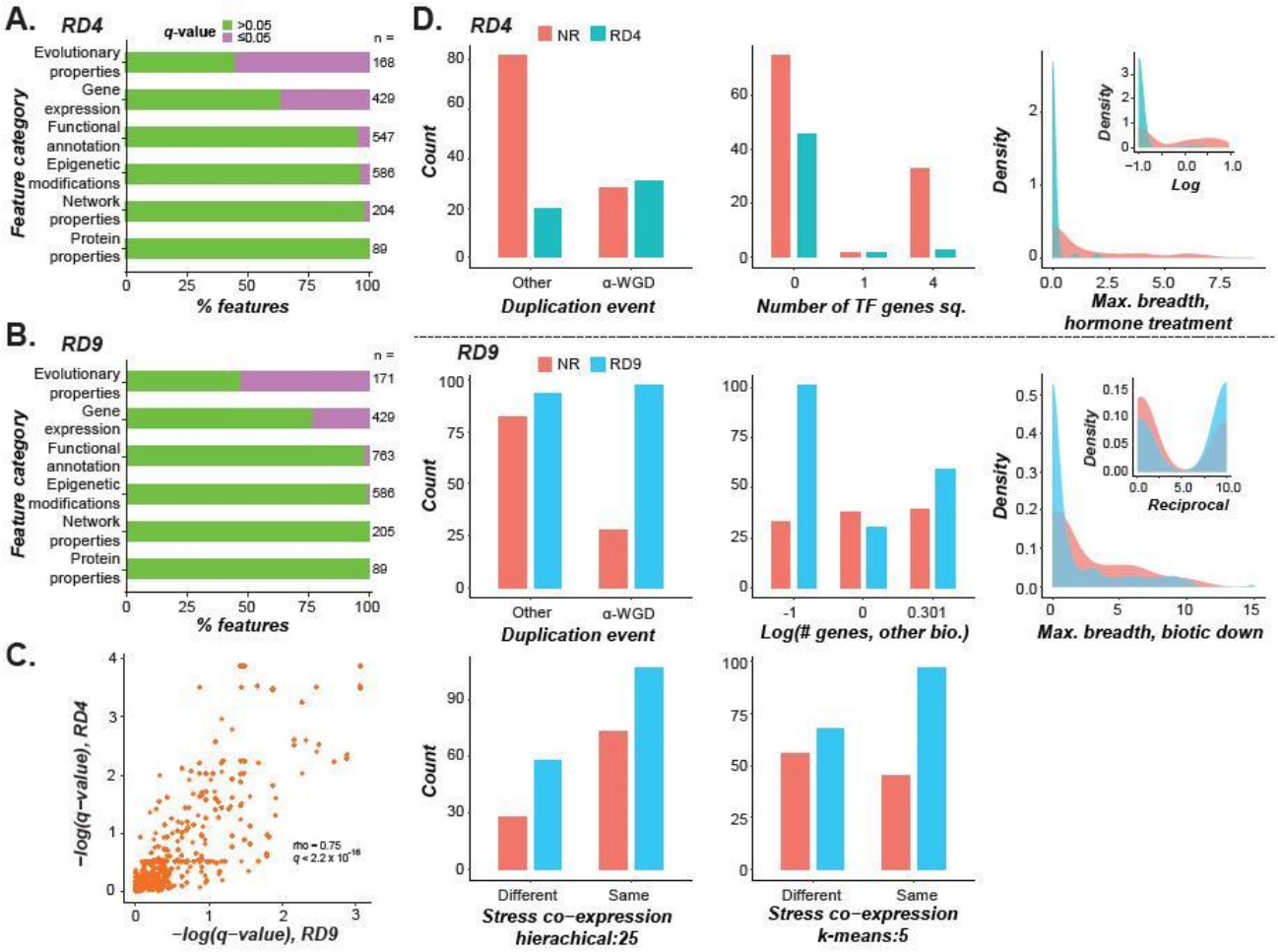

Figure 2.3: Features significantly associated with redundancy.
(A-B) Percentage of features in each feature category that were significantly associated with redundancy (Wilcoxon rank-sum test for continuous features; Fisher's exact test for binary features; all multiple-test corrected with Benjamini-Hochberg method) when using (A) RD4 and (B) RD9. (C) Correlation between RD4 and RD9 -log(q-values) obtained using the statistical tests as described in (A) and (B) for each feature. (D) Distribution of values among redundant and nonredundant gene pairs for selected features using RD4 and RD9 (separated by a dotted line). For each model, a feature is shown here if the importance score ranked between 1 and 20, was the highest in its feature category, and was significantly associated with redundancy using the statistical tests described in (A) and (B). For transformed continuous features, untransformed feature values are shown, with transformed values shown as inserts. Abbreviations: "Number of TF genes sq." is the square of the number of genes in the pair with the annotation DNA-dependent transcription factor; "Max. breadth, hormone treatment" is the maximum number of hormone treatments in which a gene in the pair is differentially expressed. "# genes, other bio." is the number of genes in a pair with the GO annotation "other biological function". "Max. breadth, biotic down" is the maximum number of genes in a pair downregulated under biotic

Figure 2.3 cont'd
stress. "Stress co-expression, hierarchical:25" and "Stress co-expression, k-means:5" refer to co-expression clusters generated from stress datasets with hierarchical (split into 25 clusters) and k-means (k=5) clustering, respectively; plots indicate whether the genes in a pair are in the same cluster or different clusters.

compared with nonredundant gene pairs, while the reverse was true for the other RD. For example, expression variation in the developmental expression dataset (after reciprocal average transformation) was higher for redundant gene pairs according to RD4 than for nonredundant gene pairs, but lower for redundant gene pairs according to RD9. We also found that tissue-specific stress responses varied by redundancy definition; the mean rank of features related to abiotic stress response for RD4 was lower for root tissue (97) than shoot tissue (120), while the opposite was true for RD9 (99 and 94, respectively). Features derived from biotic/abiotic stress and hormone treatment data were more consistently informative across definitions than those from the developmental dataset; while there were four developmental gene expression features in the top 30 for RD9, no such features ranked higher than 54 for RD4. The most important gene expression feature for RD9 was the maximum number of biotic stress conditions under which one or both genes in a pair was downregulated, with redundant gene pairs having a lower maximum than nonredundant gene pairs (**Figure 2.3D**). Thus, redundant gene pairs tend not to be downregulated under stress conditions. Previous findings indicated that duplicate genes involved in stress responses are retained at a higher rate than genes involved in other processes (Maere et al. 2005). The most important gene expression feature for RD4 was the maximum number of hormone treatments under which one or both genes in a pair was differentially expressed compared with the control, with nonredundant gene pairs having a higher maximum (**Figure 2.3D**).

Among 2,627 functional annotation features, 19 and 13 were among the top 200 for the RD4 and RD9 models, respectively. While only one of these features was selected for both models, given that functional enrichment among redundant gene pairs varies by RD (**Figure 2.S5**), it was expected that different functional annotation features would be important for predicting redundancy using different redundancy definitions. The most important gene function feature for the RD4 model was the number of genes in a pair (0, 1 or 2) annotated as DNA-dependent transcription factors (referred to as transcription factors). In the trained RD4 model, gene pairs in which both genes had this annotation were more frequently predicted as nonredundant, consistent with the feature value distributions (**Figure 2.3D**). This was somewhat unexpected as previous studies have shown that transcription factors are more likely to be retained after gene duplication than other types of genes (Blanc and Wolfe 2004). The most important functional annotation feature for RD9 was the number of genes in the pair having the annotation "other biological processes" (**Figure 2.3D**). This term, which encompasses a broad range of processes including responses to stressors or hormones, ion transport, circadian rhythm, aging, and cell growth, among many others, was an important predictor of nonredundant gene pairs.

Finally, while no network properties or protein properties were among the 20 most important features in predicting redundancy for RD4, two network properties were in the top 20 important features for RD9: presence in the same gene co-expression clusters (generated using biotic and abiotic stress datasets, two different clustering algorithms and two different numbers of clusters), with gene pairs in the same cluster more likely to be redundant (**Figure 2.3D**). Consistent with this, Chen et al. (2010) found that gene co-expression during pathogen infection was one of the most important features for predicting redundancy in Arabidopsis. In the earlier

study by Chen et al. (2010), isoelectric point, overlap in protein domain annotations, and sequence similarity were also among the features found to be important predictors of redundancy. While these features were included in our model building based on RD4 and RD9, they ranked between 26 and 166 depending on the redundancy definition (**Table S3**). Note that the redundancy definition used in the Chen et al. study was based on solely single mutant phenotypes; this likely accounts for the minimal overlap in features found to be important in predicting redundancy.

We also examined the potential causes of mis-predictions by comparing feature values between correctly and incorrectly predicted pairs, generating a score (see **Materials and Methods**) representing whether mis-predicted nonredundant pairs had feature values similar to RD9 pairs (**Figure 2.4A**). We identified several features for which incorrectly predicted nonredundant pairs had values more like RD9 gene pairs than correctly predicted nonredundant pairs, and that also had high feature importance scores, suggesting they may play a role in mis-predictions (**Figure 2.4B**). Additionally, in a principal components analysis of correctly and incorrectly predicted nonredundant pairs (**Figure 2.4C**), the top 24 features contributing to the first principal component were related to CpG methylation (**Table S4**), implicating it as a major contributor to mis-prediction.

2.3.6 Redundancy predictions for Arabidopsis gene pairs not in the benchmark dataset

With the predictive model of redundancy in place, we wanted to get an estimate of genetic redundancy broadly throughout the entire genome. For this analysis, we selected a subset of paralogous gene pairs: all of the WGD and tandem duplicate (TD) pairs in the Arabidopsis genome (7,764 total, collectively referred to as the WG/TD set; **Supplemental Data**). We also

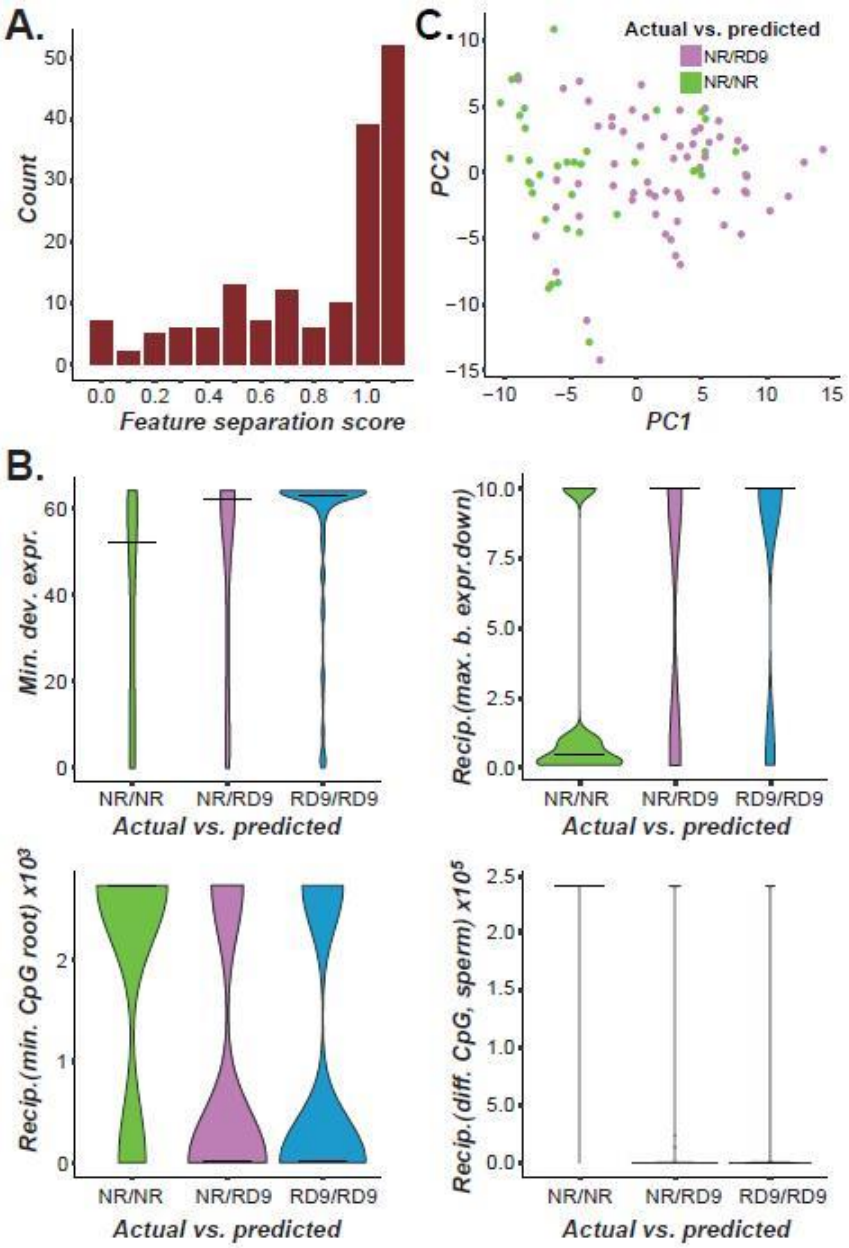

Figure 2.4: Features associated with mispredictions.
Distribution of feature separation scores for features used to build the RD9 model. To identify features that may contribute to mis-predictions, feature values were compared between (1) nonredundant gene pairs predicted as nonredundant (NR/NR), (2) nonredundant pairs predicted as redundant (NR/RD9), and (3) redundant pairs predicted as redundant (RD9/RD9). Using the median value (Med) in each class/predicted class category, we calculated a normalized feature separation score as follows: $(Med_{NR/RD9} - Med_{NR/NR}) / (Med_{RD9/RD9} - Med_{NR/NR})$. For each feature, the feature separation score represents the difference in feature values between correctly and incorrectly predicted nonredundant gene pairs, with a score of 0 meaning that correctly and incorrectly predicted pairs had similar values and a score of 1 meaning that

Figure 2.4 cont'd
incorrectly predicted pairs had values more similar to redundant gene pairs. Close to 20% of the features had a separation score of 1. (*B*) Distribution of values for selected features among the three categories of actual and predicted redundancy described in (*A*). Horizontal bars indicate the median. "Min. dev. expr." is the minimum number of tissues and developmental stages in which a gene in the pair is differentially expressed. "Recip. (max. b. expr. down)" is the reciprocal of the maximum number of biotic stress conditions in which one or both genes in the pair are downregulated. "Recip. (min. CpG root)" is the reciprocal of the minimum level of CpG methylation in root tissue for genes in the pair. "Recip. (diff. CpG sperm)" is the reciprocal of the difference in CpG methylation level in sperm cells for genes in the pair. These four features had a feature separation score close to 1 and had feature importance scores in the top 10 for RD9, implicating them in mis-predictions. (*C*) Dimensions 1 and 2 of a principal components analysis performed to identify features that were different between correctly and incorrectly predicted nonredundant pairs. The top 24 features contributing to Dimension 1 were related to CpG methylation levels (**Table S4**).

sought to model redundancy within a gene family; because a gene family consists of a group of paralogs derived from a variety of duplication mechanisms and with differing evolutionary distances, it offers a wide spectrum of relatedness among gene pairs. For this analysis, we used the protein kinase (Kin) superfamily to generate all possible combinations of gene pairs, then randomly selected 10,000 pairs for analysis (**Supplemental Data**). We expected that applying our model to both datasets would provide information about genetic redundancy at the genome-wide scale and at the more fine-grained gene family level. While both the RD4 and RD9 models showed a high degree of accuracy in predicting redundant gene pairs in cross-validation (87% and 92% of redundant gene pairs correctly predicted, respectively; **Figure 2.S6A-B**), the RD4 model predicted nonredundant gene pairs with much higher accuracy than the RD9 model (75% and 36%, respectively; **Figure 2.S6A-B**). Because of the high error rate in predicting nonredundant pairs with the RD9 model, we focused on using the RD4 model to estimate the prevalence of genetic redundancy in the Arabidopsis genome.

Although we analyzed machine learning results primarily as a binary variable (gene pairs were classified as either redundant or nonredundant), these binary predictions were generated from likelihood scores output by the machine learning pipeline. The likelihood score, referred to as a "redundancy score", ranges on a continuum from 0-1, with 0 being most likely nonredundant and 1 most likely redundant. Using this redundancy score, a threshold score was determined that would maximize the harmonic mean of precision (in this case, the proportion of true redundant pairs to predicted redundant pairs) and recall (proportion of redundant pairs predicted correctly), and this threshold was used to generate the binary predictions for the WG/TD and Kin datasets. Using the RD4 model, the majority of the 17,764 WG/TD and Kin gene pairs were predicted as redundant with redundancy scores well above the threshold (**Figure 2.5**). Among the WG/TD set as a whole, 80% were predicted as redundant (**Figure 2.5A**), with gene pairs derived from the α-WGD event more likely to be predicted as redundant (83%; **Figure 2.5B**) compared with those derived from the β-WGD event (71%; **Figure 2.5C**) and the γ and more ancient WGD events (73%, **Figure 2.5D**). As duplicate pairs evolve over time, it is expected that the degree of genetic redundancy would continue to decline. While this is true when comparing the α-WGD to older events, similar proportions of duplicate pairs from the β and more ancient events were predicted as redundant based on RD4. This may be because gene pairs derived from the more ancient γ-WGD look similar to those derived from the β-WGD in terms of *Ks* (Maere et al. 2005). However, it is surprising that so many redundant gene pairs (defined based on RD4) that duplicated 50 MYA (α-WGD), 80 MYA (β-WGD; Edger et al. 2015) or longer would be retained. Similarly, 83% of tandem duplicates and 87% of kinases were predicted as redundant based on RD4 (**Figure 2.5E** and **Figure 2.5F**, respectively).

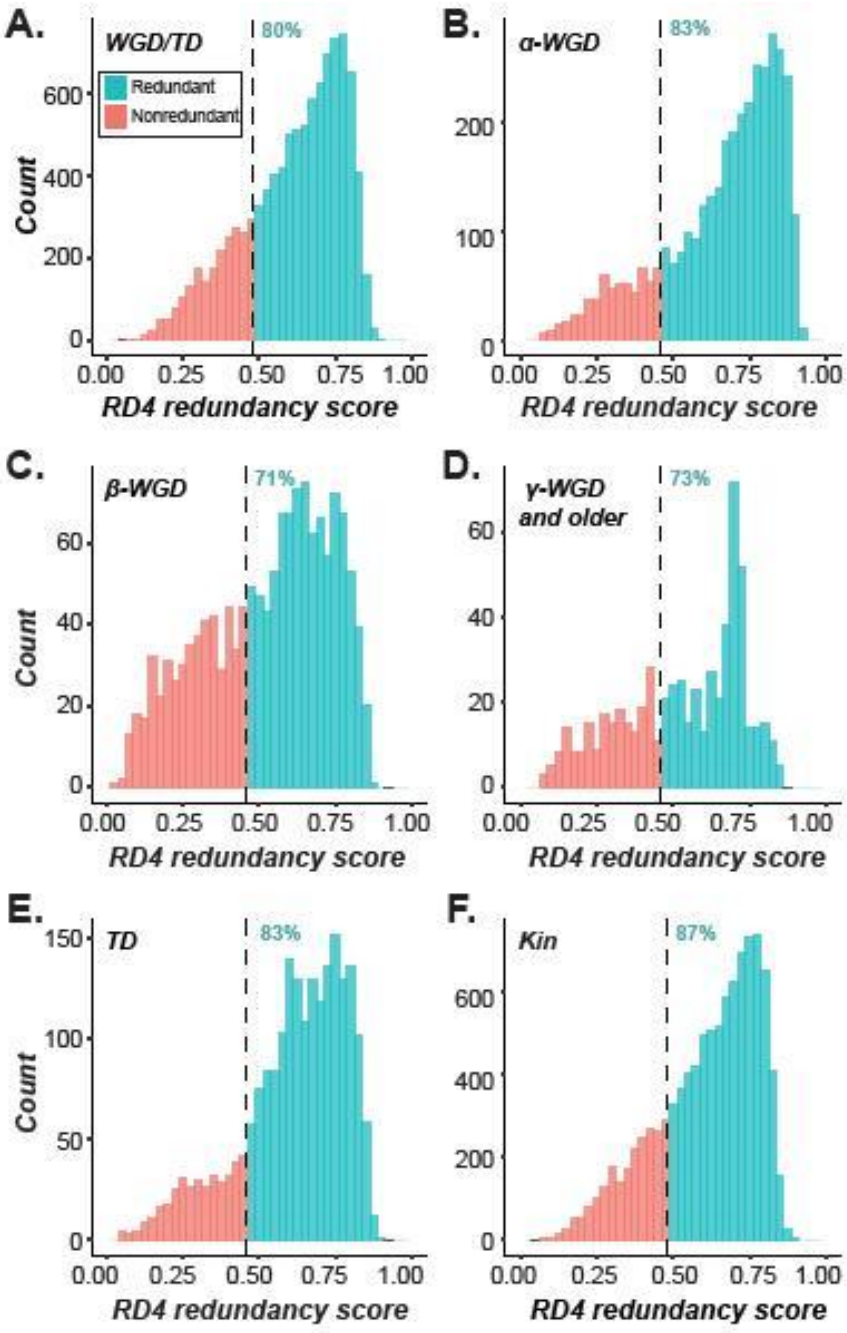

Figure 2.5: Predicted redundancy scores for gene pairs throughout the Arabidopsis genome.

(A) Predicted redundancy scores from the RD4 model for gene pairs in the genome derived from whole genome or tandem duplication (WGD and TD, respectively). These results grouped specifically by duplication event/type are shown in the following four sections of this figure. (B) Gene pairs derived from the α-WGD event, (C) gene pairs derived from the β-WGD event, (D) gene pairs derived from the γ-WGD event, (E) gene pairs derived from tandem duplication (TD),

Figure 2.5 cont'd
and (F) 10,000 randomly-selected gene pairs from the kinase superfamily (Kin). A majority of gene pairs in all of these datasets were predicted as redundant using RD4.

This percentage of redundant pair predictions was higher than previous estimates in the literature (e.g., Chen et al. 2010). It is important to note that in our WG/TD and Kin datasets, gene pairs are likely being predicted as redundant because they more closely resemble redundant gene pairs with respect to features that have the highest weight in our predictive model (e.g., WGD event). However, the model is built on experimental data that have much more power when calling a gene pair as nonredundant than calling them as redundant; demonstrating that a single mutant has an abnormal phenotype (meaning it is nonredundant) is a simpler task than definitively stating that a mutant has no abnormal phenotype and therefore is redundant with another gene. As previously proposed (Bouché and Bouchez 2001; Bolle et al. 2013), the lack of an observed severe phenotype in a single mutant may be because phenotypes are conditional, tissue-specific, and/or subtle rather than masked by genetic redundancy. Many large-scale phenotyping studies are not able to take these factors into account, and it would therefore be expected that a model built with data from such studies overestimate genetic redundancy in the genome.

While the binary classification of gene pairs as redundant or nonredundant was possible with the available data and straightforward to interpret, it is an over-simplification of the complex nature of genetic redundancy. The threshold-based definition of genetic redundancy may be convenient, but the landscape of genetic redundancy is far more nuanced, with a continuum between gene pairs with various degrees of genetic redundancy. Nonetheless, these data still allowed us to gain valuable insights into the mechanistic underpinnings of genetic

redundancy by revealing important features as discussed in the earlier sections. In addition, we anticipate the models can be iteratively improved with the future availability of more phenotype data, particularly quantitative data.

2.3.7 Validation of predictions

To validate predictions, we used a "holdout" testing set (10%, 16 and 30 pairs for RD4 and RD4, respectively, randomly selected and proportionally divided between redundant and nonredundant pairs, **Figure 2.2A**) of the benchmark data. This test set was not included in the model building process and serves to illustrate how the model will perform on new data. Applying the RD4 and the RD9 models on the test set, we obtained AUC-ROC scores of 0.73 and 0.68, respectively (**Figure 2.6A**) and AU-PRC scores of 0.62 and 0.82, respectively (**Figure 2.6B)**. Although there was a decrease in performance compared with cross-validation results (**Figure 2.2B-E**), 80% (4/5) and 68% (13/19) of redundant pairs were predicted correctly based on the RD4 and the RD9 models, respectively, and 36% (4/11) of nonredundant pairs were predicted correctly by each of these models (**Figure 2.6C-D**). Thus, the holdout testing set generally supported the utility of the RD4 and RD9 models, but the current threshold score was more conservative toward calling gene pairs as non-redundant.

Further validation was performed by identifying single and double mutants in the literature that have specifically been studied as mutant trios and have very well documented and characterized phenotypes. We selected ten of these gene pairs: five that meet our criteria for redundancy under RD9 and five we would classify as nonredundant (**Table S5**). Half of the pairs were present in our RD9 benchmark training dataset, while the other half were present in the

WG/TD and/or Kin test datasets. We compared the known redundant gene pairs from the literature to our predictions from cross-validation for pairs in the training set and predictions

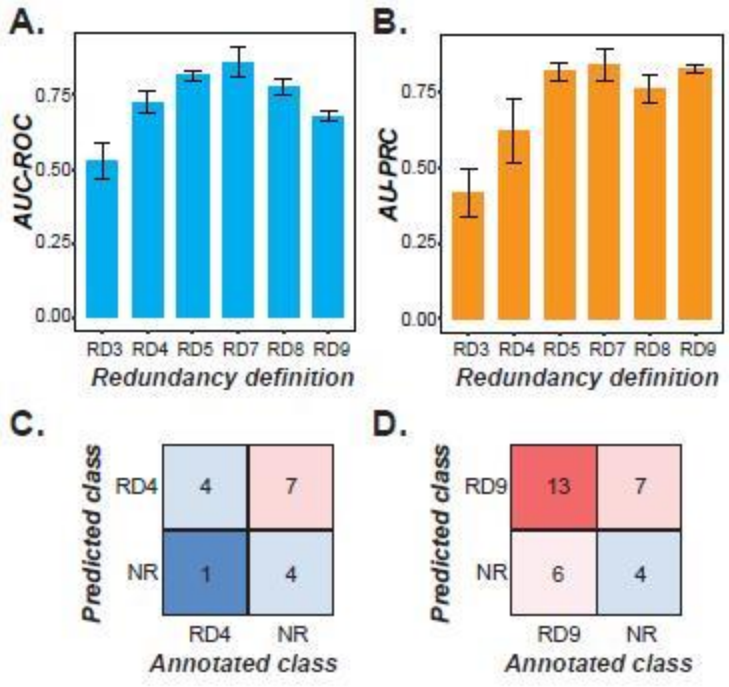

Figure 2.6: Model performance on holdout test sets.
(A) AUC-ROC and (B) AU-PRC curves for the holdout test sets for models built with each RD. Performance of the models on test sets was lower compared with performance in cross-validation, likely due to the small sample sizes of the test sets. (C-D) Confusion matrix for (C) RD4 and (D) RD9 showing the number of correctly and incorrectly predicted redundant and nonredundant gene pairs in the respective test sets. results showing that the RD9 model tends to err on the side of predicting false positives while the RD4 model is much more conservative and prone to generating false negative predictions.

from application of the trained model for pairs in the test set(s). We found that the RD9 model correctly predicted four of five redundant pairs (according to RD9) but mis-predicted all five of the nonredundant pairs as redundant. This comparison was repeated for the RD4 model with the same gene pairs. However, three of the gene pairs classified as redundant using RD9 were

classified as nonredundant using RD4 because the double mutants were not lethal. Thus, the validation set for RD4 included two redundant and eight nonredundant gene pairs. The RD4 model correctly predicted one out of the two redundant pairs (according to RD4) and four out of the eight nonredundant pairs. This was consistent with our expectations and prior results showing that the RD9 model tends to err on the side of predicting false positives while the RD4 model is much more conservative and prone to generating false negative predictions.

To determine why mis-predictions may have occurred in these specific cases, we revisited features previously identified as likely contributors to mis-prediction in general in the benchmark dataset (e.g., **Figure 2.4A-B**). For the RD9 model, one such feature was reciprocal best match. Although this feature was more strongly associated with nonredundant gene pairs in the benchmark dataset (**Figure 2.S7A**), the one RD9 pair predicted as nonredundant comprised paralogs that were not reciprocal best matches, making this a likely reason for mis-prediction. Derivation of paralogs from the α-WGD event was another such feature (**Figure 2.S7B**); three nonredundant pairs predicted as RD9 (nonredundant/RD9) were derived from the α-WGD event, indicating that this feature was a likely contributor to their mis-prediction. Another important feature was related to the number of biotic stress conditions under which genes were downregulated (referred to as biotic downregulation breadth). For this feature, the distribution of feature values among the actual/predicted classes demonstrated that all five nonredundant/RD9 pairs had values more similar to the correctly predicted RD9 pairs than to the correctly predicted nonredundant pairs (**Figure 2.S7C**). For the RD4 model, the one RD4 pair that was predicted as nonredundant had values for features related to CpG methylation (**Figure 2.S7D**), gene family size (**Figure 2.S7E**) and CHH methylation (**Figure 2.S7F**) that were more similar to those of nonredundant pairs. Additionally, all four of the nonredundant pairs predicted as RD4 had CHH

methylation in embryo tissue values that were more similar to those of RD4 gene pairs (**Figure 2.S7F**).

In total, we identified several types of features that were likely contributors to mispredictions, including duplication event (α-WGD or not), downregulation under biotic stress conditions, and gene methylation patterns. Importantly, we were thus able to identify one or more features that likely contributed to each instance of mis-prediction of both the RD4 and RD9 gene pairs used for validation, a important step in improving future iterations of the model; for example, depending on the definition being used and the importance of the accuracy of predictions (precision) compared with the importance of identifying all redundant gene pairs in a dataset (recall), certain features could be excluded from the model.

2.4 Conclusions

In this study, we optimized and utilized a machine learning approach to predict genetic redundancy among paralogs in Arabidopsis using multiple definitions of redundancy. We identified two biologically relevant and well-performing definitions of redundancy and the optimal 200 features for each definition that allowed us to best model redundancy. Our models performed well on a hold-out testing dataset, demonstrating their utility. Several features related to evolutionary properties, including lethality score, whether genes in a pair were reciprocal best matches, and the type of duplication event from which a gene pair was derived, were consistently ranked as important in generating predictions across redundancy definitions. Interestingly, evolutionary rates, such as *Ka* and *Ks*, were statistically different between redundant and nonredundant gene pairs but not highly ranked in the models, indicating that multiple factors contribute to redundancy, as revealed by machine learning models integrating multiple features.

Analysis of these evolutionary-related features demonstrated that redundant gene pairs tend to be more recent duplicates than nonredundant pairs. While it may be tempting to explain redundancy as gene pairs having not had enough time to diverge in function, many redundant pairs are derived from a WGD event estimated to have occurred ~50 million years ago, offering plenty of time for pseudogenization. This suggests that there is some selective pressure to maintain redundancy. In general, we found feature importance to be highly variable by redundancy definition, underscoring the need for testing multiple definitions depending on the biological question being addressed. For example, if one is interested in predicting which genes are lethal or have severe phenotypes a stricter definition is required than when a broader view of redundancy is being used, whereby less extreme phenotype contrasts between single and double mutants would be appropriate.

While the models provide useful information about gene features related to genetic redundancy, there is still room for improvement in terms of prediction accuracy. Performance on test gene pairs withheld from model building was generally not as good as the performance based on cross-validation, which may be due to the small size of the test sets. In addition, our more conservative trained model predicted 84% of 17,764 paralogs throughout the genome to be redundant, which is a much higher estimate than has been shown previously (Chen et al. 2010). This is likely a result of the underlying data used for model building; our models are expected to be biased towards categorization of gene pairs as redundant for the following reasons. We classified redundancy using phenotype data from the literature, including experiments that were not specifically designed to identify redundancy; there are expected to be substantial differences between experiments in how phenotypes were scored. For example, conditional or particularly subtle phenotypes may not have been examined. This likely results in misclassification of single

mutants as not having an abnormal phenotype. Because genetic redundancy was defined as a double mutant having more a severe phenotype than the corresponding single mutants, this bias will therefore lead to overestimation of genetic redundancy. Furthermore, classification of gene pairs as redundant or nonredundant, as we were able to do using the broad phenotype categories currently available on a large scale, overly simplifies a complex phenomenon. Redundancy as it exists in nature is not an all-or-nothing binary state, but rather a continuum with a wide range of biologically relevant states.

In our modeling exercise, redundancy scores derived from the model allow an approximation of this continuum, which can be further tested. One approach for testing the degree of genetic redundancy is by obtaining lifetime fitness data for single and double mutant sets. Because lifetime fitness in a mutant reflects the totality of phenotypic effects due to the introduced mutation over the entire life cycle of the individual, subtle and conditional phenotypes are likely better captured. Importantly, our current model can predict redundancy as defined by differences in some phenotypes under some specific conditions. It remains unclear the extent to which such model is relevant to predicting redundancy when it is defined based on single and double mutant fitness, the phenotypic outcome that has the most bearing on the evolutionary fate of a gene pair. Thus, in future studies the generation of lifetime fitness data would allow for a machine learning regression model that more accurately predicts degrees of genetic redundancy between genes in a pair rather than simply classifying genes as redundant or not. Such a model could be applied to gene pairs within a large gene family to compare predicted redundancy scores and reveal patterns related to redundancy maintenance and loss through evolutionary time. Analysis of features important for building the model would be expected to yield additional useful insights about mechanisms related to the evolutionary fate of gene

duplicates and the long-term retention of genetic redundancy. Taken together, our results demonstrate the utility of machine learning in combining features to generate accurate predictions of genetic redundancy and identify several evolutionary features that are important in predicting genetic redundancy across several definitions.

2.5 Materials and Methods

2.5.1 Definitions of redundant and nonredundant gene pairs

Arabidopsis mutant phenotype data were collected from Lloyd and Meinke (2012) and Bolle et al. (2013). Our benchmark dataset comprised gene trios for which a double mutant phenotype and both corresponding single mutant phenotypes were reported, with a total of 300 gene trios. A numeric phenotype severity value was assigned to each single and double mutant (**Figure 2.1A)**, with 0 representing no phenotype; 1, a conditional phenotype of any kind; 2, a cell or biochemical phenotype; 3, a morphological phenotype; and 4, a lethal phenotype. Redundancy was classified using nine definitions (RDs) of varying stringency (**Figure 2.1B**). The least stringent definition was RD9, in which any gene pair for which the double mutant phenotype severity score was higher than that of both the single mutants was defined as redundant. With this definition, the dataset contained 190 redundant gene pairs. Gene pairs were classified as nonredundant if at least one single mutant had a phenotype severity score greater than or equal to the double mutant score; the dataset contained 110 nonredundant gene pairs.

2.5.2 Feature value generation

For predictive modeling, data from six general categories were collected for each gene: functional annotations such as GO terms; evolutionary properties such as synonymous

substitution rate; protein sequence properties such as posttranslational modifications; gene expression patterns; epigenetic modifications such as histone methylation; and network properties such as gene interactions (**Table S1**). These data were processed to generate feature values for each gene pair (**Supplemental Data**), and the method used for processing depended on the data type: binary (e.g., whether or not a gene had a given protein domain), categorical (e.g., all the names of protein domains present in a given gene product) and continuous (e.g., gene expression level).

Features such as protein domain and functional annotations were treated as binary and/or categorical input data for feature generation. For processing as binary input data, each gene was assigned a score of 0 (does not have the annotation/property) or 1 (has the annotation/property); gene pair feature values were then generated by taking the number of genes in the pair (0, 1, or 2) having that annotation or property. For example, if Gene1 was annotated as having DNA binding activity but Gene2 was not, the feature value for DNA binding activity for that gene pair would be 1. Additional features were generated by taking the square, $-\log_{10}$, and reciprocal value of features processed in this way. For processing as categorical input data, all annotations of a specific type (e.g., GOslim terms) were listed for each gene. These were then used to represent similarity between genes in a pair. For example, if Gene1 had functional annotations of "DNA binding activity" and "signal transduction" and Gene2 had functional annotations of "signal transduction" and "protein binding", the number of overlapping annotations would be 1, the total number of unique annotations between the gene pair would be 3, and the percent overlap would be 33. For continuous data, gene pair feature values were generated by calculating the difference, average, maximum, minimum, and total of the values for the gene pair. For example, if Gene1 had an isoelectric point of 10 and Gene2 had an isoelectric point of 9, the difference would be 1,

the average 9.5, the maximum 10, the minimum 9, and the total value would be 19. Additional features were generated by taking the square, $\log_{10}$, and reciprocal of features processed as categorical and continuous data, and by assigning each value to one of four quartile bins generated from the untransformed feature data.

2.5.3 Functional annotation and evolutionary property features

Functional annotations included GO biological process, molecular function and cellular component annotations (The Gene Ontology Consortium et al. 2000; The Gene Ontology Consortium 2017), metabolic pathway annotations from AraCyc v.15 (Mueller et al. 2003), and predicted protein domain annotations from Pfam (Finn et al. 2016). These annotations were processed as binary and categorical data as described above. There were 2,627 features related to functional annotations after transformations were applied (**Table S1** and **Supplemental Data**).

Broadly, evolutionary properties included duplication mechanism and timing, and relationship to other genes in the genome. There were 171 features related to evolutionary properties after transformations were applied (**Table S1** and **Supplemental Data**).

To get the evolutionary rate for each gene in a pair, protein sequences (collected from NCBI; Pruitt et al. 2007) of each *A. thaliana* gene pair were searched against protein sequences from *Theobroma cacao*, *Populus trichocarpa*, *Glycine max* and *Solanum lycopersicum*, using the Basic Local Alignment Search Tool for protein sequences (BLASTP; Altschul et al. 1990). Protein sequences of the gene pair and the best hits in these four species were first aligned using MUSCLE (Edgar 2004), and then were compared to their coding nucleotide sequences to generate the corresponding coding sequence (CDS) alignment. CDS alignments were used to build gene trees using RAxML/8.0.6 (Stamatakis 2014) with parameters: -f a -x 12345 -p 12345

-# 1000 -m PROTGAMMAJTT. *Ka*, *Ks* and the *Ka*/*Ks* ratio on branches leading to each gene of a gene pair were calculated using the free-ratio model of the codeml program in PAML v. 4.9d (Yang 2007). Gene family size and lethality scores were obtained from Lloyd et al. (2015). Where lethality scores were not available, a score of 0 was assigned to known nonlethal genes and 1 was assigned to known lethal genes. Nucleotide and amino acid sequence similarity were calculated using EMBOSS Needle (McWilliam et al. 2013). *Ka*, *Ks*, *Ka*/*Ks*, gene family size, functional likelihood, lethality scores, and sequence similarity were processed as continuous data

Gene pairs were determined to have been derived from one of four types of gene duplication events using MCScanX-transposed (Wang et al. 2013): 1) segmental duplicates—paralogs located in corresponding intra-species collinear blocks; 2) tandem duplicates—paralogs next to each other; 3) proximal duplicates—paralogs close to each other, but separated by ≤ 10 non-homologous genes; 4) transposed duplicates—one of the paralogs located in inter-species collinear blocks, the other not. Segmental duplicates were additionally noted as being derived or not derived from the α- or β-WGD events. Protein sequences of *A. thaliana* were searched against protein sequences of *A. thaliana* (intra-species), *Arabidopsis lyrata*, *Brassica rapa*, *Carica papaya*, *P. trichocarpa*, and *Vitis vinifera* (inter-species) using BLASTP, with a cutoff E-value of $1x10^{-10}$. Five different sets of parameters were evaluated for MCScanX-transposed: 1) -k 50 -s 5 -m 25; 2) -k 50 -s 2 -m 25; 3) -k 25 -s 2 -m 25; 4) -k 25 -s 2 -m 50; 5) -k 25 -s 5 -m 25; where -k indicates the cutoff score of collinear blocks, -s specifies the number of matched genes required for the calling of a collinear block, and -m means the maximum number of genes allowed for the gap between two genes. The duplication mechanisms inferred using these five different sets of parameters were consistent with one another for the majority of gene pairs; 78 pairs had discrepant results, representing 0.4% of the total dataset. In these cases, the mechanism

that occurred most frequently in the results for that gene pair was assigned; if there was no majority, the mechanism was listed as N/A. Each gene pair was assigned a binary value indicating whether or not the genes were reciprocal best matches (i.e., they were one another's best hit based on nucleotide BLAST searches) and whether or not they were derived from each type of duplication mechanism (e.g., a gene pair derived from the α-WGD event would have a value of 1 for the WGD feature and for the α-WGD feature, and a value of 0 for all other duplication mechanisms).

Retention rate was based on the presence or absence of a paralog in 15 species: *A. lyrata*, *Capsella rubella*, *B. rapa*, *T. cacao*, *P. trichocarpa*, *Medicago truncatula*, *V. vinifera*, *S. lycopersicum*, *Aquilegia coerulea*, *Oryza sativa*, *Amborella trichopoda*, *Picea abies*, *Selaginella moellendorffii*, *Physcomitrella patens*, and *Marchantia polymorpha*. The retention rate for each gene was calculated as the number of genomes in which a paralog was present divided by the total number of genomes analyzed (16: *A. thaliana* plus the 15 additional species). Genome data were collected from Phytozome (Goodstein et al. 2012) for *P. patens* 318 v3.3, *M. polymorpha* 320 v3.1, *S. moellendorffii* 91 v1.0, *A. trichopoda* 291 v1.0, *O. sativa* 323 v7.0, *B. rapa* 277 v1.3, *C. rubella* 183 v1.0, *A. thaliana* 167 TAIR10, *A. lyrata* v2.1, *M. truncatula* 285 Mt4.0 v1, *V. vinifera* 145 Genoscope 12x, *A. coerulea* v3.1, *P. trichocarpa* 210 v3.0, and *T. cacao* 233 v1.1; from NCBI for *S. lycopersicum* v2.5; and from PlantGenIE (Sundell et al. 2015) for *P. abies* v1.0.

2.5.4 Gene expression and epigenetic modification features

Processed microarray gene expression datasets were obtained from Moore et al. (2019) and contained gene expression levels under biotic (Wilson et al. 2012) and abiotic stress (Kilian

et al. 2007; Wilson et al. 2012), under hormone treatment (Goda et al. 2008), at different developmental stages (Schmid et al. 2005), and at different times of day (Mockler et al. 2007). In addition to these gene expression levels, we also considered expression breadth, which represents the number of tissues and conditions under which each gene is expressed. Gene expression levels and ribosome occupancy from RNA-seq and Ribo-Seq experiments in root tissue were obtained from Hsu et al. (2016) and processed along with the microarray gene expression data as continuous data. There were 450 features related to gene expression after transformations were applied (**Table S1** and **Supplemental Data**).

Epigenetic modifications included DNA methylation, chromatin accessibility, and histone modifications. Percent CHH, CHG, and CpG methylation, gene body methylation, and histone modification data were obtained from Lloyd et al. (2015). Percent methylation values were treated as continuous data, and gene body methylation and histone modification data as binary data. Chromatin accessibility data were from Sullivan et al. (2014) and were also binary, with each gene receiving a score of 1 if it contained a DNase peak site and a score of 0 if it did not. There were 565 features related to epigenetic modifications after transformations were applied (**Table S1** and **Supplemental Data**).

2.5.5 Protein sequence and network property features

Protein sequence properties included amino acid length, isoelectric point, and posttranslational modifications. Amino acid lengths were obtained from Lloyd et al. (2015). Isoelectric points and myristoylation data were from The Arabidopsis Information Resource (Berardini et al. 2015). Amino acid length and isoelectric point were processed as continuous data. Acetylation, deamination, formylation, hydroxylation, oxidation, and propionylation data

were obtained from The Plant Proteome Database (Sun et al. 2009). Posttranslational modifications were processed as binary data: whether or not the protein product was predicted or known to have the modification. In total, 93 features were related to protein sequence properties after transformations were applied (**Table S1** and **Supplemental Data**).

Network properties were related to known or potential interactions of genes or protein products. Gene interaction data (AraNet v.1, Lee et al. 2010) and protein-protein interactions (AtPIN, Brandão et al. 2009) were processed as categorical data. Gene co-expression was calculated from the microarray datasets referenced above using multiple clustering algorithms, namely k-means, c-means and hierarchical clustering at k=5, 10, 25, 50, 100, 200, 300, 400, 500, 1000, and 2000 as described in Moore et al. (2019). These data were processed as categorical data, with each combination of clustering algorithm, dataset and k-value included as a feature; a gene pair received a value of 1 if both genes were in the same cluster and a value of 0 if they were not. There were 205 features related to network properties after transformations were applied (**Table S1** and **Supplemental Data**).

2.5.6 Identification of features distinguishing redundant and nonredundant pairs

To identify features that could distinguish between gene pairs from the redundant and nonredundant classes, we applied statistical tests to determine if feature values were significantly different between the classes. Binary gene pair features (e.g., duplication type, presence in a gene co-expression cluster) were analyzed using two-sided Fisher's exact tests with multiple testing correction using the Benjamini-Hochberg method (Benjamini and Hochberg 1995). To determine whether feature value transformations improved the ability to distinguish between classes, the reciprocal, square, and $\log_{10}$ of continuous features were included as separate features.

Continuous values were also binned into four quartiles of equal size and bin values included as features. Transformed and untransformed continuous feature values were analyzed using a Wilcoxon rank sum test (Wilcoxon 1945) with multiple testing correction performed using the Benjamini-Hochberg method. Features were considered to distinguish between redundant and nonredundant gene pairs if $q < 0.05$ after multiple testing correction (**Table S1**). Continuous feature effect sizes are the standardized z statistic (calculated from the *p*-values given by the Wilcoxon rank sum test) divided by the square root of the sample size. Binary feature effect sizes correspond to the odds ratio calculated from the enrichment table for each feature.

2.5.7 Redundancy prediction model building and optimization with machine learning

Models for predicting genetic redundancy between gene pairs were built with Random Forest, Gradient Boosting and SVM algorithms implemented in the scikit-learn machine learning package (Pedregosa et al. 2011) in Python. For Random Forest and Gradient Boosting, a grid search was performed with 10-fold cross-validation for parameter optimization; gene pairs were randomly divided into a training set (90%) and a testing set (10%) with proportional division of redundant and nonredundant gene pairs. This division was repeated 10 times, with 10 replicate models built for each iteration, for a total of 100 models. Ten-fold cross-validation was also used in model building with 100 iterations each for a total of 1000 models, with each being tested on a "holdout" testing dataset, consisting of 10% of the benchmark dataset and not included in building the model, to assess performance. The trained model was used to predict redundancy among all tandem and WGD pairs in Arabidopsis (**Supplemental Data**) and among a random sample of Arabidopsis kinase gene pairs. Using kinase family classifications from Lehti-Shiu

and Shiu (2012), all possible within-family combinations of gene pairs were generated. Ten thousand of these pairs were then randomly selected for predictions (**Supplemental Data**).

APPENDIX

Supplemental Information

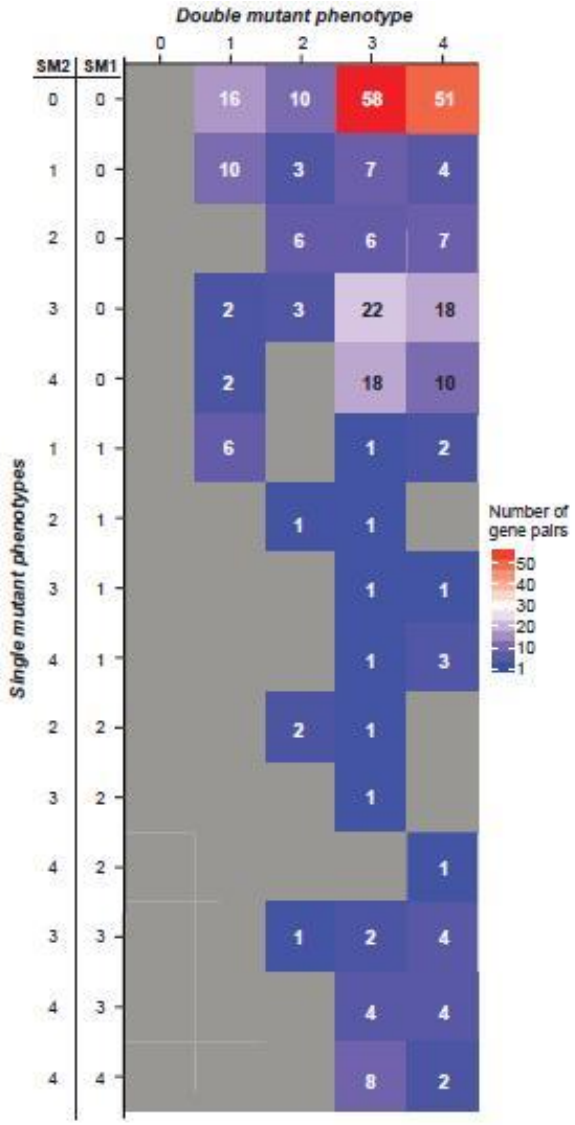

Figure 2.S1: Benchmark gene pair phenotypes.
Distribution of benchmark gene pairs among phenotype severity categories (as defined in Figure 2.1) for both mutants (SM1 and SM2) and the double mutant for each pair. The dataset is biased toward double mutants with more severe phenotypes.

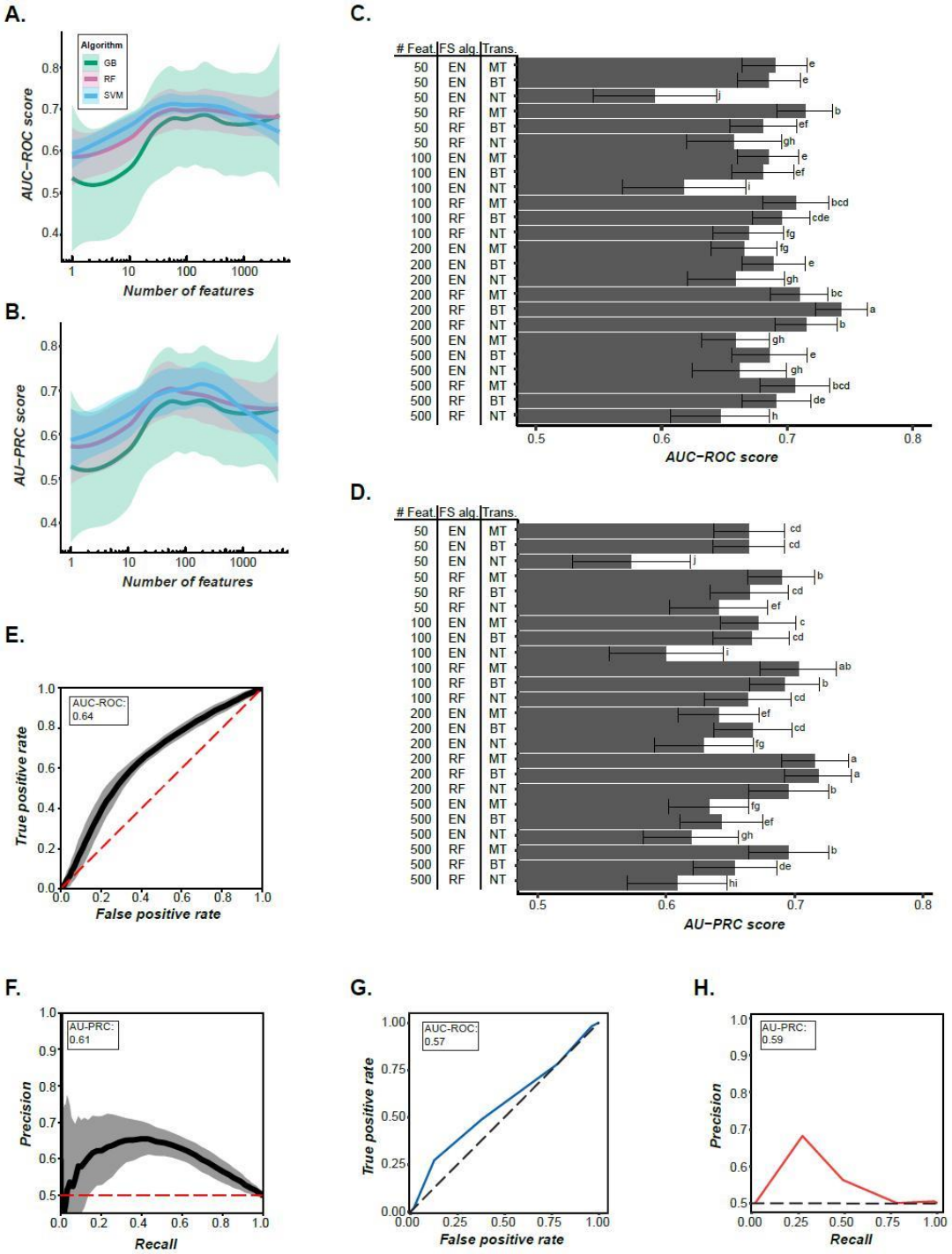

Figure 2.S2: Comparison of models built with different algorithms and numbers of features.

Figure 2.S2 cont'd
(*A*) AUC-ROC scores and (*B*) AU-PRC scores for binary classification machine learning models built using RD9 with Gradient Boosting (GB), Random Forest (RF) and Support Vector Machine (SVM) algorithms and using different numbers of features. Shading indicates the standard deviation. Using AUC-ROC as a measure, models built with SVM performed the best (ANOVA, p-value < 2×10^{-16}, and Tukey's Honestly Significant Difference [HSD] test, p-values < 0.008). Using AU-PRC, models built with SVM performed significantly better compared with those built with Gradient Boosting (ANOVA, $p < 2 \times 10^{-16}$; Tukey's HSD, p-value < 0.0001), but not with those built with Random Forest (Tukey's HSD, p-value = 0.36). (*C*) AUC-ROC scores and (*D*) AU-PRC scores for machine learning models built using different combinations of feature numbers ("# Feat."), feature selection algorithms ("FS alg."), and numbers of transformations allowed for each feature ("Trans."). Different letters indicate statistically significant differences between models according to Tukey's HSD. Using AUC-ROC as a measure, the best-performing combination was 200 features selected with Random Forest and with only the best transformation of each feature allowed (ANOVA, $p < 2 \times 10^{-16}$; Tukey's HSD, p-values < 0.0001). Using AU-PRC as a measure, this combination was significantly better than all other combinations of parameters (ANOVA, $p < 2 \times 10^{-16}$; Tukey's HSD, p-values < 2.3×10^{-4}) except for the following two combinations: 200 features selected with Random Forest, with multiple transformations of each feature allowed, and 100 features selected with Random Forest, with multiple transformations of each feature allowed (Tukey's HSD, p-values 1.00 and 0.13, respectively). (*E*) AUC-ROC curve and (*F*) AU-PRC curve of a model built with all untransformed features, demonstrating the improved performance of the optimized model in (*C*) and (*D*) with respect to both measures. (*G*) AUC-ROC and (*H*) AU-PRC for a model trained using RD5 gene pairs and half of the nonredundant pairs (randomly selected) then applied to RD9 gene pairs (excluding RD5) and nonredundant pairs that did not overlap with those used in training the RD5 model. Using these performance measures, this model did not perform as well as a model trained on RD4 then applied to a test set composed of RD9 gene pairs (after removing RD4 pairs) and a random subset of half the nonredundant gene pairs (**Figure 2.2D-E**); therefore, RD5 was not selected for further analysis.

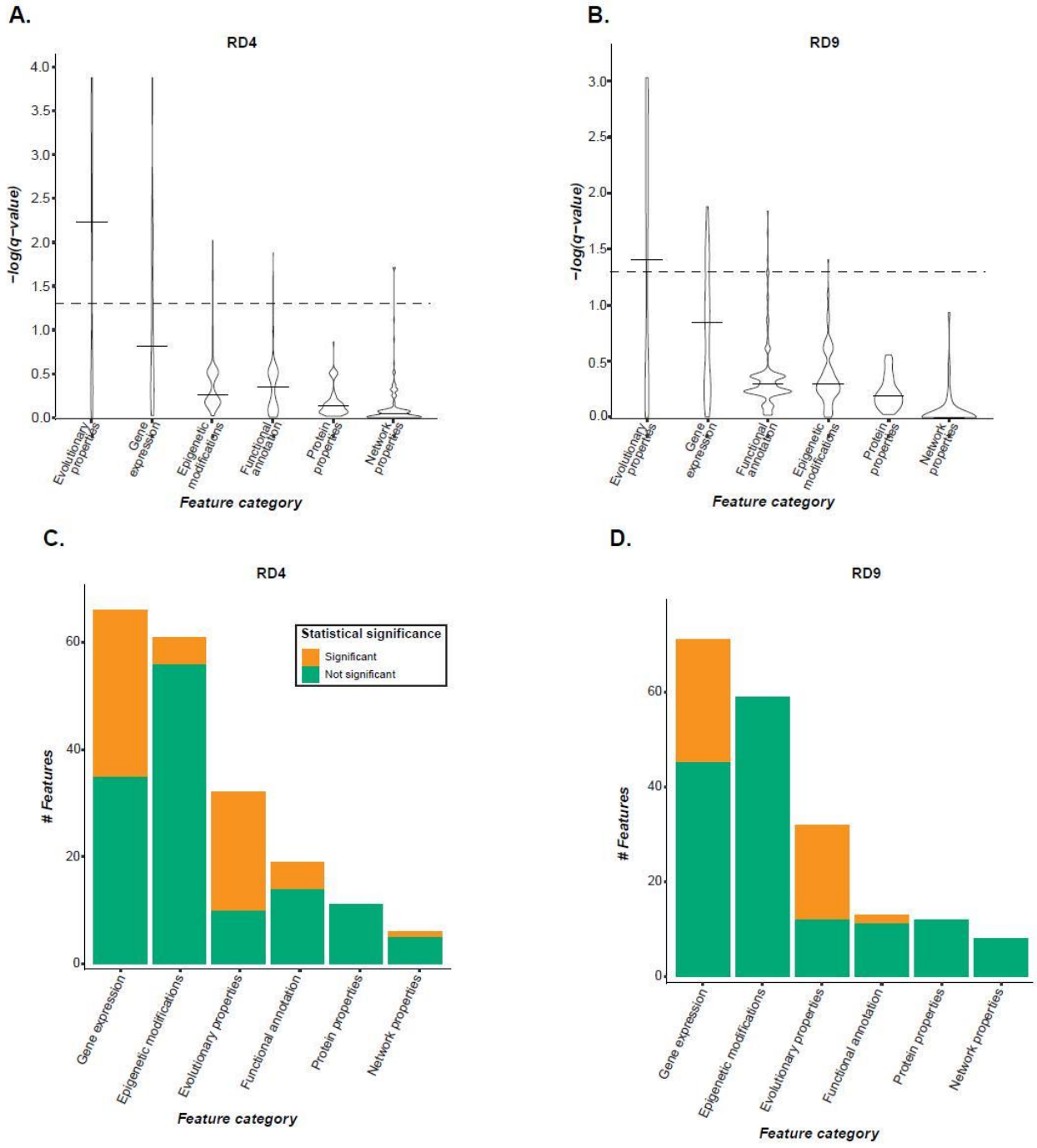

Figure 2.S3: Statistical association of features among different feature categories with redundancy.

(*A*-*B*) Distribution of -log(*q*-values) from tests of feature association with redundancy as defined using (*A*) RD4 and (*B*) RD9. Statistical significance was determined with Wilcoxon rank sum test for continuous features and two-sided Fisher's exact test for binary features; all values were

Figure 2.S3 cont'd
corrected for multiple testing with the Benjamini-Hochberg method. The dotted lines show a *q*-value of 0.05. All *p*-values, *q*-values and effect sizes are reported in **Table S1**. The median effect sizes (calculated as described in **Methods**) were 0.11 (RD4) and 0.07 (RD9) among continuous features, and 1.4 (RD4) and 1.1 (RD9) among binary features. (*C*- *D*) Distribution by feature category of the 200 features selected for model building for (*C*) RD4 and (*D*) RD9. Features that have a statistically significant association with redundancy as described above are shown in orange. Only 25% of the features selected were significantly associated with redundancy.

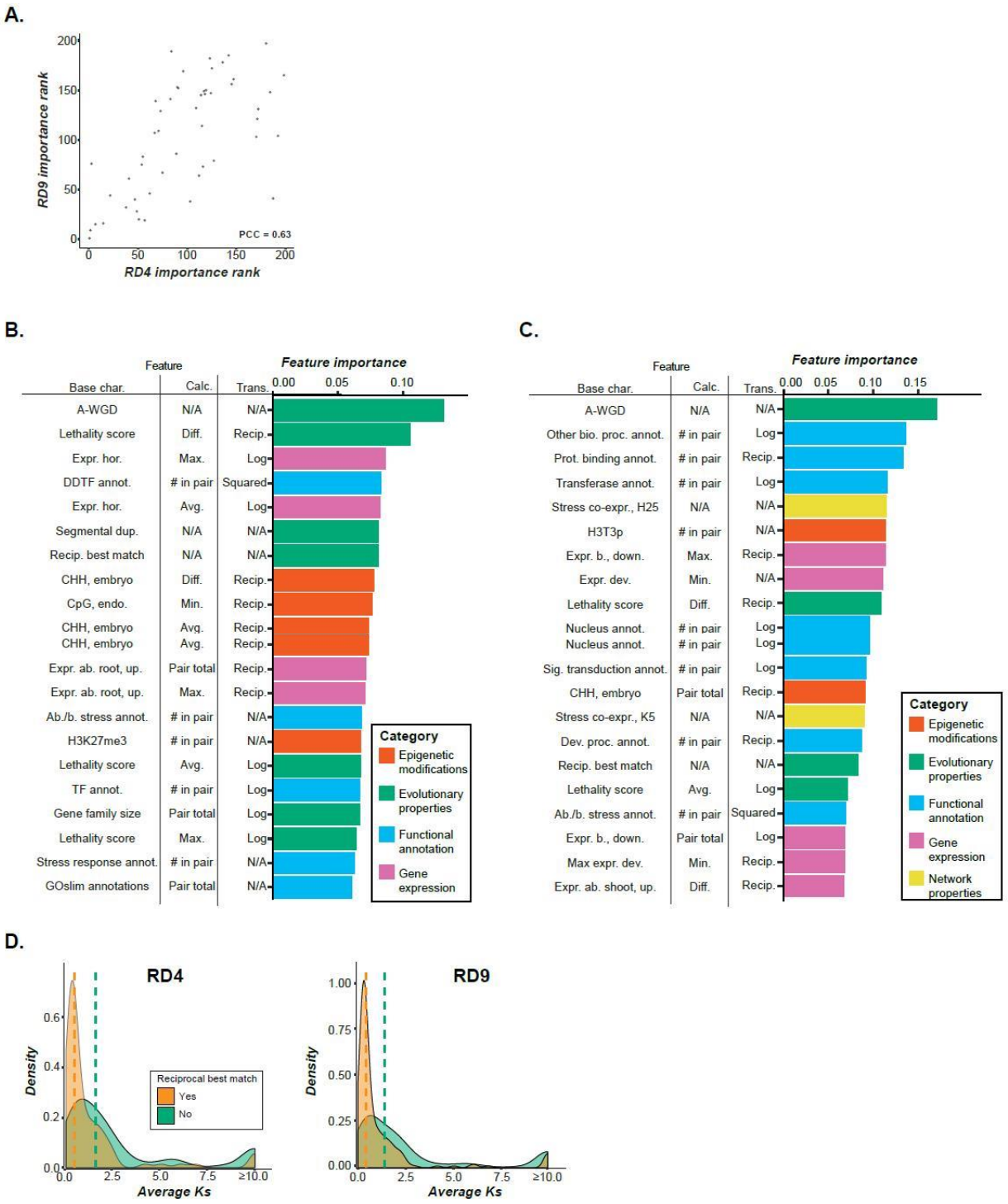

Figure 2.S4: Important features in predicting redundancy.
(*A*) Comparison of feature importance ranks of the 51 features included in both the RD4 and RD9 models. The feature importance ranks are well correlated between the two models (PCC = 0.63; $p = 6.0 \times 10^{-7}$), indicating that a core set of features is important in predicting redundancy

Figure 2.S4 cont'd
across definitions. (*B*-*C*) Feature importance scores obtained from machine learning models for (*B*) RD4 and (*C*) RD9. (*D*) Distribution of *Ks* values among gene pairs included in the RD4 (left) and RD9 (right) models that are reciprocal best matches (orange) and not reciprocal best matches (green). Dotted lines show the median *Ks* value for each group. Gene pairs that are reciprocal best matches tend to be more recent duplicates as shown by the lower *Ks* values.

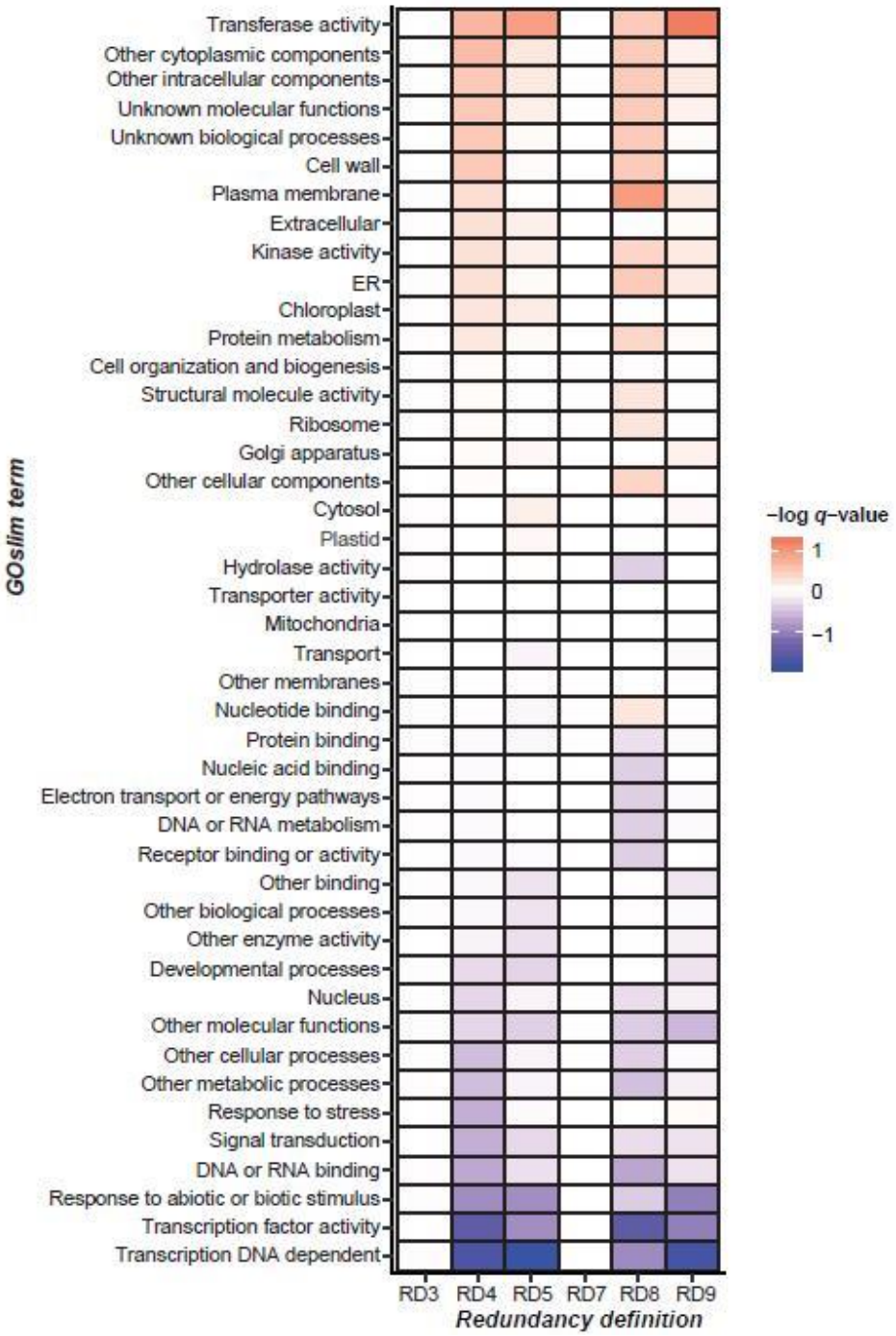

Figure 2.S5: Enrichment of GO terms among redundant gene pairs vs. nonredundant gene pairs for each redundancy definition.
Blue represents enrichment among nonredundant pairs while red represents enrichment among redundant gene pairs; lighter shades show statistically weaker associations (i.e., higher *q*-value)

Figure 2.S5 cont'd
and darker shades show statistically stronger associations (lower *q*-value). Statistically significant enrichment was seen in transcription factor activity among nonredundant pairs compared with RD4 and RD8 gene pairs, and in DNA-dependent transcription factor activity among nonredundant gene pairs compared with RD4, RD5 and RD9 gene pairs. In general, functional enrichment highly varied by RD.

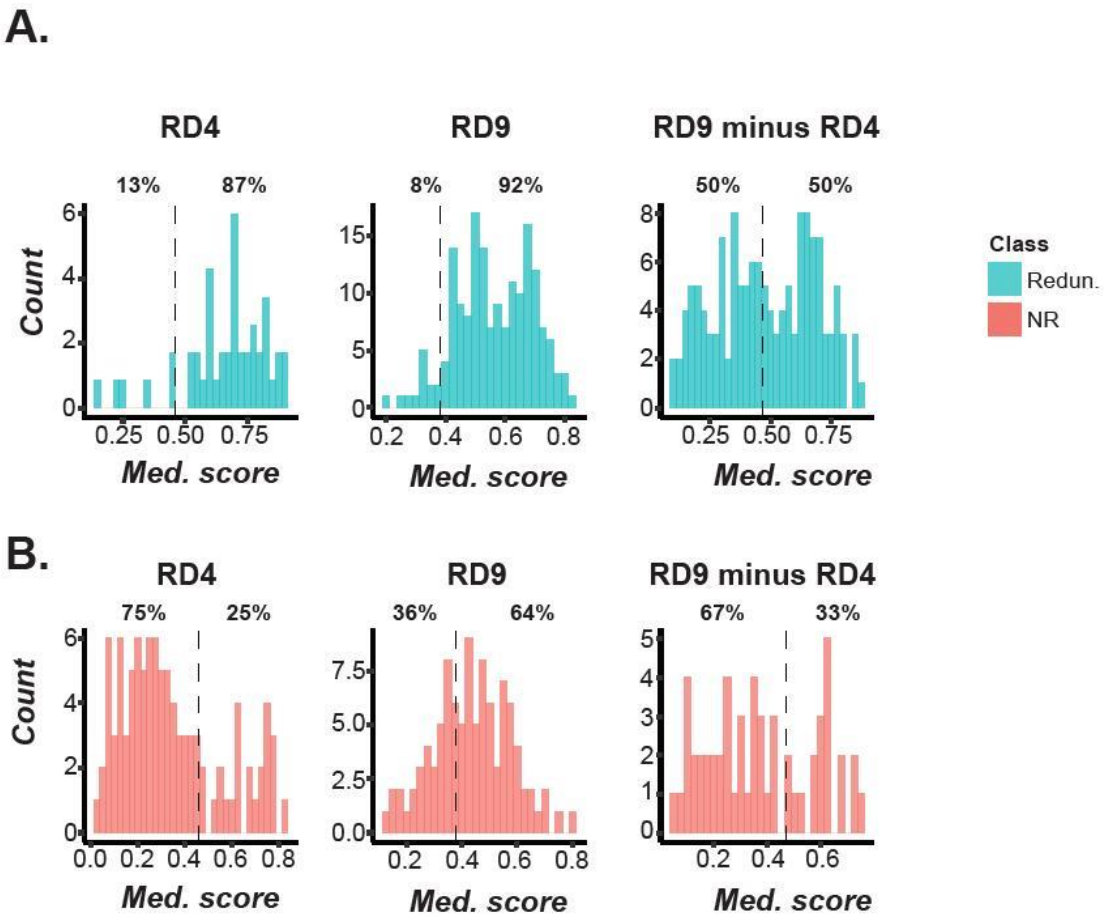

Figure 2.S6: Performance of models trained on RD4 and RD9 in cross-validation.
(*A*-*B*) Performance in cross-validation; the percentages of (*A*) redundant and (*B*) nonredundant gene pairs correctly and incorrectly predicted using different RDs are shown. Gene pairs to the left of the threshold (dotted line) were classified as nonredundant and gene pairs to the right were classified as redundant. Models were built using RD4 pairs, RD9 pairs, and RD9 pairs not included in RD4 as the redundant instances (RD9 minus RD4) with a randomly-selected half of nonredundant pairs as the nonredundant instances.

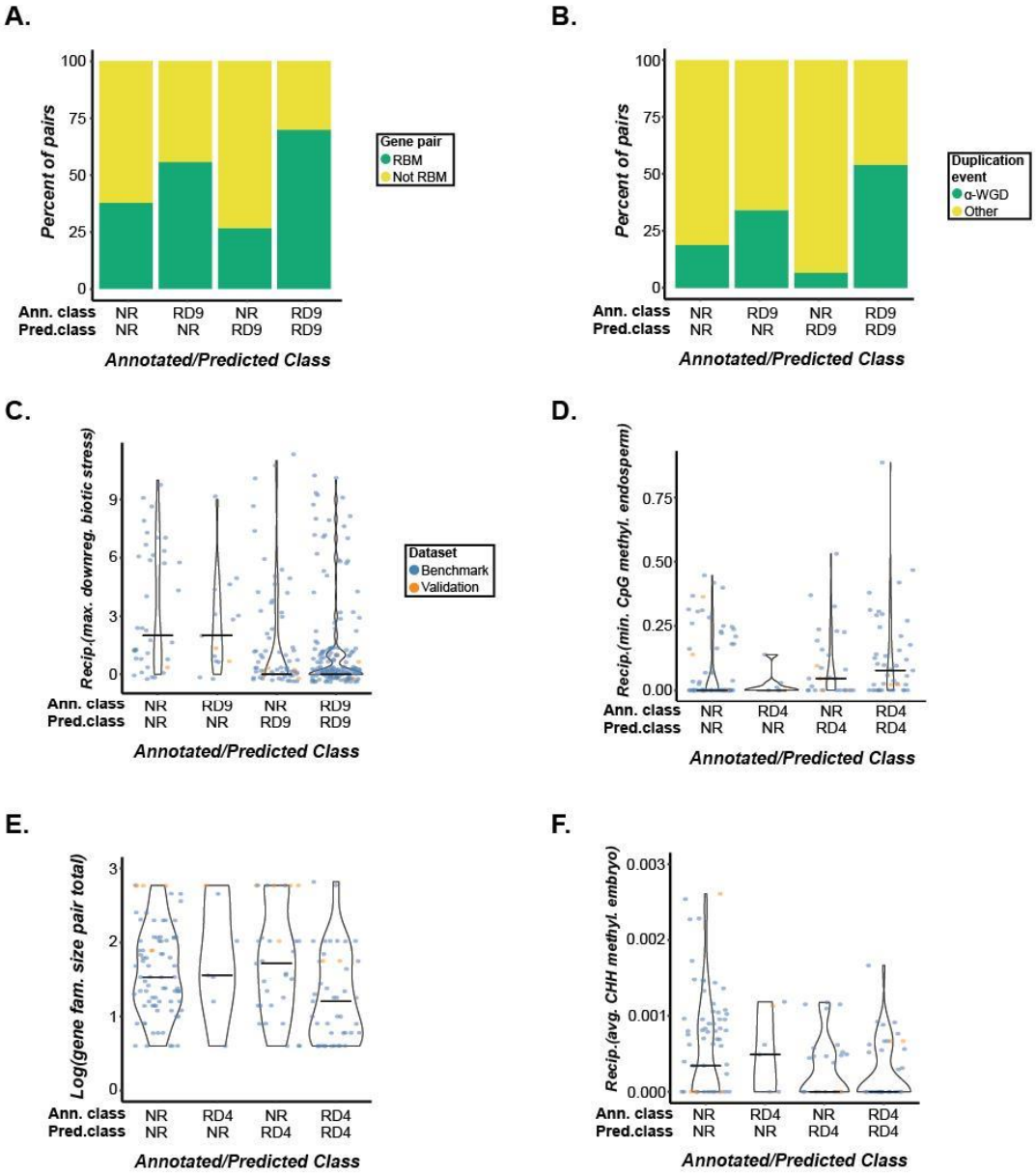

Figure 2.S7: Distribution of values among features that may contribute to mispredictions. (*A*) Distribution of reciprocal best match gene pairs (RBM) among annotated (Ann.) vs. predicted (Pred.) RD9 classes: true negatives (NR/NR), false negatives (RD9/NR), false positives (NR/RD9), and true positives (RD9/RD9), including the benchmark dataset and the 10 validation pairs identified from the literature (here referred to as validation pairs). Two NR/RD9 validation pairs were reciprocal best matches, which was observed more often for RD9/RD9 pairs than NR/NR pairs, while genes in the RD9/NR validation pair were not reciprocal best matches, likely explaining these three mis-predictions. (*B*) Distribution of α-whole genome duplication (α-WGD)-derived gene pairs among the annotated/predicted classes, including the benchmark and 10 validation pairs. Three validation NR/RD9 pairs were derived from the α-WGD event, which was observed more often for RD9/RD9 pairs than NR/NR pairs, potentially

Figure 2.S7 cont'd
contributing to their mis-prediction. (*C*) Distribution of feature values among benchmark and validation gene pairs for the maximum biotic downregulation breadth between genes in a pair; a reciprocal transformation was applied to generate reciprocal maximum biotic downregulation breadth. All five of the validation NR/RD9 pairs had high values for this feature and looked more similar to RD9/RD9 pairs than to NR/NR pairs. (*D*) Distribution of reciprocal minimum CpG methylation in endosperm cells values among benchmark and validation pairs. (*E*) Distribution of total gene family size values among benchmark and validation pairs. The RD4/NR validation pair had a high value, which was more consistent with the values of NR/NR pairs than RD4/RD4 pairs. (F) Distribution of reciprocal average CHH methylation in embryo tissue values among benchmark and validation pairs. All four of the NR/RD4 validation pairs had high values that were more similar to those of RD4/RD4 gene pairs, while the one RD4/NR pair had a low value more similar to those of NR/NR pairs.

CHAPTER 3: MODELING MUTANT FITNESS AND GENETIC REDUNDANCY IN SACCHAROMYCES CEREVISIAE

3.1 Abstract

Genetic redundancy is a phenomenon where more than one gene encodes products that perform the same function. This frequently experimentally manifests as a single gene knockout mutant that does not demonstrate a phenotypic change or a decrease in fitness compared to wild type individuals, due to the presence of a paralogous gene that may have retained the same function. A phenotype is only observed when one or more paralogs are knocked out in combination. While some factors that are associated with genes with retained, potentially redundant functions have been identified in yeast, little is known about factors contributing to long-term maintenance of genetic redundancy. Here, we leveraged the availability of fitness and multi-omics data in budding yeast *Saccharomyces cerevisiae* to build machine learning models for predicting genetic redundancy and related phenotypic outcomes (single and double mutant fitness) among paralogs. We found that single mutant fitness was particularly informative to predict double mutant fitness with a high degree of accuracy (Pearson's correlation coefficient, PCC, = 0.83). While the genetic redundancy model did not perform as well (PCC = 0.40), six of the top 10 features identified as important in predicting double mutant fitness were shared with the redundancy model. Our models provide quantitative assessments of how well existing data allow predictions of fitness and genetic redundancy. In addition, these models allow the identification of potentially biologically significant features contributing to long-term maintenance of genetic redundancy in eukaryotes.

3.2 Introduction

While there has been a rapid increase in the number of organisms with sequenced genomes as sequencing technology has improved and costs decreased, a key remaining challenge

is linking genotypes to phenotypes, and especially genes to functions. In service of this goal, large-scale reverse genetic screens have been undertaken in many organisms (e.g., Lu et al. 2011, Matynia et al. 2008, Giaever et al. 2002, Lee et al. 2003). These experiments have had varying levels of success in identifying phenotypes, as a single gene knockout mutant often does not demonstrate a phenotype that is informative with respect to its function (Bouché and Bouchez 2001, Thatcher et al. 1998). This can be due to a number of reasons, including subtle, tissue-specific, or environmentally conditional phenotypes, or genetic interactions (discussed in Brookfield 1992; Bouché and Bouchez 2001; Bolle et al. 2013).

Saccharomyces cerevisiae has been a popular eukaryotic model system for reverse genetics for decades because of its small size, short generation time, and genetic tractability (Botstein and Fink 1988). This means that knockout studies can be conducted at a scale that is not feasible with physically larger or physiologically more complicated organisms. The factors that can lead to a lack of apparent phenotype in single mutants can be overcome in this system: the environment can be manipulated with relative ease (e.g., increasing the incubation temperature, addition of antibiotics or limiting essential elements in the growth medium); fitness phenotypes in the form of colony size can be precisely quantified; and there are no complications of tissue-specific function in a unicellular organism.

Importantly, an approach called synthetic genetic array (SGA) has been developed in yeast for studying double mutants (DMs) in a systematic way (Tong 2001). Utilizing haploid single mutants (SMs), SGA involves crossing each SM (the query mutant) with all other possible SMs (array of mutants) to generate a comprehensive library of haploid DMs. Normalized colony size (i.e., fitness) of DMs can be compared to the corresponding SMs and to the wild-type (WT) to identify interactions among pairs of genes. These can include positive genetic interactions,

where the DM has greater fitness than expected compared to the SMs, or negative interactions, where the DM has lower fitness. There are several notable and sometimes overlapping subsets of negative genetic interactions, including synthetic lethality and genetic redundancy. Synthetic lethality refers to a situation where a DM is lethal while both of the corresponding SMs are viable (Dobzhansky 1946). The term genetic redundancy typically refers to paralogs that have retained some or all of the same function following the duplication event from which they are derived (e.g., Liu et al. 2008, Mendonca et al. 2011, Rutter et al. 2017). Genetic redundancy manifests experimentally as any other type of negative genetic interaction, where the DM has a more severe fitness decrease compared with either of the corresponding SMs (this can include synthetic lethality). The difference is that among redundant gene pairs, this effect is specifically seen due to the phenotypic buffering effect of having an evolutionarily related "backup" gene to perform the function when one is knocked out.

The idea of having multiple genes performing the same function as an evolutionary failsafe makes some intuitive sense. However, retention of functionally redundant genes over millions of years is seemingly an evolutionary paradox, because selective pressure would be expected to relax on one copy, allowing mutations to accumulate and loss of the ancestral function over time (Brookfield 1992). However, there are genetically redundant gene pairs in the yeast genome that have been retained for hundreds of millions of years (Vavouri et al. 2008). Features common among retained paralogs have been relatively well-studied (Seoighe and Wolfe 1999; Scannell et al. 2007), but there is less information about factors associated with and mechanisms contributing to maintained genetic redundancy, i.e., retained paralogs with overlapping functions (Guan et al. 2007; Li et al. 2010).

A previous study used SGA to identify of hundreds of thousands of positive and negative genetic interactions in yeast (Costanzo et al. 2016), a resource that can be used to better understand factors associated with different types of genetic interactions. Using these published phenotype data, we built upon previous efforts to understand the biological underpinnings of genetic redundancy. This was accomplished using machine learning to build predictive models of phenotypic severity in single and double mutants, and of genetic redundancy among gene pairs as a function of single and double mutant fitness. Because we used a large dataset where phenotypes are known, we were able to first assess the accuracy of the models on the known data and then identify potential biologically important features identified by the machine learning model as important in making those predictions. The single mutant fitness model confirmed the validity of using predictive model building to identify biologically important features; as expected, essential ("housekeeping") functions were identified as important in predicting severe phenotypes. Models were then built for double mutant fitness and genetic redundancy, and the features that were important in generating those predictions analyzed to identify biologically relevant factors which may be contributing to double mutant phenotype severity and genetic redundancy.

3.3 Results and Discussion

3.3.1 Distribution of fitness scores

We first evaluated the structure of the single mutant (SM) and double mutant (DM) fitness data. Our analyses included the 5327 SMs having a reported fitness value in the 26°C datasets of Costanzo et al. (2016; see **Methods**). Fitness values for these mutants ranged from 0.13 to 1.14 (**Figure 3.1A**), with a value of 1 representing no change from the WT fitness. The

data in our set were extremely unbalanced; the values were clustered tightly around 1, with a median SM fitness value of 0.99 (mean = 0.93, standard deviation [SD] = 0.16; **Figure 3.1A**). These results indicated that under normal growth conditions, the vast majority of genes in the yeast genome can be knocked out with little or no fitness penalty. This is consistent with previous findings (e.g., Giaever et al. 2002).

DMs included in our analysis comprised paralogous gene pairs with reported fitness values for the DM and both corresponding SMs in the 26°C datasets of Constanzo et al. (2016; see **Methods**). There were 8160 of these DMs, with fitness values ranging from -0.09 to 1.28 (**Figure 3.1B**). Like the SM values, the DM fitness values were tightly clustered around 1, with a median DM fitness value of 0.96 (mean = 0.87, SD = 0.21; **Figure 3.1B**). Together with the SM fitness values, this result suggests that the majority of genes in the genome can be knocked out either alone or in combination with the closest paralogous gene with little or no consequence to the fitness of the organism at 26°C. There are several possible explanations for this observation: many genes may have conditional phenotypes (e.g., only present under stress conditions); fitness could be affected in a way that is not captured by colony size (such as a difference in cell number); or there may be an extremely high degree of genetic redundancy involving three or more paralogs. This last point depends greatly on how paralogs are defined; according to our definition, 46% of the genes in our DM dataset are members of a gene family with three or more members (**Figure 3.S1**). However, using only paralogous genes which have been confirmed in the literature (according to *Saccharomyces* Genome Database; Cherry et al. 1998), our dataset contains no gene families with more than three paralogs and only five gene families containing three paralogs. Thus, it is unlikely that a high degree of redundancy among three or more paralogs would account for the lack of phenotypes observed in both SMs and DMs.

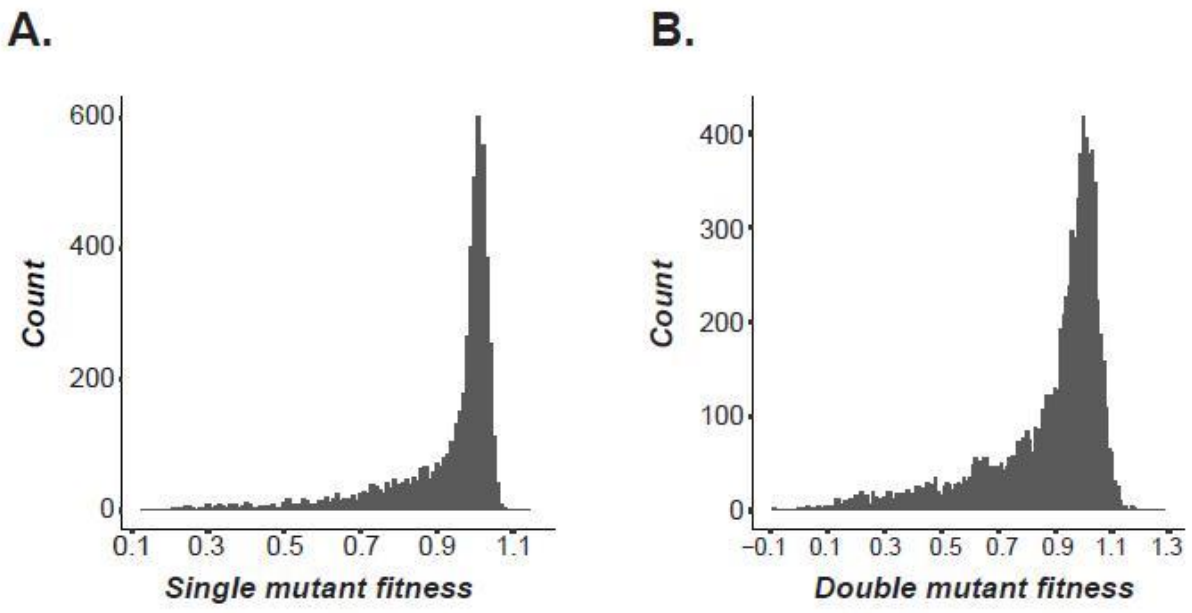

Figure 3.1: Distribution of mutant fitness values.
Distribution of (A) single and (B) double mutant fitness values in our dataset. Single mutants were those for which fitness data were available in the 26°C dataset from Costanzo et al. (2016); double mutants were those for which fitness data were available in the Costanzo et al. dataset for a double mutant and both corresponding single mutants, where the genes in the double mutant pair are paralogs.

3.3.2 Predictions of single mutant fitness

To ultimately model genetic redundancy in yeast to identify biological factors contributing to it, we first tested the validity of this approach on a simpler problem: prediction of SM fitness. We applied machine learning to build a regression model of SM fitness in yeast using 11,661 features spanning five categories (**Supplemental Data** and **Methods**), including functional annotations (e.g., GO terms), evolutionary properties (e.g., presence of homologous genes in other species), protein sequence properties (e.g., PTMs), gene and protein expression (e.g., stress conditions under which a gene is differentially expressed vs. control), and network properties (e.g., gene interactions). In the context of machine learning, each SM is here

considered an instance, i.e., one observation. Data used in machine learning comprised the 5327 SM instances, the corresponding SM fitness values as the predicted variable, and the gene features corresponding to each SM. The instances were divided into a training and a testing set (90% and 10% of instances, respectively; see **Methods**). The support vector regression (SVR) algorithm was applied to build a predictive model of SM fitness from the training set using ten-fold cross-validation (see **Methods**), and the resulting model was then applied to the testing set. Model performance on the training and testing sets was evaluated using Pearson's correlation coefficient (PCC; 0 in a random model and -1 or 1 in models with perfect negative or positive correlation, respectively). The model had a PCC = 0.40 on the training set and 0.35 on the testing set (**Figure 3.2A)**. To determine whether this would be the performance of the model if the features were not actually predictive of fitness (i.e., if this result was simply due to chance), we randomized the fitness values and reran the model. The PCC was significantly lower than with the original data (0.029 and -0.036 on training and testing sets, respectively; **Figure 3.2B**), indicating that the features do have predictive value for single mutant fitness, and therefore some likely biologically relevant association with this trait.

3.3.3 Feature importance in predicting SM fitness

Output of our SVR model included feature importance scores, which represent the contribution of each feature to predictions; high importance scores have larger weight. Among features used for SM fitness predictions, eight of the top ten predictors (determined by importance score rank) were features related to specific genes having interactions with other genes (**Figure 3.2C**). Generally, a feature such as interaction with a specific gene is highly sparse—only a small number of genes in the dataset would be interacting, and this held true in

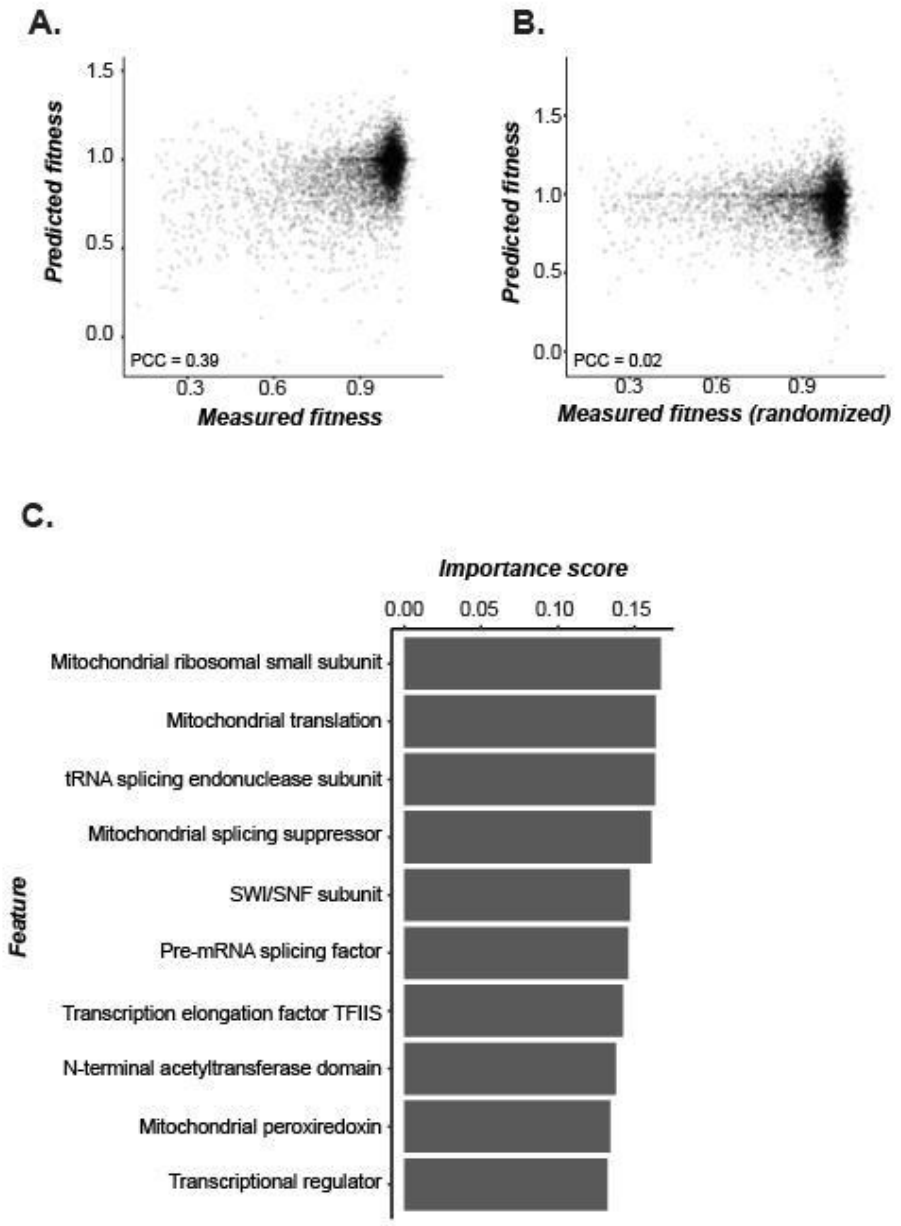

Figure 3.2: Performance and important features of single mutant fitness model.
Performance of (A) single mutant fitness model (Spearman's ρ = 0.36) and (B) single mutant fitness model with randomized fitness values (Spearman's ρ = 3.8x10-3). Models were built with the Support Vector Regression (SVR) algorithm. Randomized fitness values simulated a model where features and fitness values were not related in any way beyond random chance, demonstrating that a relationship did exist between SM fitness and our features. (C) Features with the ten highest-ranked importance scores in predicting SM fitness. Mitochondrial translation (ranked 2nd) was a Gene Ontology (GO) term; N-terminal acetyltransferase domain (ranked 8th) is a PANTHER protein domain; all others represent interaction with a gene which has been given

Figure 3.2 cont'd
a simplified descriptor. In order from rank 1-10 the gene names are: MRPS12, SEN2, MSS51, SNF11, CDC40, DST1, PRX1, and OPI1.

our dataset. Each of the important gene interaction features was explanatory for less than 0.3% of gene pairs in the dataset (13 genes or fewer out of our set of 5327 had each of these interactions; **Figure 3.S2**). Thus, while these features were informative for explaining the fitness effects of some single mutants which had these interactions, the predictive power of the features was not generalizable throughout the genome.

To determine potential biological reasons why these specific features were identified as the most important in generating predictions, we looked at the functional annotations of the genes with which interactions were predictive of fitness. In general, the features were related to the levels and integrity of mRNA and proteins. Specifically, they included a splicing factor, splicing suppressor, and splicing endonuclease; a transcription factor and transcriptional regulator; a chromatin remodeling complex; mitochondrial translation function and a mitochondrial ribosomal small subunit; and N-terminal acetyltransferase activity, which typically occurs at the ribosome (Polevoda et al. 2008). To determine the directionality of the predictive power (e.g., whether knocking out a gene with N-terminal acetyltransferase activity is associated with increased or decreased fitness), we next compared the distributions of fitness values between SMs that had gene interactions identified by the model as important and those that did not (**Figure 3.S3**).

While the difference was more dramatic for some features than for others, SMs for genes with the interactions tended to have decreased fitness compared to those that did not have them. As mentioned above, these features mainly represent key housekeeping functions without which

a cell cannot metabolize and/or grow. For example, four of the features were directly related to mitochondrial function (**Figure 3.S3A, B, D, I**), with an additional feature (**Figure 3.S3J**) annotated as having disrupted mitochondrial metabolism when knocked out (Luévano-Martínez et al. 2013). It is therefore plausible that altering their functions would be related to decreased fitness, demonstrating the biological validity of our approach. One might hypothesize that the prevalence of mitochondrial-related genes and functions may be a result of decades of extensive study of mitochondria, leading to a much higher number of genes having annotations related to mitochondria than to other functions, making these features less sparse and therefore more informative. While we knew that these features were sparse with respect to the number of genes in the dataset, we hypothesized that genes having annotations related to mitochondria may still be disproportionately high. To determine whether importance among features such as GO terms and gene interaction was simply a function of sparsity, we assessed whether there was a linear relationship between the feature importance score and the number of genes having the annotation or interaction represented by that feature. Using this measure, importance scores do not seem to be dependent on the sparsity of the feature (PCC = -0.02, p = 0.04; **Figure 3.S4A**).

3.3.4 Predictions of DM fitness and comparison of important features

As the SM fitness model demonstrated the validity of regression-based machine learning to predict SM fitness and evaluate identified features as biologically significant, we next sought to determine whether this approach would be useful for a related and more complex trait. We therefore built a model of DM fitness, deriving 17,199 feature values (**Supplemental Data**) from the corresponding SMs for each DM (see **Methods**). The DM fitness model performed much better than the SM model, with PCC = 0.73 and 0.80 for the training and testing sets,

respectively (**Figure 3.3A**). Randomization of the fitness values to approximate predictions of fitness with uninformative features resulted in a much worse model (PCC = 0.01; **Figure 3.3B**), demonstrating that there is a relationship between the features and fitness that is stronger than random chance.

While feature importance analysis in the SM fitness model served as a demonstration of the validity of our methods, we here used feature importance scores to identify potential biological factors contributing to DM fitness. In contrast with the SM feature importance scores, the ten highest ranked features (**Figure 3.3C**) did not include any specific gene interactions, but rather included GO terms (eight features), protein-protein interactions (one feature), and protein domains (one feature). The GO terms included transcriptional regulation under zinc starvation, function in fermentation, snoRNA localization, localization to the Golgi, RNA polymerase assembly, and both the GO and GOslim terms for mitochondrial translation. Also included was the number of GO terms that were shared between genes in a pair. Similar to the results from the SM fitness model, these are key cellular functions, which may explain the association with decreased fitness. Also consistent with the SM fitness model, there was no correlation between feature sparsity and importance rank (PCC = 0.05, *p*-value = 4.3×10^{-6}; **Figure 3.S4B**), meaning these features were not selected as important simply as that type of artifact of the data structure.

We next looked at the distribution of values for important features in relation to DM fitness to determine the directionality of their predictive power (i.e., whether higher feature values were related to higher or lower DM fitness; **Figure 3.S5**). For the important GO terms identified, we found that DMs in which one gene had the annotation generally had lower fitness values, while DMs in which neither had the annotation generally had higher fitness values. For the important features representing shared GO terms and protein-protein interactions (referred to

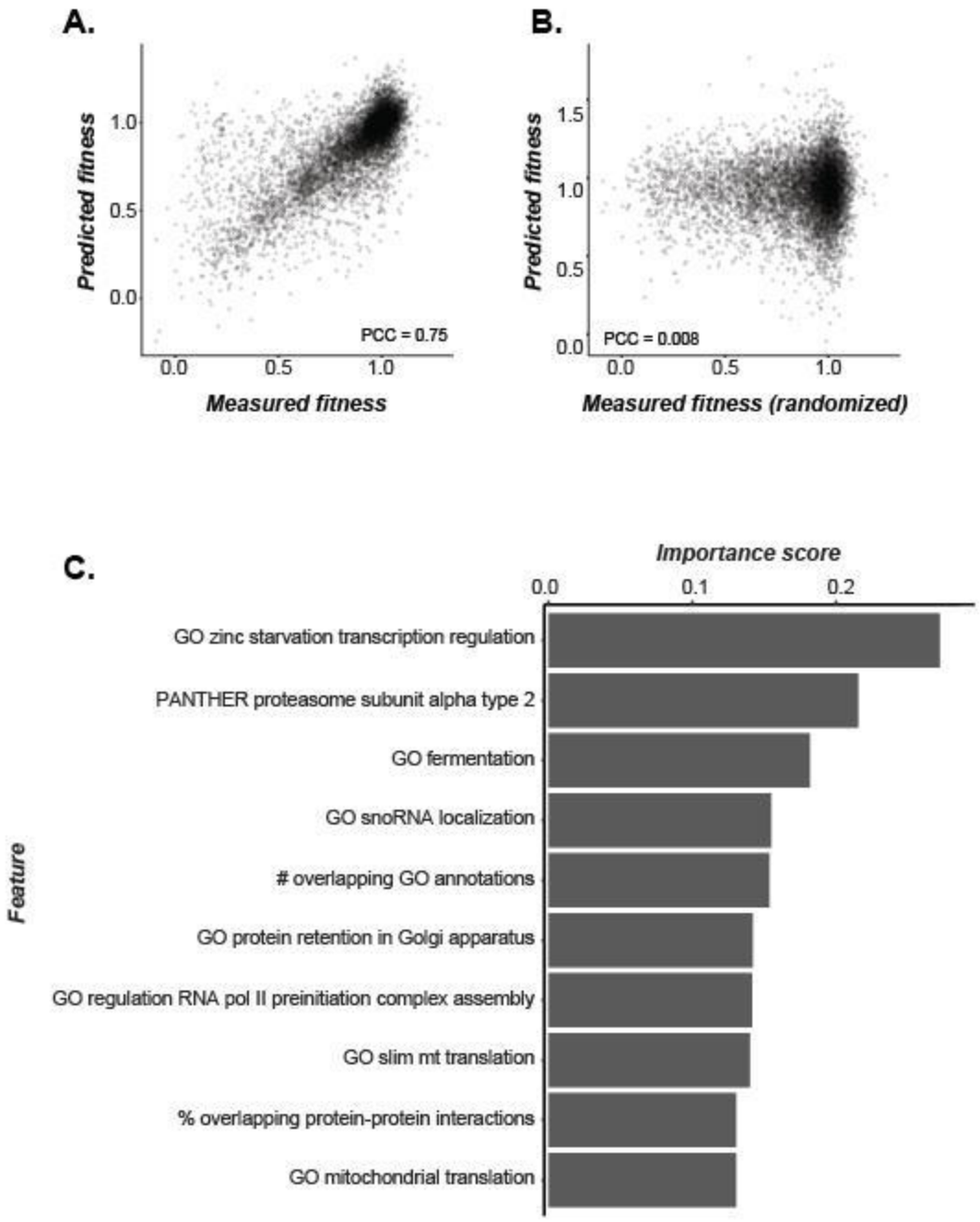

Figure 3.3: Performance and important features of double mutant fitness model.
Performance of (*A*) DM fitness model (Spearman's ρ = 0.71 and (*B*) DM fitness model with fitness scores permuted to represent no association between features and fitness values (Spearman's ρ = 0.009). (*C*) Features with the ten highest-ranked importance scores in predicting DM fitness. In contrast with the SM fitness model, important features were primarily GO terms rather than interactions with specific genes.

as “# overlapping GO terms” and “% protein-protein interactions”, respectively), we expected that genes having a high percentage of the same protein interactors and high number of shared GO terms may be performing similar functions or involved in the same pathway(s), and therefore would have a strong fitness effect when both genes were knocked out. This was not the case, however; there was no significant correlation between percent shared PPI or number of shared GO terms and DM fitness (Spearman’s rank sum rho = -0.03 and -0.04, respectively; **Figure 3.S5E, I**). Taken together, we found that, consistent with the SM fitness model, key housekeeping functional annotations are predictive of reduced fitness, but that due to the sparsity of genes having these annotations, the predictions are accurate for a small number of genes.

3.3.5 Prediction of double mutant fitness using single mutant fitness

With one DM fitness model in place, we wondered whether the high level of gene interactions in yeast (Costanzo et al. 2016) would limit the utility of SM fitness data in predicting DM fitness. Where two genes do not have epistatic interactions, the fitness of the corresponding DM can be modeled as the sum of the fitness decreases of both SMs, making prediction of DM fitness based on SM fitness a very straightforward proposition. However, gene interactions such as genetic redundancy can lead to non-additive fitness effects in a DM compared to the corresponding SMs, which would be expected to complicate the problem of DM fitness prediction significantly. To answer this question, we incorporated features derived from SM fitness into the DM fitness model: for each DM, we included the difference, average, maximum, minimum and total SM fitness for the genes in each pair as features. Inclusion of these SM fitness score-derived features (SMF features) did improve the model (PCC = 0.80 and 0.84 for training and testing sets, respectively; **Figure 3.4A-B**). The improvement in model performance

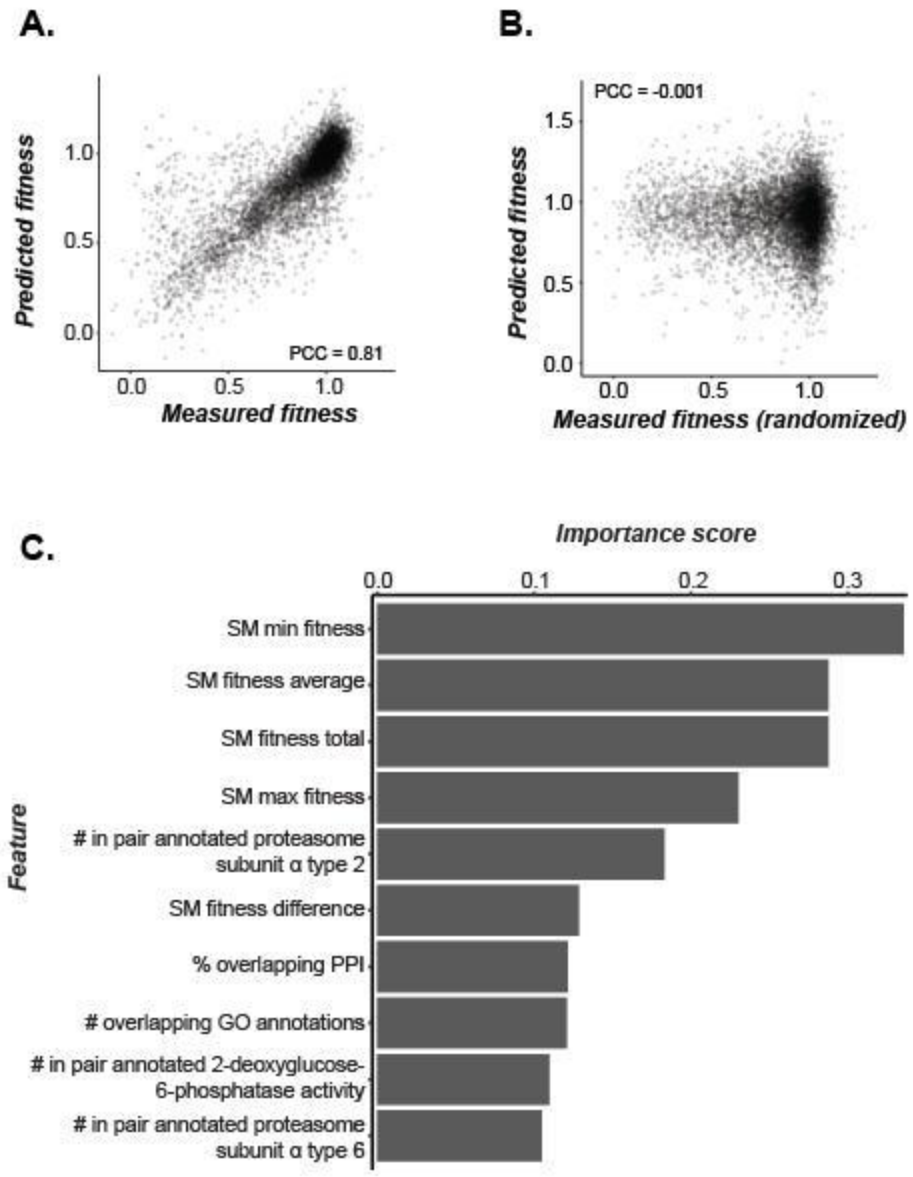

Figure 3.4: Performance and important features of double mutant fitness model including single mutant fitness-derived features.
Performance of (*A*) DM fitness model with single mutant fitness (SMF) features included (Spearman's ρ =) and (*B*) DM fitness model with SMF features included and fitness scores permuted (Spearman's ρ =). SMF features for each gene pair comprised the maximum, minimum, average, difference, and total values for the corresponding SM fitness values. (*C*) Features with the ten highest-ranked importance scores in predicting DM fitness with SMF features included. SMF features were extremely important predictors of DM fitness.

indicated that the SMF features were important in predicting DM fitness, which was confirmed by the SMF features occupying five of the six top feature importance ranks for this model (**Figure 3.4C**).

As mentioned above, due to the prevalence of genetic interactions in yeast, this importance of SMF features in predicting DM fitness was not expected, but suggests an important relationship between SM and DM fitness. To determine how effective this relationship could be in predicting DM fitness, we built another model using only the five SMF features (SMF feature model). This model performed better still than the previous model, with PCC = 0.84 and 0.83 for the training and testing sets, respectively (**Figure 3.5A**), and randomization of the DM fitness values demonstrated that this performance was not achieved by chance (**Figure 3.5B**). The higher performance of the model with five features compared to the model with ~17,000 features may be counterintuitive, as a strength of machine learning in general is the combination of features to achieve more accurate predictions than could be attained with any one feature. However, the presence of many uninformative features can decrease the performance of the model. The sparsity of most of the functional annotation data may have made many of them uninformative, explaining the decreased performance of the model incorporating them.

To determine the contribution of each individual SMF feature to the DM predictions, we calculated the correlation between DM fitness and each of the SMF features. Despite our expectation that genetic interactions would equate to a nonlinear relationship between the fitness of DMs with the corresponding SMs, DM fitness showed a positive linear correlation with minimum, average, total, and maximum SM fitness, and a negative linear correlation with difference in SM fitness (Spearman's rank sum rho = 0.86, 0.88, 0.88, 0.63, and -0.65, respectively; *p*-values < 2.2×10^{-16}; **Figure 3.5C-G**). In general, a very strong linear correlation

of an individual feature with the predicted variable approximates the predictive power of the feature on its own, as was the case here (**Figure 3.S6**).

The difference in correlation strength/direction among the five SMF features indicated that the way in which SM fitness was represented affected the predictive power. The two features most strongly correlated with DM fitness were the average and the total SM fitness, features which contain fitness information about both of the SMs with a single number. Maximum and minimum SM fitness features by definition represent only one of the two SMs for each gene pair, and were not as well correlated with DM fitness. While the decrease in correlation was very small in minimum SM fitness, maximum SM fitness had a comparatively much weaker correlation with DM fitness. This suggests a result relevant to gene interactions: that SMs with lower fitness are less likely to have positive paralogous gene interactions than are SMs with high fitness.

3.3.6 Predicting degrees of genetic redundancy and comparison of important features

We next sought to determine whether we could accurately model genetic redundancy as a form of negative gene interaction, incorporating both SM and DM fitness data for a given gene pair. Redundancy scores were calculated based on the change in fitness between WT and single/double mutants, where the score would be 1 in the case of full redundancy and 0 would be nonredundant (see **Methods**). Using this definition of genetic redundancy, the redundancy scores in our dataset were concentrated around 0, with a median score of 0.0023 (mean = 0.01; standard deviation = 0.12; **Figure 3.6**), suggesting few instances of genetic redundancy are present in the yeast genome.

As with SM and DM fitness, we established a regression model to predict redundancy,

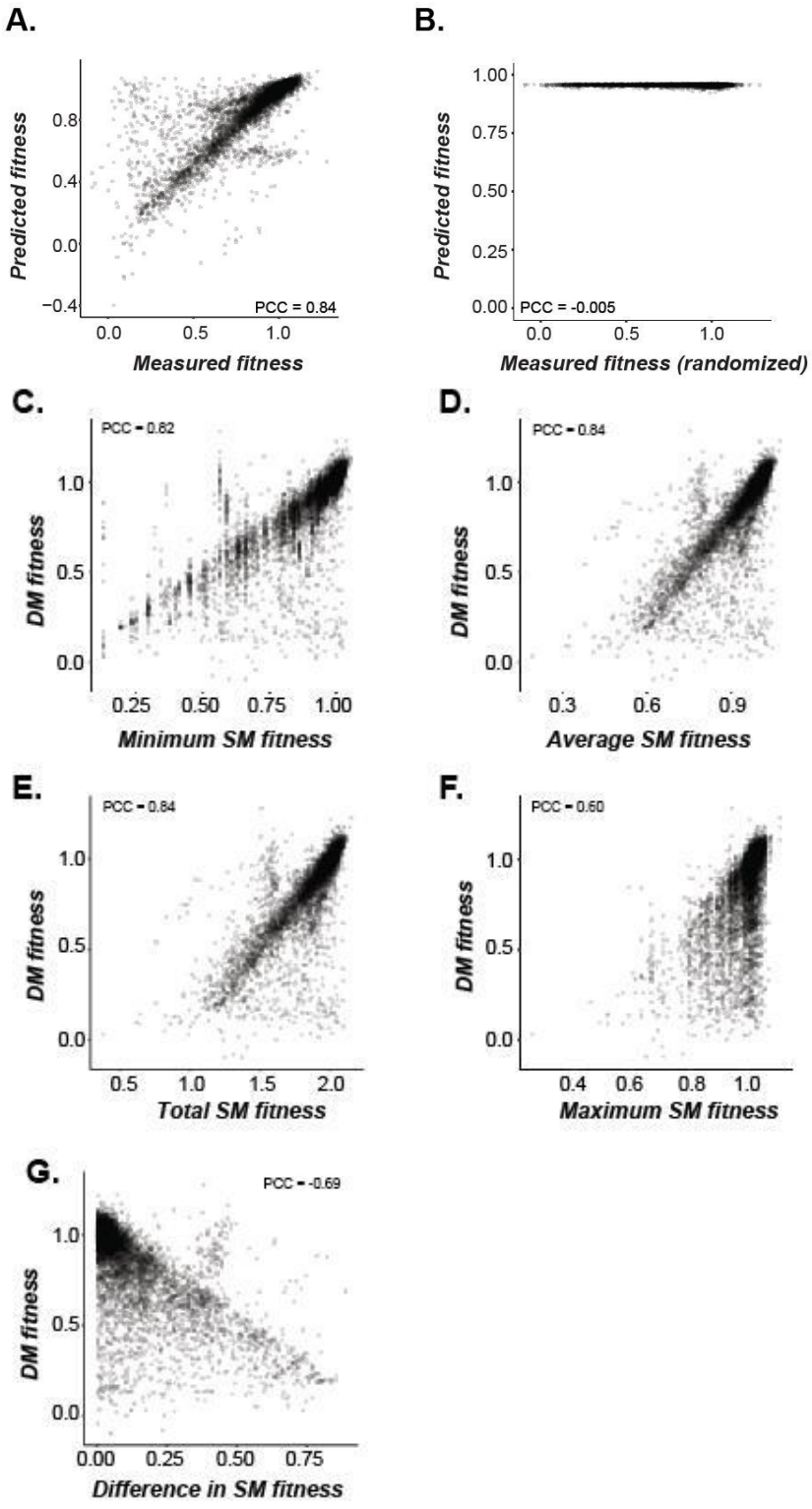

Figure 3.5: Performance of model built only with and correlation of fitness values with single mutant fitness-derived features.

Figure 3.5 cont'd
Performance of (*A*) DM fitness model with SMF features only (Spearman's ρ = 0.88) and (*B*) DM fitness model with SMF features only and fitness scores randomized (Spearman's ρ = -0.008). (*C-G*) Correlation of individual SMF features with DM fitness.

and used the same gene pairs and features used for the DM fitness model. In addition, five principal component scores derived from GO terms and protein domains (as a means of dimensional reduction for those sparse feature types) were included as features. The performance of the redundancy model was comparable to the performance of the SM fitness model, with PCC of 0.32 and 0.40 on the training and testing sets, respectively (**Figure 3.7A**). However, the model with randomized redundancy values here again demonstrated that there was predictive power in

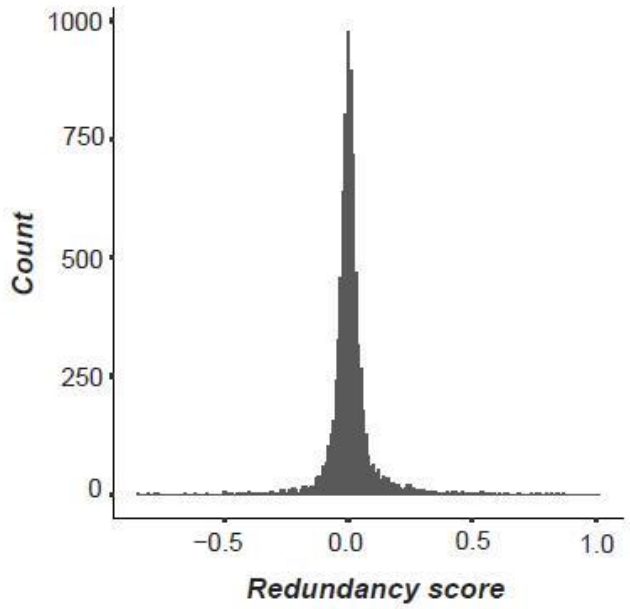

Figure 3.6: Distribution of genetic redundancy scores.
Distribution of genetic redundancy scores in our dataset. A value of 0 represents nonredundancy and a value of 1 represents full redundancy, with degrees of partial redundancy represented in between.

our features beyond random correlation (**Figure 3.7B**).

To identify potential biological factors associated with redundancy, we next analyzed the feature importance scores for the redundancy model. We hypothesized that there would be a relatively high degree of overlap in the most important features for predicting DM fitness and redundancy, as redundancy was directly calculated from fitness scores. This did turn out to be the case, with six of the ten most important features in the redundancy model shared with the DM fitness model (**Figure 3.7C**), and a high degree of correlation in importance score ranks between the DM fitness and redundancy models (**Figure 3.7D**).

Maximum SM fitness was the most important feature in predicting redundancy, despite being the least important of the SMF features for predicting DM fitness. Given the high feature importance scores and correlation of SMF features with DM fitness in that model, it was

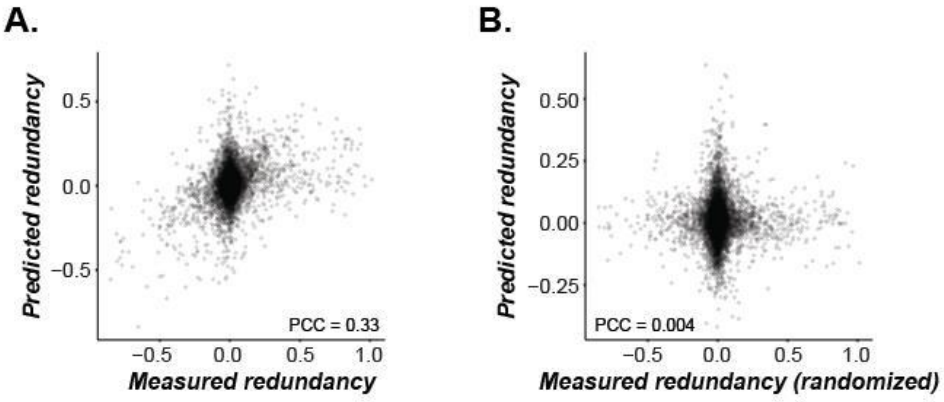

Figure 3.7: Performance and important features of redundancy model.
Performance of (A) redundancy model (Spearman's $\rho = 0.20$) and (B) redundancy model with fitness scores randomized (Spearman's $\rho = 0.006$).

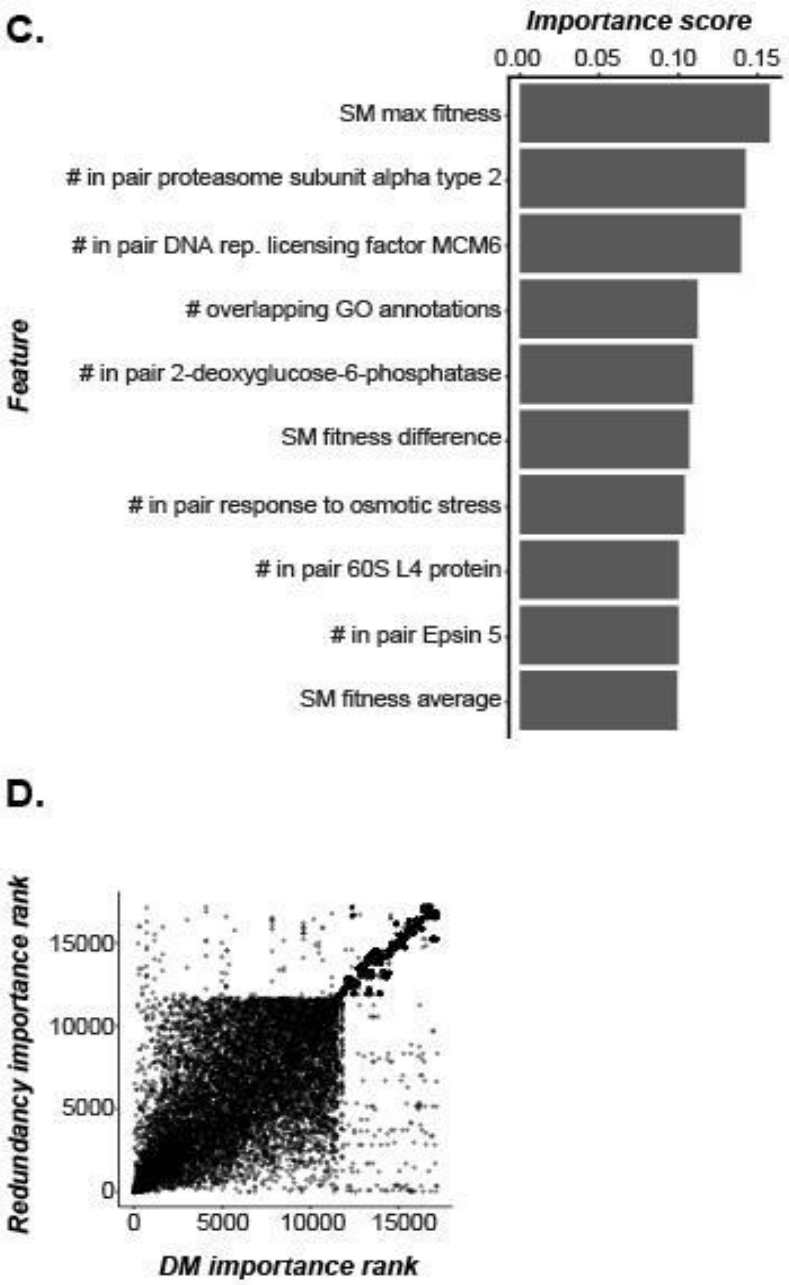

Figure 3.7 cont'd
(*C*) Features with the ten highest-ranked importance scores in predicting redundancy. (*D*) Correlation of feature importance scores in the DM fitness and redundancy models showing that features important in predicting DM fitness also tend to be important in predicting redundancy (PCC/Spearman's $\rho = 0.84$).

expected that important SMF features in the redundancy model would also be correlated with redundancy. However, none of the SMF features were very well correlated with redundancy due to the presence of many outliers. PCC was 0.18, -0.16, and 0.22 for maximum, difference in and average SM fitness, respectively (**Figure 3.S7A, F, J**). The high importance scores appear to be an effect of the skewedness of the data; the vast majority of SM fitness scores were clustered tightly around 1, so the maximum and average SM fitness scores were mostly high, and the difference in SM fitness scores was mostly low. As the redundancy scores were clustered around 0, there would appear to be an association of low redundancy scores with high maximum SM fitness, high average SM fitness, and low difference in SM fitness. This is simply because that is how the data were distributed, not because there was any correlation or predictive power.

Several of the features with high importance scores did show some real differences between high and low redundancy gene pairs. The number of genes in the pair that had a GO annotation of response to osmotic stress was related to redundancy (**Figure 3.S7G**). Gene pairs that had either 0 or 1 gene in the pair with this annotation had low redundancy scores, while instances where both genes had the annotation had high redundancy scores. This makes biological sense in that genes with the same annotation would be more likely to be redundant than gene pairs without the same functional annotations. A similar case was seen with other GO annotations, namely an annotation of 60S L4 ribosomal protein for both genes in the pair co-occurring with a high redundancy score (**Figure 3.S7H**) and an annotation of epsin 5 for only one gene in a pair co-occurring with a low redundancy score (**Figure 3.S7I**). This was once again likely due to the uneven distribution of data—for each feature, there was one gene pair with a high redundancy score and the GO annotation, making these features not generally predictive of genetic redundancy, but highly effective at accurately predicting a very small

number of gene pairs. In fact, for binary features such as GO annotations in this model, the feature importance score was essentially a function of the how many gene pairs had the annotation multiplied by the difference in median redundancy score among gene pairs with and without the annotation (**Figure S8**). Thus, feature importance in this case appears to be a combination of the structure of the data (feature sparsity, as discussed in **Section 3.3.3**) and true differences in redundancy score among gene pairs with specific annotations, such as the 60S L4 ribosomal protein and epsin 5. Furthermore, ribosomal genes are known to have many copies in yeast (Komili et al. 2007), and epsin is a family of highly conserved membrane proteins (Sen et al. 2012), pointing to potential broader biological significance of these results in the context of genetic redundancy.

3.4 Conclusions

In this study we used machine learning models to predict mutant fitness and genetic redundancy in yeast. We established several extremely accurate regression models for the prediction of DM fitness phenotypes. As DM fitness can easily be found experimentally, the importance of this model is not so much for generating predictions; that the predictions are so good simply shows that the model is working accurately. The biological significance of our models was also supported by the SM fitness model, as several features related to mitochondrial function were found to be important in predicting SM fitness. Thus, our experiments demonstrated the validity of using a machine learning approach to model fitness phenotypes and identify features that are important contributors to yeast SM and DM fitness, which are the basis for calculating genetic redundancy. We also discovered that SM fitness is a very reliable predictor of DM fitness, which was not previously known, and identified several GO annotations

that are highly accurate in predicting redundancy scores among a subset of gene pairs, implicating those processes as potentially important in retained redundancy.

While this study found some exciting results, there certainly are further avenues that could be explored with respect to prediction of redundancy. The features we used for prediction of redundancy among gene pairs were mainly different ways of utilizing the same features used in the SM fitness model, but there are many other interesting features specific to pairs of genes that may shed more light on the relationship between genes in a pair. Future work in this area would use DM-specific features such as the rate of synonymous and nonsynonymous substitution and gene co-expression values. Additionally, a classification-based model could be used to predict fitness and redundancy as categorical rather than continuous variables; for example, SM and DM fitness could be categorized as increased fitness/same fitness/decreased fitness with respect to the wild type, and genetic interactions could be represented as negative epistasis/positive epistasis (redundancy)/no gene interaction. Approaching fitness and redundancy as a classification problem would potentially allow for more accurate predictions as well as important features that are generalizable to specific, biologically relevant states of fitness and genetic interaction, rather than specific features important only to a few gene pairs along a continuum of values. It would also be beneficial to incorporate yeast fitness data from additional conditions, such as increased temperature or nutrient limitation, to identify redundancy among genes that are only conditionally observed.

On the computational side, additional model hyper-parameter optimization should be conducted using grid search. Additionally, the algorithm used to build this model (SVR) is somewhat limited in terms of parsing individual feature contributions to the model as a whole and to specific gene pairs. A decision tree-based model could be built, which would allow for a

more nuanced investigation of features that are important in predicting redundancy generally and in each gene pair instance. Analysis of features important in correctly predicted and mis-predicted gene pairs would be expected to uncover features with biological significance in genetic redundancy and allow us to further optimize the model, for example by removing features that consistently contribute to mis-predictions.

We here learned that a regression-based machine learning approach can be used to identify features that are biologically relevant in predicting fitness phenotypes and genetic redundancy in yeast. Single mutant fitness was unexpectedly important in predicting double mutant fitness, with a correlation suggesting that single mutants with low fitness are less likely to have gene interactions in general. Finally, contrary to expectations, we determined that in standard yeast growth conditions (at 26°C on rich medium) it is difficult to predict genetic redundancy as a continuous variable; future predictive analyses may benefit from the inclusion of fitness data from additional conditions and approaching genetic redundancy as one of several categorical types of gene interactions.

3.5 Materials and Methods

3.5.1 Data source

Saccharomyces cerevisiae fitness data for single and double mutants (grown at 26°C on Yeast Extract Peptone Dextrose medium, as described by Kuzmin et al. (2014)) were obtained from Costanzo et al. (2016). For predictive modeling, data from five general categories were collected for each gene: functional annotations; evolutionary properties; protein sequence properties; gene and protein expression; and network properties (**Supplemental Data**).

Functional annotation data were obtained from the Saccharomyces Genome Database (SGD, yeastgenome.org) and included gene ontology (GO) terms (v20170913), GO slims (v20170114), protein domains (v201610), and biochemical pathways (v. 15). From the protein domain data, the PANTHER domain annotations were used. Evolutionary data from SGD comprised the number of orthologs for each gene among bacteria and among genomes throughout the domain eukarya, with a particular focus on fungi (v201609). Protein properties data obtained from SGD included subcellular localization (Huh et al. 2003); molecular weight, amino acid length, and isoelectric point (v20160916); and regulatory protein binding levels (Venters et al. 2011). Post-translational modifications (PTMs) were from iPTMnet (Huang et al. 2018) release 5.0. Expression data from SGD included protein abundance in wild type (WT) yeast (Huh et al. 2003), protein abundance in WT and mutant yeast under stress conditions (Chong et al. 2015), mRNA half-lives (Geisberg et al. 2014), and gene over- or underexpression in stress conditions (Waern and Snyder 2013). Network property data comprised physical interactions (v20170114).

3.5.2 Features for single mutant fitness predictions

For single mutant (SM) fitness predictions, the features are listed in **Supplemental Data**. Most of the data were processed in each of two ways: as binary (presence/absence) and as continuous (total number) data. As an example, each gene had a binary feature value for each GO term, if the gene had that annotation or not, and a continuous feature value for all of the GO terms collectively, the number of GO annotations for the gene. Data processed in this manner included the following categories: GO terms, GO slim terms, PANTHER protein domains, biochemical pathways, species with a homologous gene, subcellular localization, PTMs, physical

interactions, overexpression in stress conditions, and underexpression in stress conditions. An additional continuous feature, differential expression in stress conditions, combined the over- and underexpression terms. Each PTM was also represented by a continuous value for each gene comprising the number of sites at which the protein product had that PTM. Continuous features with no binary feature counterparts included molecular weight, amino acid length, isoelectric point, protein abundance, mRNA half-life, and level of regulatory protein binding.

3.5.3 Features for double mutant and genetic redundancy predictions

The majority of features for double mutants (DMs) were derived from SM gene pair features, and the manner in which these were calculated depended on the type of SM feature (**Supplemental Data**). Categories of gene properties processed as binary for SMs include GO terms, GO slim terms, PANTHER protein domains, biochemical pathways, species with a homologous gene, subcellular localization, PTMs, physical interactions, and differential expression in stress conditions. These gene properties processed as binary for SMs were used to generate both continuous and categorical features for DMs. For each gene property category, continuous DM features were generated by determining the percent and number of overlapping properties between each gene pair. For each gene property within each category, categorical DM features were generated by taking the total number of genes in the pair with that property (0, 1 or 2). Additional DM features were generated from continuous SM features by taking the difference, average, maximum, minimum, and sum of the two SM values in a DM pair. This same processing was applied to SM fitness values for use in the DM model as detailed below.

3.5.4 Definition of genetic redundancy

From the 26°C datasets of Constanzo et al. (2016), we identified a subset of paralogous gene pairs (defined as gene pairs with a BLASTP Expect-value < 1×10^{-5} ; Altschul et al. 1990) for which fitness data were available for the DM and both corresponding SMs. DMs in this subset were used in modeling DM fitness and redundancy scores. We define genetic redundancy (*R)* as a state where the difference in DM fitness compared to WT is greater than the sum of the difference between both corresponding SMs compared to WT, i.e., there is a synergistic rather than additive effect of the double mutations:

(Eq. 1) $$R = (F_{WT} - F_{DM}) - (F_{WT} - F_{SM1}) - (F_{WT} - F_{SM2})$$

where *F* represents the fitness value of the indicated genotype. The mutant fitness values were normalized (*F'*) relative to the fitness of WT. Equation 1 was thus simplified to:

(Eq. 2) $$R = (1 - F'_{DM}) - (1 - F'_{SM1}) - (1 - F'_{SM2})$$

R was calculated for the DMs in our dataset to generate continuous redundancy scores.

3.5.5 Machine learning models

Predictive models of SM fitness, DM fitness, and genetic redundancy were implemented in the ScikitLearn machine learning package in Python (Pedregosa et al. 2011). Support Vector Regression (SVR, Drucker et al. 1997) was used to build regression models to predict fitness and redundancy scores as continuous values. For building a model, the gene (for the SM fitness

model) or gene pair (for the DM fitness and redundancy models) instances were randomly divided into a training set (90%) for model building and a testing set (10%) that was withheld from model building for independent validation. During training, 10-fold cross-validation was used to build the models, meaning that a total of 10 models were built with the training data; with each model, 90% of the training set instances were used for model building and the other10% to were used to cross-validate the performance of the model. After cross-validation, the trained model was then applied to the testing set. Performance results are reported from cross-validation and from the testing set.

APPENDIX

Supplemental Information

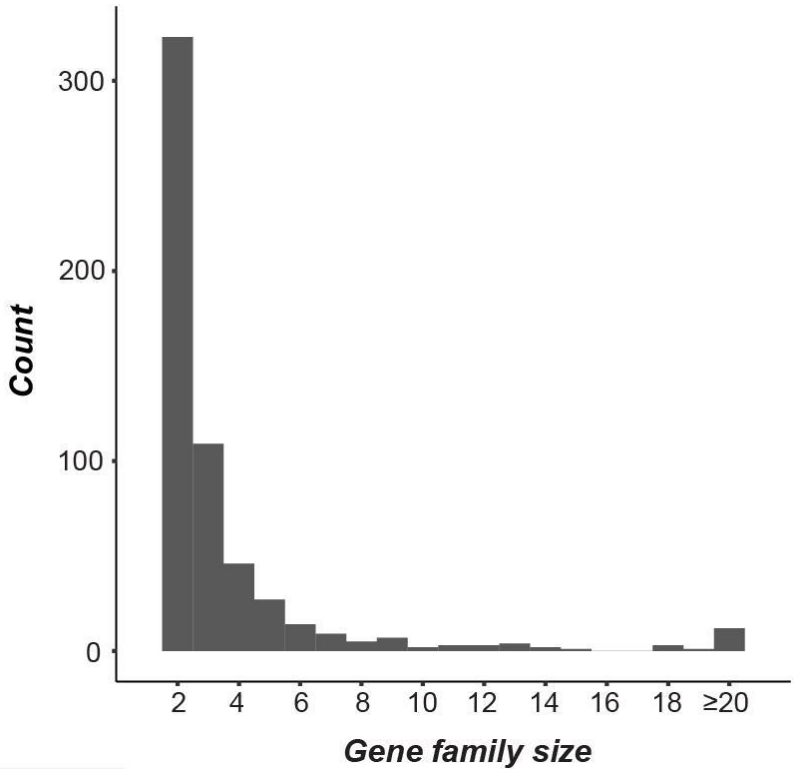

Figure 3.S1: Distribution of gene family sizes.
Distribution of gene family sizes for the paralogs in our dataset. The paralogous gene set comprises 2287 unique genes in 571 gene families. Forty-three percent of the gene families have more than two members, meaning that 61% of the paralogous genes in our dataset have three or more paralogs. This conflicts with results in the Saccharomyces Genome Database, which shows only three confirmed gene families in our dataset with more than two members, likely due to our use of a sequence similarity-based definition of paralogs.

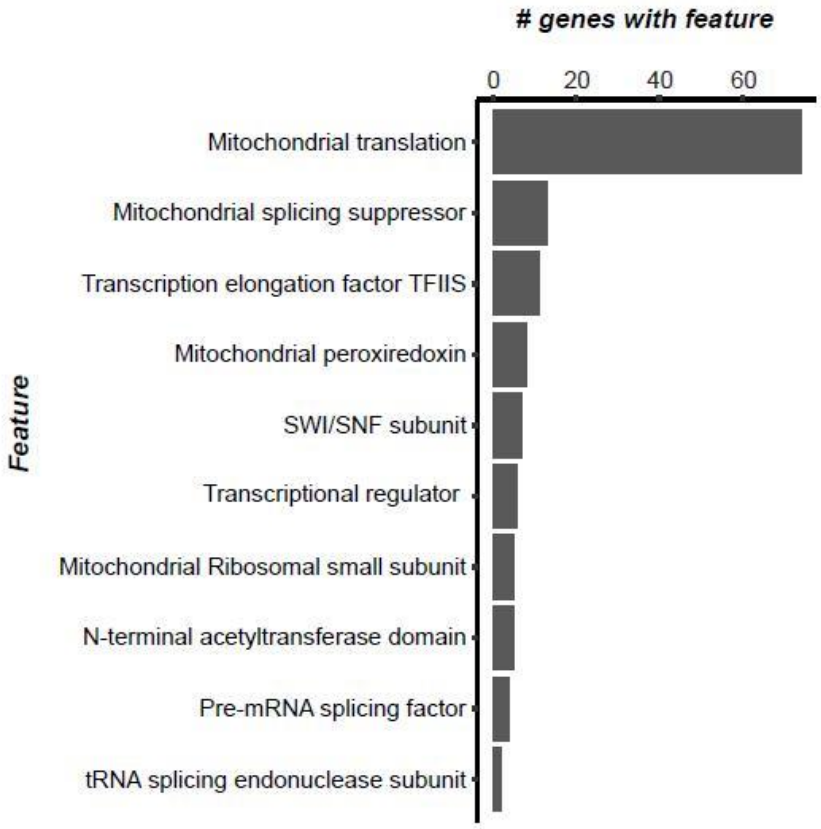

Figure 3.S2: Feature importance scores for SM fitness model.
Number of genes having each of the ten most important features in predicting SM fitness. Mitochondrial translation (ranked 2nd) was a Gene Ontology (GO) term; N-terminal acetyltransferase domain (ranked 8th) is a PANTHER protein domain; all others represent interaction with a gene which has been given a simplified descriptor. In order from rank 1-10 the gene names are: MRPS12, SEN2, MSS51, SNF11, CDC40, DST1, PRX1, and OPI1.
Many of the gene interaction features in our dataset are sparse, meaning few of the genes have these interactions and they were not expected to be particularly informative in ML.

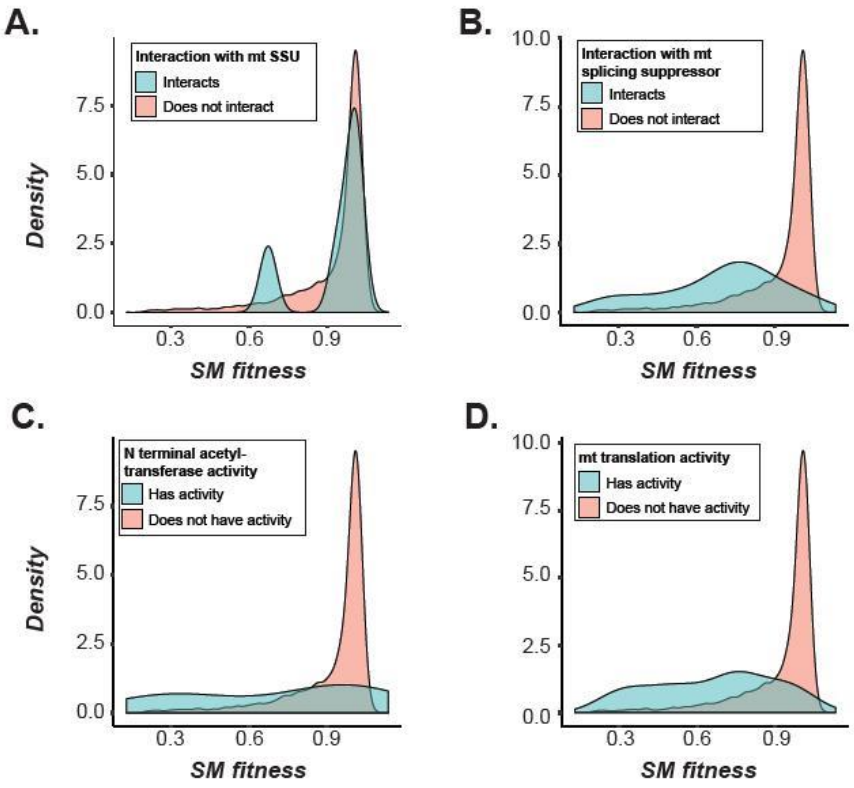

Figure 3.S3: Distribution of SM fitness values among gene pairs with selected annotations. Distribution of SM fitness values among genes with and without the annotations and interactions shown to be important in predicting SM fitness (see **Figure 3.2C** and **Figure 3.S2**). In general, genes without the annotations and interactions had lower SM fitness values.

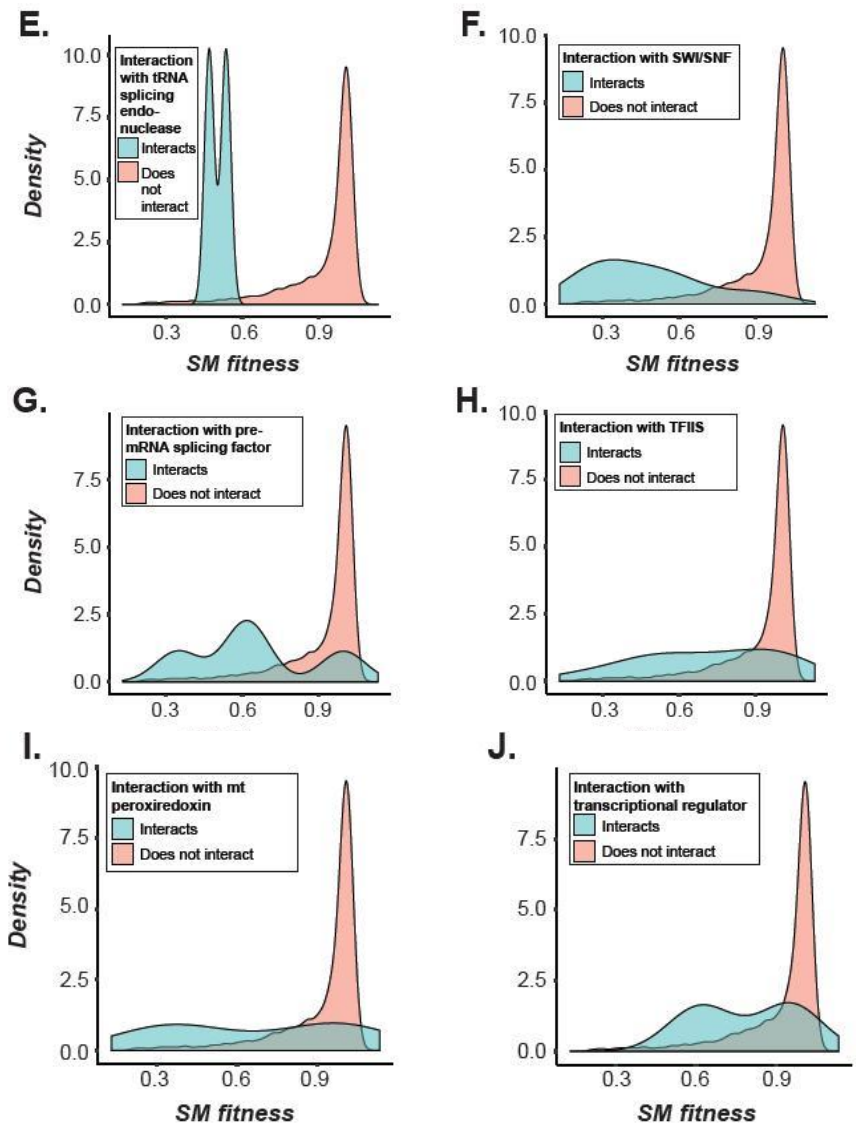

Figure 3.S3 cont'd
Distribution of SM fitness values among genes with and without the annotations and interactions shown to be important in predicting SM fitness (see **Figure 3.2C** and **Figure 3.S2**). In general, genes without the annotations and interactions had lower SM fitness values.

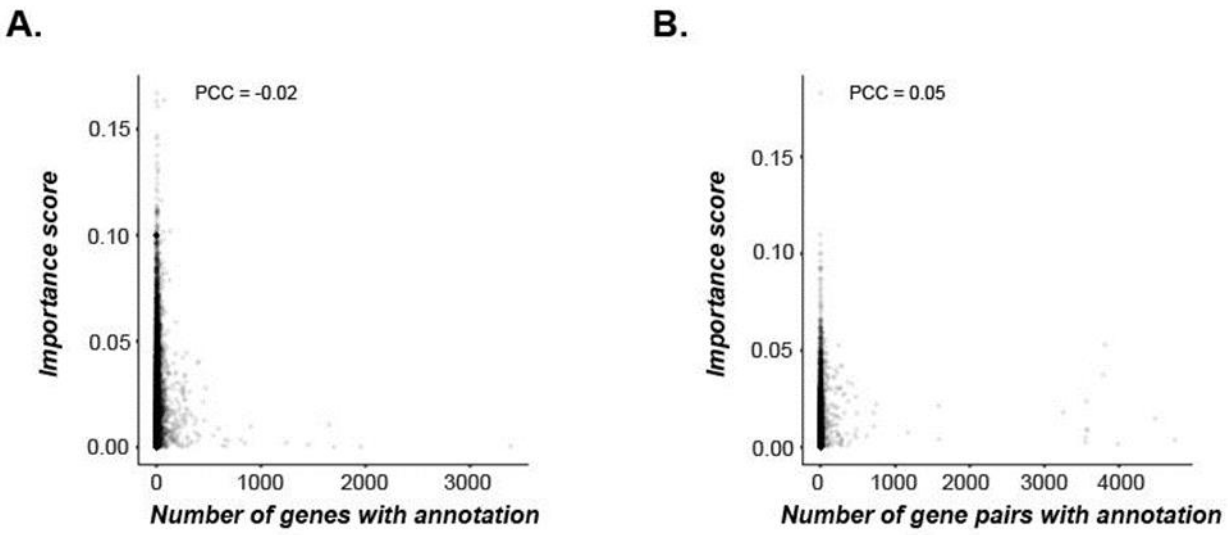

Figure 3.S4: Feature importance scores vs. number of gene pairs with annotation.
(*A*) For each binary feature in the SM fitness model (e.g., GO term or protein domain annotation; interaction with a specific gene), comparison of the feature importance score with the number of genes having that annotation or interaction. (*B*) For features in the DM fitness model corresponding to the features shown in *A*, comparison of the feature importance score with the number of gene pairs in which both genes have that annotation or interaction. In both cases, there is no correlation, indicating that feature importance is not simply a function of the prevalence of a particular feature in our dataset.

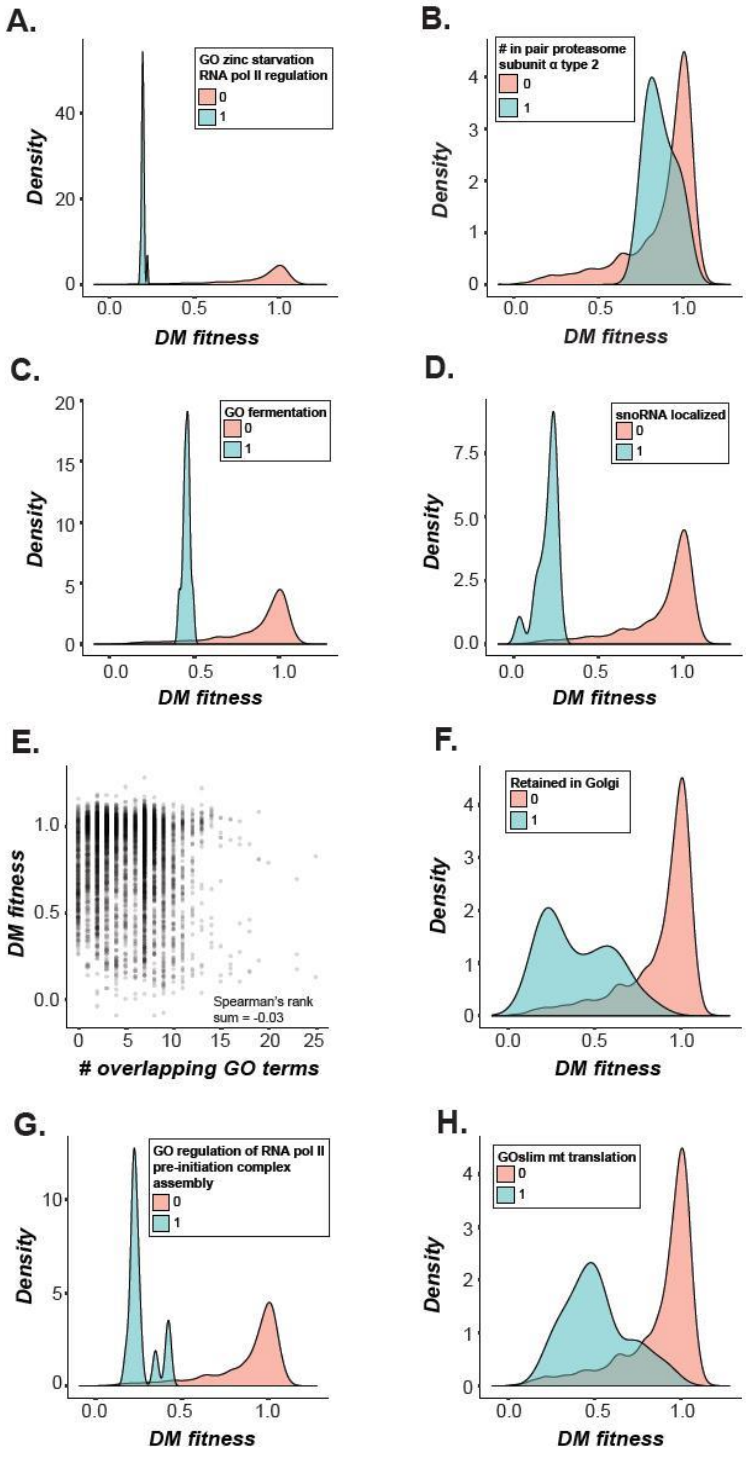

Figure 3.S5: Distribution of DM fitness values among gene pairs with selected annotations. (*A-D, F-H, J*) Distribution of DM fitness values among genes with and without the annotations and interactions shown to be important in predicting DM fitness (see **Figure 3.3C**). In general, gene pairs in which one gene without the annotations and interactions had lower SM fitness

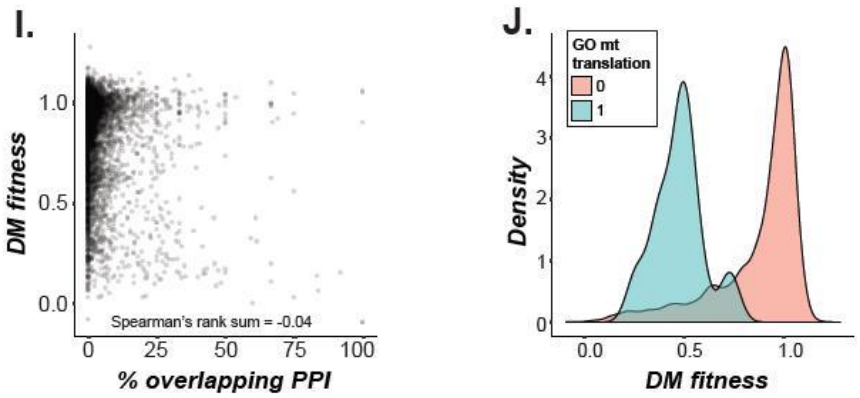

Figure 3.S5 cont'd
values. (*E*) Distribution of DM fitness values vs. number of overlapping GO terms and (*I*) percent overlapping protein-protein interactions among genes in a pair. There was no significant relationship between DM fitness and either of those features, contrary to our expectations.

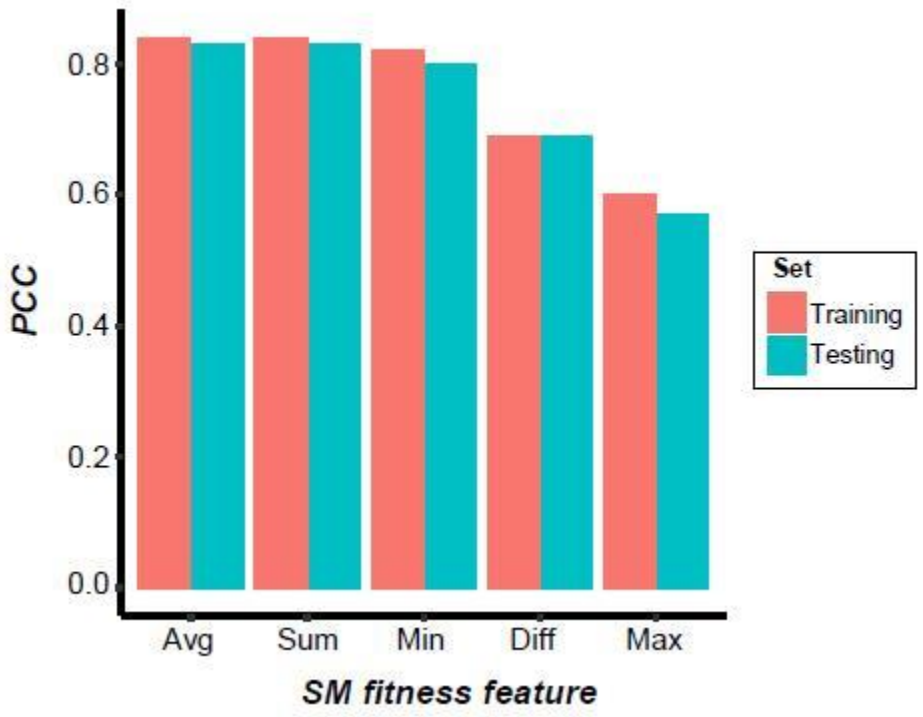

Figure 3.S6: Performance of DM fitness models built with single features.
PCC scores representing the performance of prediction of DM fitness using each of the SMF features separately. These results are consistent with the correlation strength between DM fitness and each SMF feature.

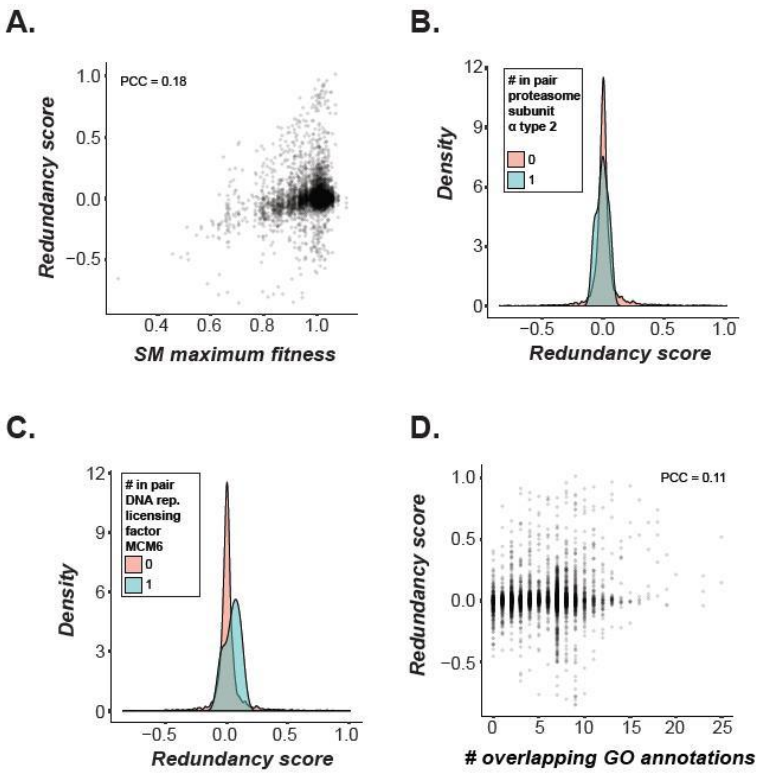

Figure 3.S7: Distribution of redundancy scores among gene pairs with selected annotations. (*A*) Distribution of redundancy scores vs. maximum SM fitness for genes in a pair; (*D*) vs. number of overlapping GO terms. Neither of the correlations were very strong, highlighting the importance of a model combining multiple features with weak predictive power to generate more accurate predictions. (*B*) and (*C*) Distribution of redundancy scores among gene pairs with 0, 1 or 2 genes in the pair having an annotation shown to be important in predicting redundancy (see **Figure 3.7C**).

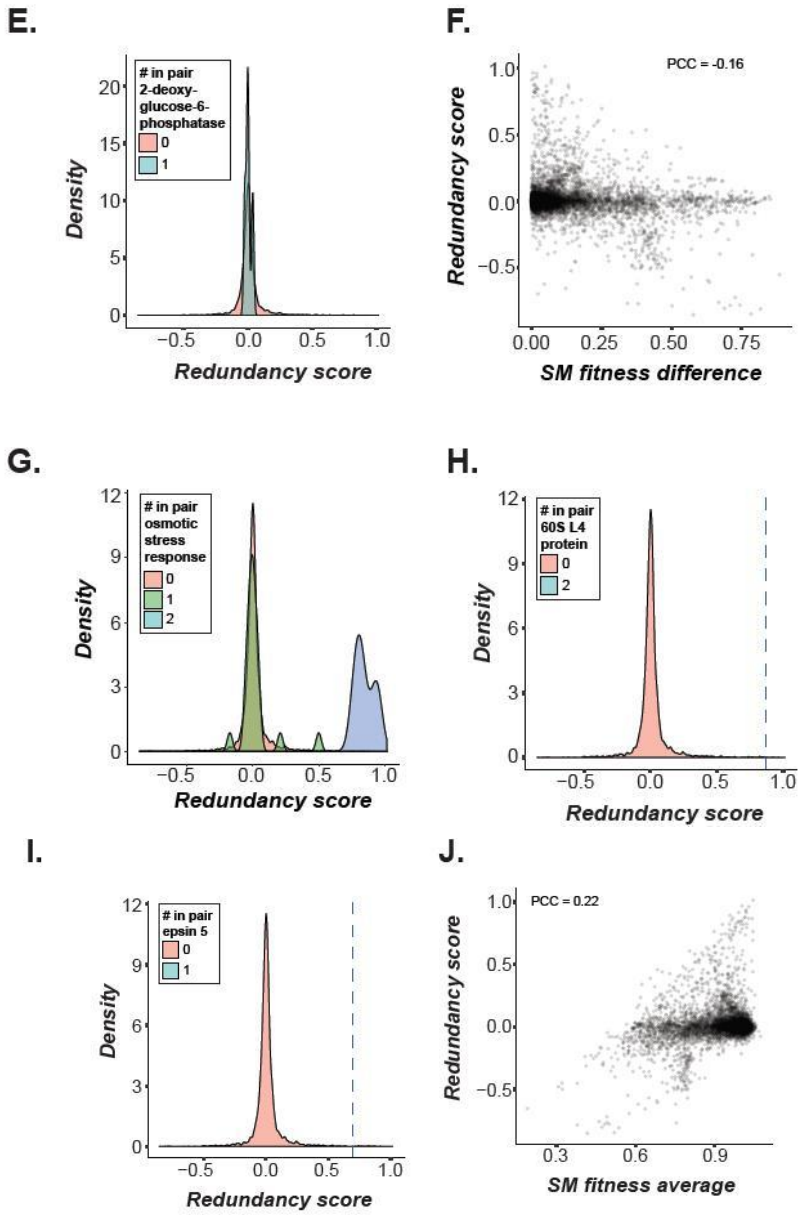

Figure 3.S7 cont'd
(*E*) and (*G-I*) Distribution of redundancy scores among gene pairs with 0, 1 or 2 genes in the pair having an annotation shown to be important in predicting redundancy (see **Figure 3.7C**). There were few generalizable patterns, although some features (*G-I*) did show a very small number of gene pairs which had the annotation and a high redundancy score, likely leading to the high importance value assigned to those features. (*F*) Distribution of redundancy scores vs. difference in SM fitness between genes in a pair and (*J*) vs. average SM fitness for genes in a pair. Neither of the correlations were very strong, highlighting the importance of a model combining multiple features with weak predictive power to generate more accurate predictions.

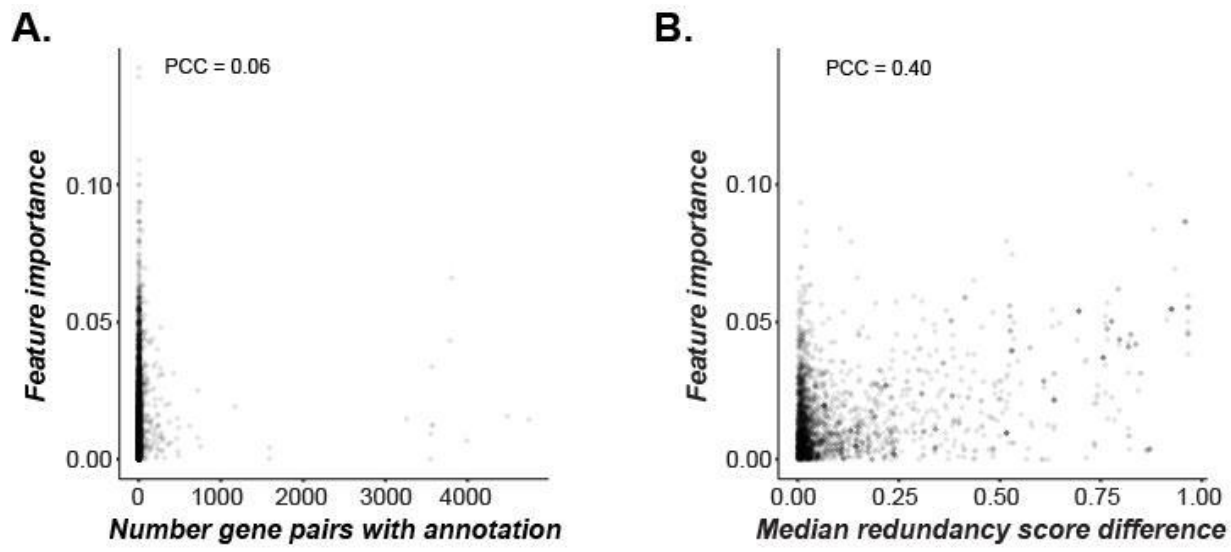

Figure 3.S8: Feature importance scores vs. number of gene pairs with annotation and difference in redundancy score among genes in a pair.
(*A*) For features in the redundancy model derived from binary features (e.g., number of genes in a pair with a GO term or protein domain annotation), comparison of the feature importance score with the number of gene pairs in which both genes have that annotation or interaction. There is no correlation, indicating that feature importance is not simply a function of the number of genes in the dataset with a particular annotation. (*B*) For features shown in (*A*), median redundancy score was calculated among gene pairs in which both genes had the annotation and among gene pairs in which one or neither gene had the annotation; the difference in median values was then compared with feature importance scores. There was some correlation between feature importance and the median redundancy score difference, indicating that, as would be expected, feature importance was affected by the ability of an individual feature to differentiate between high and low redundancy scores.

CHAPTER 4: CONCLUSIONS

In studying genetic redundancy over the last several years, I have come to appreciate it for the nuanced and evolutionarily fascinating phenomenon that it is. The Shiu Lab, together with Jeffrey Conner at Kellogg Biological Station and Patrick Krysan at University of Wisconsin-Madison, is currently conducting experiments to dig deeper into these nuances, defining genetic redundancy as a continuous rather than binary trait. They are accomplishing this by measuring lifetime fitness of hundreds of single and double mutant trios in Arabidopsis, both in growth chambers and in the field. The Shiu Lab has also collaborated with the Kramer Lab at MSU, using their Center for Advanced Algal and Plant Phenotyping to conduct comprehensive photosynthetic phenotyping of some of our mutants, generating truly staggering amounts of data. With these quantitative phenotypes, degrees of genetic redundancy among gene pairs can be calculated, producing a veritable treasure trove of data for some lucky student(s) to analyze. I look forward to seeing the next iteration of the Arabidopsis redundancy model, using genetic redundancy scores derived from this project to increase the accuracy of predictions and uncover additional insights into the evolutionary underpinnings of redundancy.

When I joined the Shiu Lab, its graduate student population was at something of an inflection point. One had recently finished up and moved on, and two more would do the same within the next six months. I had the privilege to work alongside two other grad students for another couple of years before they graduated too, and in that same year, my last one, we gained two new students. Every scientist in academia is familiar with this— frequent turnover as students come and go, with all the new ideas and feelings of excitement, loss, joy, and sadness that brings. In just a few short years, largely driven by graduate students, the lab added several areas of research to its repertoire and began mastering new, more powerful machine learning methods than the ones used and discussed here. It’s been thrilling and humbling to see the pace

at which computational biology is moving, and I thank new and future graduate students everywhere for showing me that there will always be bright and enthusiastic new people just waiting for a chance to push the limits of what's possible.

REFERENCES

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. J Mol Biol. 215(3):403–410.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 57(1):289–300.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis Information Resource: Making and Mining the "Gold Standard" Annotated Reference Plant Genome. Genesis. 53(8):474–485.

Birchler JA, Veitia RA. 2010. The Gene Balance Hypothesis: implications for gene regulation, quantitative traits and evolution. New Phytol. 186(1):54–62.

Blanc G, Wolfe KH. 2004. Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution. Plant Cell. 16(7):1679–1691.

Bolle C, Huep G, Kleinbölting N, Haberer G, Mayer K, Leister D, Weisshaar B. 2013. GABI-DUPLO: A collection of double mutants to overcome genetic redundancy in Arabidopsis thaliana. Plant J. 75(1):157–171.

Botstein D, Fink GR. 1988. Yeast: An experimental organism for modern biology. Science. 240(4858):1439–1443.

Bouché N, Bouchez D. 2001. Arabidopsis gene knockout: phenotypes wanted. Curr Opin Plant Biol. 4(2):111–117.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 422(6930):433–438.

Brandão MM, Dantas LL, Silva-Filho MC. 2009. AtPIN: Arabidopsis thaliana Protein Interaction Network. BMC Bioinformatics. 10(454).

Briggs GC, Osmont KS, Shindo C, Sibout R, Hardtke CS. 2006. Unequal genetic redundancies in Arabidopsis - a neglected phenomenon? Trends Plant Sci. 11(10):492–498.

Brookfield J. 1992. Can genes be truly redundant? Curr Biol. 2(10):553–554.

Brookfield JFY. 1997. Genetic redundancy. In: Advances in Genetics vol. 36. p. 137–155.

Chen H-W, Bandyopadhyay S, Shasha DE, Birnbaum KD. 2010. Predicting genome-wide redundancy using machine learning. BMC Evol Biol. 10(357).

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, et al. 1998. SGD: Saccharomyces Genome Database. Nucleic Acids Res. 26(1):73–79.

Chong YT, Koh JLY, Friesen H, Duffy K, Cox MJ, Moses A, Moffat J, Boone C, Andrews BJ. 2015. Yeast proteome dynamics from single cell imaging and automated analysis. Cell. 161(6):1413–1424.

Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, et al. 2016. A global genetic interaction network maps a wiring diagram of cellular function. Science. 353(6306).

DeSmet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van De Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci U S A. 110(8):2898–2903.

Dobzhansky T. 1946. Genetics of natural populations. XIII. recombination and variability in populations of Drosophila pseudoobscura. Genetics. 31(May):269–290.

Drucker H, Surges CJC, Kaufman L, Smola A, Vapnik V. 1997. Support vector regression machines. Adv Neural Inf Process Syst. 9:155–161.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci U S A. 112(27):8362–8366.

El-Brolosy MA, Kontarakis Z, Rossi A, Kuenne C, Günther S, Fukuda N, Kikhi K, Boezio GLM, Takacs CM, Lai SL, et al. 2019. Genetic compensation triggered by mutant mRNA degradation. Nature. 568(7751):193–197.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44(D1):D279–D285.

Gabriel ML. 1960. Primitive Genetic Mechanisms and the Origin of Chromosomes. Am Nat. 94(877):257–269.

Geisberg J V., Moqtaderi Z, Fan X, Ozsolak F, Struhl K. 2014. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. Cell. 156(4):812–824.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, et al. 2002. Functional profiling of the Saccharomyces cerevisiae genome. Nature. 418:387–391.

Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li W, Ogawa M, Yamauchi Y, Preston J, Aoki K, et al. 2008. The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J. 55(3):526–542.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40(D1):D1178–D1186.

Guan Y, Dunham MJ, Troyanskaya OG. 2007. Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics. 175(2):933–943.

Guckes KR, Cecere AG, Wasilko NP, Williams AL, Bultman KM, Mandel MJ, Miyashiro T. 2019. Incompatibility of vibrio fischeri strains during symbiosis establishment depends on two functionally redundant hcp genes. J Bacteriol. 201(19):1–14.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H. 2008. Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. Plant Physiol. 148(2):993–1003.

Haruta M, Burch HL, Nelson RB, Barrett-Wilt G, Kline KG, Mohsin SB, Young JC, Otegui MS, Sussman MR. 2010. Molecular Characterization of Mutant Arabidopsis Plants with Reduced Plasma Membrane Proton Pump Activity. J Biol Chem. 285(23):17918–17929.

Hsu PY, Calviello L, Wu H-YL, Li F-W, Rothfels CJ, Ohler U, Benfey PN. 2016. Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. Proc Natl Acad Sci U S A. 113(45):E7126–E7135.

Huang H, Arighi CN, Ross KE, Ren J, Li G, Chen SC, Wang Q, Cowart J, Vijay-Shanker K, Wu CH. 2018. IPTMnet: An integrated resource for protein post-translational modification network discovery. Nucleic Acids Res. 46(D1):D542–D550.

Huh, K. W, Falvo, V. J, Gerke, C. L, Carroll, S. A, Howson, W. R, et al. 2003. Global analysis of protein localization in budding yeast. Nature. 425(6959):686–691.

Hunter R. L., Markert C. L. 1957. Histochemical Demonstration of Enzymes Separated by Zone Electrophoresis in Starch Gels. Science. 125(3261):1294–1295.

Jammes F, Song C, Shin D, Munemasa S, Takeda K, Gu D, Cho D, Lee S, Giordo R, Sritubtim S, et al. 2009. MAP kinases MPK9 and MPK12 are preferentially expressed in guard cells and positively regulate ROS-mediated ABA signaling. Proc Natl Acad Sci U S A. 106(48):20520–20525.

Jiang W, Liu Y, Xia E, Gao L. 2013. Prevalent Role of Gene Features in Determining Evolutionary Fates of Whole-Genome Duplication Duplicated Genes in Flowering Plants. Plant Physiol. 161(4):1844–1861.

Kempin SA., Savidge B, Yanofsky MF. 1995. Molecular Basis of the cauliflower Phenotype in Arabidopsis. Science. 267(5197):522–525.

Khan NA, Haynes RH. 1972. Genetic redundancy in yeast: Non-identical products in a polymeric gene system. MGG Mol Gen Genet. 118(3):279–285.

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K. 2007. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J. 50(2):347–363.

Klappenbach JA, Dunbar JM, Schmidt TM. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol. 66(4):1328–1333.

Komili S, Farny NG, Roth FP, Silver PA. 2007. Functional Specificity among Ribosomal Proteins Regulates Gene Expression. Cell. 131(3):557–571.

Kosetsu K, Matsunaga S, Nakagami H, Colcombet J, Sasabe M, Soyano T, Takahashi Y, Hirt H, Machida Y. 2010. The MAP kinase MPK4 Is Required for Cytokinesis in Arabidopsis thaliana. Plant Cell. 22(11):3778–3790.

Krysan PJ, Jester PJ, Gottwald JR, Sussman MR. 2002. An Arabidopsis mitogen-activated protein kinase kinase kinase gene family encodes essential positive regulators of cytokinesis. Plant Cell. 14(5):1109–1120.

Kuzmin E, Sharifpoor S, Baryshnikova A, Costanzo M, Myers CL, Andrews BJ, Boone C. 2014. Chapter 10: Synthetic Genetic Array Analysis for Global Mapping of Genetic Networks in Yeast. In: Yeast Genetics: Methods and Protocols. Vol. 1205. p. 143–168.

Lacroute F, Piérard A, Grenson M, Wiame JM. 1965. The biosynthesis of carbamoyl phosphate in Saccharomyces cerevisiae. J Gen Microbiol. 40(1):127–142.

Last RL, Bissinger PH, Mahoney DJ, Radwanski ER, Fink GR. 1991. Tryptophan mutants in Arabidopsis: The consequences of duplicated tryptophan synthase β genes. Plant Cell. 3(4):345–358.

Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY. 2010. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nat Biotechnol. 28(2):149–156.

Lee SS, Lee RYN, Fraser AG, Kamath RS, Ahringer J, Ruvkun G. 2003. A systematic RNAi screen identifies a critical role for mitochondria in C. elegans longevity. Nat Genet. 33(1):40–48.

Lehti-Shiu MD, Shiu S-H. 2012. Diversity, classification and function of the plant protein kinase superfamily. Philos Trans R Soc Lond B Biol Sci. 367(1602):2619–2639.

Li C, Lin H, Chen A, Lau M, Jernstedt J, Dubcovsky J. 2019. Wheat VRN1, FUL2 and FUL3 play critical and redundant roles in spikelet development and spike determinacy. Dev. 146(14):1–11.

Li J, Yuan Z, Zhang Z. 2010. The cellular robustness by genetic redundancy in budding yeast. PLoS Genet. 6(11).

Liu B, Kanazawa A, Matsumura H, Takahashi R, Harada K, Abe J. 2008. Genetic redundancy in soybean photoresponses associated with duplication of phytochrome A gene. Genetics. 180(2):995–1007.

Lloyd J, Meinke D. 2012. A Comprehensive Dataset of Genes with a Loss-of-Function Mutant Phenotype in Arabidopsis. Plant Physiol. 158(3):1115–1129.

Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H. 2015. Characteristics of Plant Essential Genes Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes. Plant Cell. 27(8):2133–2147.

Lu Y, Savage LJ, Larson MD, Wilkerson CG, Last RL. 2011. Chloroplast 2010: A database for large-scale phenotypic screening of arabidopsis mutants. Plant Physiol. 155(4):1589–1600.

Luévano-Martínez LA, Appolinario P, Miyamoto S, Uribe-Carvajal S, Kowaltowski AJ. 2013. Deletion of the transcriptional regulator opi1p decreases cardiolipin content and disrupts mitochondrial metabolism in Saccharomyces cerevisiae. Fungal Genet Biol. 60:150–158.

Lynch M, Marinov GK. 2015. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A. 112(51):15690–15695.

Ma Z, Zhu P, Shi H, Guo L, Zhang Q, Chen Y, Chen S, Zhang Z, Peng J, Chen J. 2019. PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. Nature. 568(7751):259–263.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A. 102(15):5454–5459.

Matynia A, Anagnostaras SG, Wiltgen BJ, Lacuesta M, Fanselow MS, Silva AJ. 2008. A high through-put reverse genetic screen identifies two genes involved in remote memory in mice. PLoS One. 3(5).

McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R. 2013. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res. 41(Web Server Issue):W597–W600.

Mendonca AG, Alves RJ, Pereira-Leal JB. 2011. Loss of Genetic Redundancy in Reductive Genome Evolution. PLoS Comput Biol. 7(2).

Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, Mcentee C, Kay SA, Chory J. 2007. The Diurnal Project: Diurnal and Circadian Expression Profiling, Model-based Pattern Matching, and Promoter Analysis. Cold Spring Harb Symp Quant Biol. 72:353–363.

Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, Lehti-Shiu MD, Last RL, Pichersky E, Shiu S-H. 2019. Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci U S A. 116(6):2344–2353.

Mortimer RK. 1969. Genetic Redundancy in Yeast. Genetics. 61(1):Supplement 329-334.

Mueller LA, Zhang P, Rhee SY. 2003. AraCyc: A Biochemical Pathway Database for Arabidopsis. Plant Physiol. 132(2):453–460.

Nowak MA, Boerlijst MC, Cooke J, Smith JM. 1997. Evolution of genetic redundancy. Nature. 388(6638):167–171.

Panchy N, Lehti-Shiu M, Shiu S-H. 2016. Evolution of Gene Duplication in Plants. Plant Physiol. 171(4):2294–2316.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 12:2825–2830.

Peters H, Wilm B, Sakai N, Imai K, Maas R, Balling R. 1999. Pax1 and Pax9 synergistically regulate vertebral column development. Development. 126(23):5399–5408.

Pickett FB, Meeks-Wagner DR. 1995. Seeing Double: Appreciating Genetic Redundancy. Plant Cell. 7(9):1347–1356.

Polevoda B, Brown S, Cardillo TS, Rigby S, Sherman F. 2008. Yeast Nα-terminal acetyltransferases are associated with ribosomes. J Cell Biochem. 103(2):492–508.

Pruitt KD, Last RL. 1993. Expression patterns of duplicate tryptophan synthase β genes in Arabidopsis thaliana. Plant Physiol. 102(3):1019–1026.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35(D1):D61–D65.

Rutter MT, Wieckowski YM, Murren CJ, Strand AE. 2017. Fitness effects of mutation: testing genetic redundancy in Arabidopsis thaliana. J Evol Biol.:1–12.

Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci. 104(20):1–6.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of Arabidopsis thaliana development. Nat Genet. 37(5):501–506.

Sen A, Madhivanan K, Mukherjee D, Claudio Aguilar R. 2012. The epsin protein family: Coordinators of endocytosis and signaling. Biomol Concepts. 3(2):117–126.

Seoighe C, Wolfe KH. 1999. Yeast genome evolution in the post-genome era. Curr Opin Microbiol. 2(5):548–554.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30(9):1312–1313.

Su S-H, Krysan PJ. 2016. A double-mutant collection targeting MAP kinase related genes in Arabidopsis for studying genetic interactions. Plant J. 88(5):867–878.

Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, Sandstrom R, Thurman RE, Neph S, Reynolds AP, et al. 2014. Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Rep. 8(6):2015–2030.

Sun Q, Zybailov B, Majeran W, Friso G, Olinares PDB, van Wijk KJ. 2009. PPDB, the Plant Proteomics Database at Cornell. Nucleic Acids Res. 37(D1):D969–D974.

Sundell D, Mannapperuma C, Netotea S, Delhomme N, Lin Y-C, Sjödin A, Van de Peer Y, Jansson S, Hvidsten TR, Street NR. 2015. The Plant Genome Integrative Explorer Resource: PlantGenIE.org. New Phytol. 208(4):1149–1156.

Thatcher JW, Shaw JM, Dickinson WJ. 1998. Marginal fitness contributions of nonessential genes in yeast. Proc Natl Acad Sci U S A. 95(1):253–257.

The Gene Ontology Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 45(D1):D331–D338.

The Gene Ontology Consortium, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, et al. 2000. Gene Ontology: tool for the unification of biology. Nat Genet. 25(1):25–29.

Thomas JH. 1993. Thinking about genetic redundancy. Trends Genet. 9(11):395–399.

Tischler J, Lehner B, Chen N, Fraser AG. 2006. Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. 7(8):1–13.

Tong AHY. 2001. Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. Science. 294(5550):2364–2368.

Vavouri T, Semple JI, Lehner B. 2008. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. Trends Genet. 24(10):485–488.

Venters BJ, Wachi S, Mavrich TN, Andersen BE, Jena P, Sinnamon AJ, Jain P, Rolleri NS, Jiang C, Hemeryck-Walsh C, et al. 2011. A Comprehensive Genomic Binding Map of Gene and Chromatin Regulatory Proteins in Saccharomyces. Mol Cell. 41(4):480–492.

Waern K, Snyder M. 2013. Extensive Transcript Diversity and Novel Upstream Open Reading Frame Regulation in Yeast. G3; Genes|Genomes|Genetics. 3(2):343–352.

Wang H, Ngwenyama N, Liu Y, Walker JC, Zhang S. 2007. Stomatal Development and Patterning are Regulated by Environmentally Responsive Mitogen-Activated Protein Kinases in Arabidopsis. Plant Cell. 19(1):63–73.

Wang Y, Li J, Paterson AH. 2013. MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. Bioinformatics. 29(11):1458–1460.

Weintraub H. 1993. The MyoD Family and Myogenesis: Redundancy, Networks, and Thresholds. Cell. 75:1241–1244.

Wilcoxon F. 1945. Individual Comparisons by Ranking Methods. Biometrics Bull. 1(6):80–83.

Wilson TJ, Lai L, Ban Y, Ge SX. 2012. Identification of metagenes and their Interactions through Large-scale Analysis of Arabidopsis Gene Expression Data. BMC Genomics. 13(237).

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 24(8):1586–1591.

Yao SG, Ohmori S, Kimizu M, Yoshida H. 2008. Unequal genetic redundancy of rice PISTILLATA orthologs, OsMADS2 and OsMADS4, in lodicule and stamen development. Plant Cell Physiol. 49(5):853–857.

Zhang J. 2012. Genetic Redundancies and Their Evolutionary Maintenance. In: Advances in Experimental Medicine and Biology 751. p. 279–300.