Revealing retention mechanisms of duplicated genes and surveying triple mutant genetic interaction space in yeast

by

Elena Kuzmin

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Molecular Genetics
University of Toronto

© Copyright by Elena Kuzmin 2016
Revealing retention mechanisms of duplicated genes and surveying triple mutant genetic interaction space in yeast

Elena Kuzmin

Doctor of Philosophy
Graduate Department of Molecular Genetics
University of Toronto
2016

Abstract

Recent genome sequencing efforts have revealed astounding genetic diversity between individuals in the human population. The current challenge is to assign functions to the thousands of genes in our genomes and to understand how individual differences in genes contribute to disease. Much progress has been made in identifying causal loci of single gene disorders for both common and rare genetic diseases. However, many phenotypes are influenced by more complex genetic interactions involving larger sets of genes and genetic variants. Thus, deciphering the genotype-phenotype relationship requires that we expand our focus beyond pairwise genetic interactions to include complex, higher-order genetic interactions involving more than two genes. In this thesis I used the simple model organism, the budding yeast, and a method for automated yeast genetics called Synthetic Genetic Array (SGA) analysis, to construct double and triple mutants, and to measure the resulting genetic interactions by using colony size as a proxy for cell fitness. I first applied this methodology to dissect and characterize the mechanisms of functional divergence of duplicated genes. I also applied it more generally to estimate the prevalence of complex genetic interactions in a eukaryotic cell. My results suggest that triple mutant genetic interactions occur frequently in the yeast genome and are therefore expected to have a major impact on our understanding of the relationship between genotype and phenotype.
This map of complex genetic interactions provides a unique dataset for studying how the content of our genomes is shaped and how this relates to complex disease.
Acknowledgments

First and foremost I would like to thank my supervisors Charlie Boone and Brenda Andrews. Their vision and dedication to scientific research is inspirational. I am also grateful for the opportunities presented to me by contributing to collaborative projects, book chapters, summer mentorship programs and developing undergraduate and graduate teaching course materials. Over these six years I grew as a scientist and a person in large part because of my relationships with them.

I am lucky to have worked closely with such talented individuals as Chad Myers, Michael Costanzo and Benjamin VanderSluis. Our regular meetings and discussions have contributed to the progress that I achieved. Chad’s guidance has been instrumental in the direction of the project. Michael’s knowledge of the biology pertaining to genetic interactions helped me greatly with data interpretation. Working together with Ben on data analysis has been a pleasure. Together we tackled many challenges and his humour has never failed to brighten up my day.

It was a privilege to be surrounded and work with other members of the Donnelly Centre and the Department of Molecular Genetics, especially my committee members Howard Lipshitz and Marc Meneghini, past and present members of the Boone and Andrews labs, especially, Anastasia Baryshnikova, Helena Friesen, Sara Sharifpoor, Kerry Andrusiak, Oren Kraus, Mike Cox, Mojca Mattiazzi-Usaj, Traver Hart, Guihong Tan, Erin Styles, Jolanda van Leeuwen, Vince Messier and the team of genetic network technicians who helped me a great deal throughout the project, namely Ermira Shuteriqi, Bryan-Joseph San Luis and Matej Usaj. A special thanks goes to Yiqun Chen for such great technical help with strain construction and screening. I am also grateful to Sondra Bahr for getting me started in the lab as a rotation student and always being there for me throughout my time here. My thesis writing days were immeasurably better by having shared an office with Mark Spensley and Mike Schertzberg, who lightened this tumultuous time.

Finally, I thank my family who supported me and had an unfafltering belief that I could do it, even when I had doubts; my husband together with whom we persevered through graduate school and tried to figure it all out; and my grandparents who inspired and motivated me and it is to them that I dedicate this thesis.
Table of Contents

Acknowledgments .................................................................................................................... iv

Table of Contents .................................................................................................................... v

List of Tables ........................................................................................................................... viii

List of Figures .......................................................................................................................... ix

List of Appendices .................................................................................................................. xii

Chapter 1 .................................................................................................................................. 1

1 Introduction ............................................................................................................................ 2
  1.1 Genotype to phenotype and missing heritability ............................................................. 2
  1.2 Defining genetic interactions ......................................................................................... 4
  1.3 Digenic interaction classes ............................................................................................. 6
  1.4 Yeast as model organism ............................................................................................... 9
  1.5 Mutant collections ......................................................................................................... 9
  1.6 Synthetic genetic array (SGA) analysis ......................................................................... 10
  1.7 Genome-wide genetic interaction studies ..................................................................... 14
  1.8 Mapping genetic interactions in other eukaryotes ....................................................... 15
  1.9 Higher-order genetic interaction studies ..................................................................... 17
  1.10 Gene duplication is important for network robustness and species evolution ............ 22
  1.11 Mechanisms of gene duplication ................................................................................. 22
  1.12 Segmental gene duplication ......................................................................................... 24
  1.13 Whole genome duplication ......................................................................................... 25
  1.14 Prevalence of gene duplication in eukaryotes ............................................................ 27
  1.15 Conceptual frameworks for evolution of duplicated genes ......................................... 28
  1.16 Experimental approaches to study fates of duplicated genes ..................................... 32
  1.17 Rationale ....................................................................................................................... 33

Chapter 2 .................................................................................................................................. 35

2 Retention of duplicated genes unravelled through complex genetic interactions ........... 36
  2.1 Introduction .................................................................................................................... 36
  2.2 Materials and Methods ................................................................................................. 36
2.2.1 Constructing a yeast paralog double mutant collection ........................................ 36
2.2.2 Constructing single mutant control query collection ........................................... 41
2.2.3 Synthetic genetic array (SGA) analysis technique for triple mutants .................... 42
2.2.4 Estimating single and double mutant fitness on standard growth conditions .......... 43
2.2.5 Quantitative fitness-based model of genetic interactions .................................... 44
2.2.6 Selecting strains for array miniaturization ......................................................... 44
2.2.7 Confirmation of trigenic interactions by tetrad and random spore analyses ............ 45
2.2.8 Estimating the effect of environmental stress conditions on single and double mutant fitness 45
2.2.9 Correlation with evolutionary and physiological features .................................... 48
2.2.10 Simulating paralog divergence ............................................................................ 49
2.2.11 Testing hypotheses about asymmetry, structural and functional overlap by modeling duplicate divergence .......................................................... 50

2.2.12 Results .................................................................................................................... 51
2.3.1 Strain construction and estimation of single and double mutant query fitness on standard growth conditions ................................................................. 51
2.3.2 Optimization of Synthetic Genetic Array (SGA) for triple mutants ....................... 52
2.3.3 Quantitative fitness-based model for trigenic interactions .................................... 53
2.3.4 Evaluation of control screens ............................................................................... 56
2.3.5 Array miniaturization for high-throughput screening of triple mutants ................. 57
2.3.6 Evaluation of reproducibility of genetic interactions ............................................ 63
2.3.7 Validation of trigenic interactions .......................................................................... 65
2.3.8 Triple mutant interaction fraction: functional divergence metric .......................... 67
2.3.9 SSD and WGD paralogs contribute differently to mutational and environmental robustness ........................................................................................................ 73
2.3.10 Conditional triple mutant interactions reveal functional relationships of duplicated genes 74
2.3.11 Paralog retention varies by cellular process ............................................................ 75
2.3.12 Physiological and evolutionary features correlate with triple mutant interaction fraction 76
2.3.13 Asymmetric evolution and retention of functional redundancy ............................ 78
2.3.14 Modeling paralog divergence ............................................................................... 80

2.4 Discussion .................................................................................................................. 82

Chapter 3 ......................................................................................................................... 85
3 Surveying triple mutant interaction space in yeast ............................................ 86
  3.1 Introduction ........................................................................................................... 86
  3.2 Materials and Methods .......................................................................................... 87
    3.2.1 Constructing a singleton double mutant collection with control query strains ....... 87
    3.2.2 Estimation of single and double mutant query fitness on standard growth conditions 92
    3.2.3 Triple mutant SGA ......................................................................................... 93
    3.2.4 Precision-recall analysis ................................................................................. 93
  3.3 Results .................................................................................................................. 94
    3.3.1 Gene pair selection ......................................................................................... 94
    3.3.2 Functional relation of double and triple mutant interactions ......................... 96
    3.3.3 Triple mutant interaction space is extensive ................................................ 101
    3.3.4 Genetic interaction network rewiring between distantly related species .......... 105
  3.4 Discussion ............................................................................................................. 107

Chapter 4 .................................................................................................................... 109

4 Thesis Summary and Future Directions ................................................................. 110
  4.1 Thesis Summary ..................................................................................................... 110
  4.2 Future Directions .................................................................................................. 112
    4.2.1 Enhancing the understanding of functional relationship of duplicated genes through conditional genetic interaction screens ................................................................. 112
    4.2.2 Empirically showing asymmetric evolution of paralogs due to structural entanglement 113
    4.2.3 Dissecting mechanisms of triple mutant interactions .................................... 114
    4.2.4 Understanding the contribution of higher order genetic interactions in human disease and as it relates to missing heritability .............................................................. 115

References .................................................................................................................... 116
List of Tables

Table 1.1 Prevalence of gene duplication in eukaryotes ......................................................... 27

Table 2.1. Environmental stress conditions for fitness assessment ...................................... 48

Table 2.2. Validation of $cln1A cln2A$ trigenic interactions by tetrad and random spore analyses ................................................................................................................................. 66

Table 3.1. Selection criteria for gene pairs for testing triple mutant space ......................... 89

Table 3.2. The extent of genetic interaction space .......................................................... 102
List of Figures

Figure 1.1. Multiplicative model for quantifying fitness-based genetic interactions .................. 5

Figure 1.2. Biological interpretation of genetic interactions ................................................... 8

Figure 1.3. Synthetic Genetic Array (SGA) Methodology ......................................................... 13

Figure 1.4. Biological interpretation of trigenic interactions .................................................... 21

Figure 1.5. Summary of mechanisms of gene duplication ....................................................... 23

Figure 1.6. Phylogeny and relative time of whole genome duplication (WGD) ......................... 26

Figure 1.7. Evolutionary fates of duplicated genes ................................................................. 28

Figure 1.8. Duplicate divergence by sub- or neo-functionalization ........................................... 31

Figure 2.1. Overview of query strain construction ................................................................. 37

Figure 2.2. Representative tetrad analysis for double mutant construction ......................... 41

Figure 2.3. Conditional SGA for estimating fitness in environmental stress conditions .............. 47

Figure 2.4. Quantification of fitness of query strains using high-density arrays ...................... 51

Figure 2.5. Optimization of selection steps for triple mutant SGA ......................................... 53

Figure 2.6. Comparison of τ to S-score approaches for quantifying trigenic interactions ............ 56

Figure 2.7. Evaluation of control screens .................................................................................. 57

Figure 2.8. Diagnostic array characteristics ............................................................................. 59

Figure 2.9. Technical comparison of array types ..................................................................... 61

Figure 2.10. Trigenic and digenic interaction frequency of preliminary paralog pairs across different array types ......................................................................................... 61

Figure 2.11. Functional analysis across array types .................................................................. 62
Figure 3.7. Relation of double and triple mutant interaction networks……………………..103

Figure 3.8. Pairwise interplay of digenic interaction network attributes in relation to the average trigenic interaction degree (residual) with partialling out the third variable……………………104

Figure 3.9. Triple mutant interactions of *S. cerevisiae* homologs of *S. pombe* SL pairs as query genes………………………………………………………………………………………………………106
List of Appendices

Appendix 1: List of constructed mutants

Appendix 2: Diagnostic array composition

Appendix 3: List of genetic interactions on standard conditions

Appendix 4: Fitness standard on standard and alternative growth conditions for queries

Appendix 5: List of genetic interactions on alternative growth conditions for queries

Appendix 6: Genetic interaction profile correlations

Appendix 7: Triple mutant interaction fraction

Appendix 8: List of trigenic interactions on alternative conditions
Chapter 1

Portions of this chapter have been adapted from:


I wrote or substantially edited all sections presented in this Chapter.
1 Introduction

1.1 Genotype to phenotype and missing heritability

Genome sequencing projects have revealed a massive catalogue of genes and extensive genetic diversity in a variety of organisms. The current challenge is to understand the functional role of each gene and how genes work together to modulate cellular processes, which will allow us to gain insight into the relationship between genotype and phenotype.

Systematic studies, such as the Human Genome Project, the Encode Project and the Roadmap Epigenomics Project, have made significant headway in characterizing the human genome (Lander et al. 2001; International Human Genome Sequencing 2004; Encode Project Consortium 2012; Roadmap Epigenomics et al. 2015). Subsequent systematic genetic mapping studies of single nucleotide polymorphisms (SNPs) discovered 1.42 million variants and developed tools to functionally annotate them facilitating our search for biomedically important genes relevant to disease diagnosis and therapy (Sachidanandam et al. 2001; Boyle et al. 2012). Despite substantial progress in identifying causal loci of single gene disorders for both common and rare genetic disease, many traits and disease states are more complex and are likely associated with many different genetic variants affecting numerous different genes (Zuk et al. 2012). Genome-wide association studies (GWAS) have identified thousands of loci carrying single nucleotide polymorphisms associated with hundreds of human traits and diseases, such as adult macular degeneration, Crohn’s disease, control of fetal haemoglobin, type 2 diabetes, autoimmune diseases, height, kidney disease, psychiatric disorders and others (Lander 2011).

Despite the advances in the identification of disease-associated loci, current GWAS can only account for up to ~ 20-25% of heritability, leaving the heritability for the majority of traits unexplained (Lander 2011). Importantly, it has been postulated that genetic interactions (non-additive combinatorial effects of genetic variants) could explain this ‘missing heritability’ and shed light on the genotype-phenotype problem (Zuk et al. 2012). For example, GWAS identified 71 loci associated with Crohn’s disease but only 21.5% of heritability could be explained assuming the additive nature of these loci (Franke et al. 2010). On the other hand, incorporating genetic interactions into this model has the potential to increase the explained component of heritability to 80% (Zuk et al. 2012). Consistent with these observations, large-scale studies of the genetic basis of a variety of traits in yeast segregants from a cross between two yeast strains
(BY and RM) showed that the contribution to heritability of two-locus interactions varied among traits but was as large as 50% (Bloom et al. 2013). Similarly, variation found in the genome of the wine strain, RM11-1a, results in the complexity underlying quantitative traits associated with chemical resistance and mitochondrial function (Ehrenreich et al. 2010). Several challenges remain in identifying genetic interactions and mapping sources of variation of human traits in patient cohorts, such as population size. For instance, up to 500,000 individuals would be required to identify interactions between two independent loci linked to Crohn’s disease (Zuk et al. 2012). Nonetheless, genetic interactions are becoming widely recognized as important biological phenomena for understanding the complexity of the genotype to phenotype problem as it relates to human disease.

Variation in phenotypic outcomes in different individuals is also caused by genetic determinants that act as modifiers (Hartman et al. 2001). A genetic modifier is a type of genetic interaction in which one genetic locus exerts an effect on the phenotype resulting from another locus (Kousi and Katsanis 2015). Individuals carrying the same primary mutations display drastic differences in phenotypic expression depending on their modifier loci. Some differences in disease expression are caused by one major modifier locus called a ‘panmodifier’ such as CCDC28B, which is associated with Bardet-Biedl syndrome (Badano et al. 2006). In addition, variability in penetrance/expressivity of a disease can also be due to multiple modifier loci. One example is cystic fibrosis, which overwhelmingly results from homozygosity for F508del mutation in the CFTR. However, major differences in lung function are seen in patients, and reflect functional variation in MBL2, TGFβ1, IL-8, IFRD1 and ENDRA, which are thought to affect disease severity by influencing the host’s tolerance to infection (Garred et al. 1999; Drumm et al. 2005; Vanscoy et al. 2007; Hillian et al. 2008; Darrah et al. 2010). Modifier loci are prevalent in human populations but knowledge is lacking regarding how variants interact to modulate phenotype in different individuals (Gibson 2011).

These findings are consistent with work in genetic model systems where strain specific differences in phenotypes are often attributed to genetic background. An analysis of the genomes of two laboratory yeast strains of *Saccharomyces cerevisiae*, Σ1278b and S288c, showed that while most (94%) deletion mutants have similar phenotypes (essentiality status is the same in both backgrounds), a small fraction (6%) exhibited a background-specific or conditional essentiality whereby loss-of-function mutations resulting in inviability did not represent
completely overlapping gene sets between the two strains (Dowell et al. 2010). Genetic background variation in Σ1278b also accounts for strain-specific growth programs that result in substantial morphological differences characterized by invasive growth, pseudohyphal growth and biofilm formation phenotypes (Ryan et al. 2012b). Such conditional phenotypes result from the interplay of multiple modifier loci, and suggest that complex modifiers may underlie the genetic basis of physiological states found in natural populations.

1.2 Defining genetic interactions

Genetic interactions have long been recognized as playing a fundamental role in shaping phenotypes of living organisms. They are thought to underlie such diverse biological phenomena as the evolution of sex and recombination, speciation and as mentioned above, complex disease manifestations (Kondrashov 1995; Keightley and Otto 2006; Altshuler et al. 2008; Stern 2013). The concept of a genetic interaction was recognized over a century ago, when William Bateson first coined the term ‘epistasis’ while investigating Mendelian segregation ratios and the deviation from such simple ratios leading to novel phenotypes (Bateson 1909). The term itself translates into ‘standing upon’ which he used to describe an observation in rabbits and mice in which the effect of a mutation influencing fur colour was masked by another mutation (Bateson 1909). In 1918 Ronald Fisher attempted to map discrete genetic data such as segregation of alleles onto a continuous range of quantified traits and noted that some quantitative phenotypes are unpredictable from simple addition of the effects of two loci and referred to such non-additive effects as ‘epistacy’ (Fisher 1918). Thus, the term ‘epistasis’ has distinct meanings for Mendelian vs. quantitative geneticists but a more common all-encompassing term ‘genetic interaction’ has arisen to denote any phenotype that is unexpected given the known combined effects of individual mutations (Phillips 1998).

The identification of genetic interactions relies on determining the deviation from an expected mutant phenotype. This seemingly simple task presents several challenges and various mathematical approaches have been developed to derive the expectation. Additive and multiplicative models have been among the most widely used approaches depending on whether the phenotype measurement is linear or logarithmic, respectively (Phillips 1998; Phillips 2000; Mani et al. 2008). The phenotype associated with a given genotype can be due to adding its individual single-locus effects, which would form the expected outcome and any deviation
would be called a ‘genetic interaction’. On the other hand, a more widely adopted neutrality function for fitness-based genetic interactions predicts the combined mutant fitness effects by relying on a multiplicative function, which formulates the expectation based on the product of the effects associated with individual mutant loci [Figure 1.1 (Mani et al. 2008; Costanzo 2012)].

\[
\varepsilon_{AB} = f_{AB} - (f_A f_B)
\]

**Figure 1.1. Multiplicative model for quantifying fitness-based genetic interactions.** The deviation from the expectation (\(\varepsilon\)) is captured by subtracting the product of observed single mutant phenotypes \((f_A f_B)\) from the observed double mutant phenotype \(f_{AB}\). The genetic interaction then falls on a gradient: there is no genetic interaction, when the observed and the expected double mutant fitness are equal, negative when the double mutant fitness is surprisingly lower than expected and positive when the observed double mutant fitness unexpectedly exceeds the predicted phenotype.
1.3 Digenic interaction classes

Based on the above definition, genetic interactions can be classified into two broad categories, referred to as negative and positive interactions. In the case of cell fitness, a negative genetic interaction is detected when a double mutant is less fit than expected based on the combined effects of individual mutations (Costanzo 2012). An extreme example of a negative genetic interaction is synthetic lethality, which occurs when two different mutations, each of which fails to alter cellular fitness as a single mutant, combine to cause inviability. This phenomenon was first observed in natural populations of *Drosophila pseudoobscura* in which recombination of genes by crossing-over exposed manifestations of recessive alleles (Dobzhansky 1946). Synthetic lethal interactions have been used to characterize various biological processes in many organisms (Sturtevant 1956; Novick and Botstein 1985; Bender and Pringle 1991) and theoretically can be used to study any gene of interest. Such interactions are of particular importance because they identify non-essential genes that impinge on the same essential process such as duplicated genes, functional homologs or members of parallel pathways (Guarente 1993) (Figure 1.2a). Mutations in essential genes that cause synthetic lethal interactions often involve genes that function in distinct steps of the same pathway essential for viability. In this case mutations in individual genes are not lethal but each mutation is hypomorphic, and the double mutant combination reduces pathway activity below a level required for cell viability (Figure 1.2b) (Guarente 1993; Finger and Novick 2000). Thus, negative genetic interactions reveal important functional connections between genes and the pathway architecture of a eukaryotic cell.

Positive interactions represent another broad class of genetic interactions. Two mutations interact positively when the fitness of a double mutant is less severe than expected from the product of phenotypes of the individual mutants. There are a number of different types of positive genetic interactions. For example, symmetric positive interactions are associated with single and double mutant phenotypes that are quantitatively indistinguishable, whereas asymmetric positive interactions describe cases in which the strength of phenotypic consequences varies among single and double mutants (Costanzo 2012). For example, genetic suppression is an asymmetric positive interaction that often involves genes encoding products that function in distinct pathways, where one pathway can compensate for defects in another (Guarente 1993). An example of a symmetric positive genetic interaction is ‘classic epistasis’, as defined by Bateson,
in which gene products are involved in successive steps of a biochemical pathway or gene regulatory cascade. In the case of the former, if a loss of function in one step produces a mutant phenotype, a mutation in another step will not exaggerate the effect, a pattern seen in regulatory pathways controlling amino acid biosynthesis (Figure 1.2c) (Hinnebusch and Fink 1983). An analogous interpretation extends to subunits of the same protein complex, where perturbations of one subunit results in dissolution and loss of function of the entire protein complex, rendering mutation of additional subunits of the same complex inconsequential (Figure 1.2e) (Drees et al. 2005; St Onge et al. 2007; Breslow et al. 2008). Positive interactions may also be seen in a pathway composed of a series of positive regulators. In this case, the mutant phenotype caused by perturbing the first pathway member may be suppressed by a gain of function in the second member, thus restoring the pathway activity (Guarente 1993). A loss of a negative regulator could also result in a positive interaction if combined with loss-of-function mutations in genes encoding components of the target pathway, particularly when up-regulation of the target gene is toxic to the cell (Figure 1.2d) (Baryshnikova et al. 2010b). Finally, it is important to note that the vast majority of both positive and negative interactions occur between pathways and protein complexes revealing broad phenotypic connections between functional modules and biological processes (Baryshnikova et al. 2010b; Costanzo et al. 2010; Bellay et al. 2011)
Figure 1.2. Biological interpretations of genetic interactions. Negative and positive genetic interactions occur in a variety of cellular contexts. Some examples of negative genetic interactions include duplicated genes, members of parallel pathways or functional homologs that provide back-up compensation. A) Deletion of genes involved in one of two parallel pathways does not influence fitness but a double perturbation in members of both pathways exerts a deleterious effect. B) Reduction of signaling output may be observed in a double mutant due to loss-of-function mutations in individual components of a pathway that reduce the level of a particular biological activity below a threshold required for survival. C) Positive genetic interactions may be observed when a loss-of-function perturbation of one gene in a pathway produces a mutant phenotype that is not exaggerated by mutation of another component of the same pathway (also called ‘classical epistasis’). D) Positive interactions may also result when a deletion of a negative regulator, which releases the inhibition of its effector resulting in cellular toxicity, but viability is restored upon a partial loss-of-function mutation in the effector. E) Finally, mutation of members of the same protein complex often results in a positive genetic interaction, if perturbation of any single component causes a loss-of-function phenotype. Wild-type scenario is highlighted with a grey box. A deleted gene is crossed-out; a conditional loss-of-function allele is depicted with a lower case letter.
1.4 Yeast as model organism

Yeast has been a workhorse for fundamental genetics research and is an ideal organism for the post-genomic era of experimental biology. The *Saccharomyces cerevisiae* genome was the first eukaryotic genome to be fully sequenced, paving the way for the functional characterization of all its genetic elements. The development of functional genomics in yeast was supported by the establishment of the first model organism database, the *Saccharomyces* Genome Database (SGD) (Botstein and Fink 2011). Ongoing sequencing efforts of a diverse set of *S. cerevisiae* strains have advanced the frontiers of ‘comparative genomics’ research enhancing our understanding of genome evolution and variation in natural populations (Liti et al. 2009; Engel and Cherry 2013). Moreover, ~1,000 yeast genes have orthologs in the human genome and a large fraction of these genes can functionally complement the cognate yeast mutations, enabling studies focused on understanding the functional effects of human genetic polymorphisms in a simple model organism (Kachroo et al. 2015). Thus, yeast has empowered the scientific community to answer complex questions using a tractable genetic system.

1.5 Mutant collections

The *S. cerevisiae* genome sequencing project revealed the existence of ~6,000 protein coding open reading frames (Goffeau et al. 1996). The challenge then became to understand the functional role of each gene and how genes work together to modulate different functions. A powerful way to understand gene function is by phenotyping mutants that lack a gene of interest (Winzeler et al. 1999). For this purpose the yeast community constructed a collection of gene deletion mutants in which every yeast open reading frame (ORF) was replaced with an antibiotic-selectable marker flanked by unique 20 bp sequence identifiers, which serve as molecular barcodes (Winzeler et al. 1999; Giaever et al. 2002). The PCR-mediated gene deletion strategy made use of yeast’s unique property of exhibiting a high rate of homologous recombination (Baudin et al. 1993; Wach et al. 1994; Lorenz et al. 1995). Each gene was deleted within a diploid strain and tested for viability in the haploid meiotic progeny, identifying ~1,000 essential genes and ~5,000 viable deletion mutants (Giaever et al. 2002). The strain-specific barcodes are particularly useful for the application of microarray- or sequencing-based measurement of the relative abundance of thousands of different mutants grown competitively in...
heterogeneous pools (Hillenmeyer et al. 2008; Douglas et al. 2012). In this manner, the deletion collection contributed to the functional characterization of the yeast genome.

For detailed phenotypic analysis of loss-of-function mutations in essential genes the community has generated libraries of conditional mutants. The collections include temperature-sensitive alleles, which provide a finely-tuned control of gene function by incremental temperature changes (Ben-Aroya et al. 2008; Li et al. 2011), tetracycline-repressible mutants, which rely on doxycycline-regulatable promoter to shut-off the expression of an essential gene of interest (Mnaimneh et al. 2004), and hypomorphic DAmP (decreased abundance by mRNA perturbation) alleles that destabilize the transcript by disrupting the natural 3’UTR by an insertion of an antibiotic-resistance marker (Schuldiner et al. 2005). Essential gene strain collections have facilitated exploration of the roles of highly conserved essential pathways, interconnections between various processes of the secretory system, transcriptional regulation, and functional annotation of previously uncharacterized essential genes (Mnaimneh et al. 2004; Schuldiner et al. 2005; Ben-Aroya et al. 2008; Li et al. 2011).

Recent advances have expanded the diversity of the existing collections. A prototrophic deletion library has filled a gap in our knowledge of the yeast metabolic network by providing tools for metabolic assessment of deletion mutants without the auxotrophic markers associated with the original deletion collection (VanderSluis et al. 2014). The development of β-estradiol inducible expression system has laid the groundwork for the construction of a collection that would allow studies of multiplex regulation in yeast (McIsaac et al. 2013). Finally, the implementation of novel genome engineering tools using Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) systems provides the foundation for simple and effective directed mutagenesis and allelic replacement (DiCarlo et al. 2013). Modifications to existing resources and novel genome editing strategies promise significant advances in functional characterization of the cell.

1.6 Synthetic genetic array (SGA) analysis

The availability of mutant libraries has made yeast genetics amenable to automated analysis. To achieve this goal, our group developed a technique called Synthetic Genetic Array (SGA)
analysis, which is a high-throughput method for constructing haploid yeast double mutants and the identification of genetic interactions (Figure 1.3) (Tong et al. 2001). SGA enables large-scale mating and meiotic recombination via a series of replica pinning steps, automating yeast genetics and introducing any marked allele of interest into ordered arrays of yeast mutant strains composed of non-essential gene deletions or conditional alleles of essential genes (Figure 1.3a). The resulting output arrays of haploid double mutants can be subsequently probed for a multitude of phenotypes, such as fitness, which is approximated in SGA by the simple readout of colony size (Figure 1.3b).

Colony size is a suitable phenotype serving as a surrogate measure of fitness in yeast. Fitness has been defined using two components (Houle 2006): 1) “the ability of an individual or population to leave viable and reproductively effective progeny, relative to the abilities of other individuals or populations; 2) the average contribution of an allele or genotype to the next or subsequent generations, compared with those of other relevant alleles or genotypes”. SGA analysis combines both of these fitness components since colony size measurements at the final selection step is a reflection of how reproductively successful the heterozygous diploid was i.e. how well it was able to sporulate and produce meiotic progeny and how successful those meiotic progeny with a specific combination of alleles were at establishing a colony. In effect, colony size offers a simple and yet powerful measure to deduce the adaptive significance and evolutionary dynamics resulting from genetic perturbations in question.

More specifically, SGA involves crossing an arrayed set of mutant strains marked with a dominant drug resistance gene (kanMX4), such as the deletion mutant collection, to a query strain carrying a mutation of interest marked with a different drug resistance (e.g. natMX4). In addition, the query strain also encodes a mating type-specific reporter, STE2pr_Sp_his5, which drives the expression of the Schizosaccharomyces pombe his5 gene, and complements S. cerevisiae his3Δ1, thereby enabling selective germination of meiotic progeny of only one mating type after sporulation of diploids heterozygous for the desired perturbations. The effectiveness of the haploid selection step is ensured by the accompanying deletions of the CAN1 and LYP1 loci that encode arginine and lysine permeases, respectively (Ahmad and Bussey 1986; Sychrova and Chevallier 1993). They are utilized to select against diploid cells that have been carried over from the sporulation mix. Since canavanine and thialysine are toxic analogs of arginine and lysine, respectively, diploid cells that are heterozygous for deletion alleles of the CAN1 and
LYPI transporters are selectively killed, while the desired haploids cells are can1Δ and lyp1Δ, and continue to be propagated until the final double mutant selection (Tong and Boone 2006). These genetic properties combined with robotic technologies contribute to the high-throughput of the SGA methodology.
Figure 1.3. Synthetic Genetic Array (SGA) Methodology. A MATα query mutant strain carries a mutation of interest linked to a natMX4 (filled red circle) conferring it resistance to the antibiotic nourseothricin (natR). It also carries deletions of the arginine and lysine permease genes, CAN1 and LYP1 (green outline), respectively, to select against diploids following the sporulation step. STE2pr_Sp_his5, in place of CAN1, expresses Schizosaccharomyces pombe his5 gene from the MATα-specific promoter (STE2pr). Typically, SGA query strain is crossed to an ordered array of MATα strains carrying a non-essential gene deletion (filled blue circle), each of which is marked by another dominant selectable marker, kanMX4, which confers kanamycin resistance (kanR). The resulting heterozygous diploids are sporulated and then MATα haploid meiotic progeny are selectively germinated and successive replica pinning steps are used to select for natR and kanR double mutants. Inset: A cropped plate image from the final SGA selection step. The query (queryΔ) or array (xxxΔ) gene deletions do not result in any observable fitness defects, but the deletion in both genes is lethal – a phenomenon termed synthetic lethality. The mutant is represented in quadruplicate on the array and is highlighted with a white box.
1.7 Genome-wide genetic interaction studies

Genome-wide genetic interaction studies have vastly contributed to our understanding of the genetic wiring of a eukaryotic cell. The first genome-scale SGA study was carried out using 132 query mutant strains related to actin-based cell polarity, cell wall biosynthesis, chromosome segregation and DNA synthesis and repair that were interrogated against the full non-essential gene deletion collection (Tong et al. 2004). This analysis identified ~ 4,000 synthetic sick and lethal interactions between ~1,000 genes enabling an initial characterization of genetic network properties and topology. Although genetic interactions are generally rare, the genetic network was shown to be a small world consisting of dense local neighbourhoods where functionally related genes were often connected by synthetic lethal/sick negative interactions suggesting an extensive genetic robustness. The genetic interaction network appeared rich in functional information providing insight into the biological roles of unknown genes.

Since then, SGA has been applied to a much larger set of query genes and evolved from qualitative to quantitative analysis of fitness. In the most recent large-scale experiment, SGA analysis was applied to examine ~5 million double mutants and identify ~170,000 negative and positive genetic interactions (Costanzo et al. 2010). The resulting global genetic network revealed genetic interaction profiles that enabled the annotation of novel pathway and complex memberships, and provided an initial glimpse into the functional organization of a eukaryotic cell by highlighting functionally enriched clusters as well as broad connections between diverse pathways and functional modules in the cell. In addition, these studies emphasized a largely orthogonal relationship between protein-protein and genetic interaction networks. SGA analysis has also been applied to subsets of functionally related genes to map interactions and genetic networks underlying specific biological processes (Schuldiner et al. 2005; Collins et al. 2007; Wilmes et al. 2008; Fiedler et al. 2009; Aguilar et al. 2010; Zheng et al. 2010).

SGA has inspired the development of alternative high-throughput methods for studying genetic interactions. For example, synthetic lethality analysis by microarrays (SLAM) (Ooi et al. 2003) and diploid-based SLAM (dSLAM) (Pan et al. 2004) generate mutants by directly disrupting a query gene by integrative transformation of pooled haploid or heterozygous diploid deletion mutants, respectively. dSLAM exploits a modified heterozygous yeast deletion collection, which contains the SGA reporter in every strain, allowing for haploid double mutant selection. Barcode
microarray abundance is used as a proxy for the relative growth profiles of mutants. These approaches were effectively used to define a DNA helicase genetic network and identify parallel pathways involved in the signal transduction of a DNA replication stress to the DNA damage checkpoint pathway (Ooi et al. 2003; Pan et al. 2004). A related approach is genetic interaction mapping (GIM), which conducts all the steps of SGA in liquid pooled cultures and similarly to SLAM and dSLAM, measures genetic interaction using a bar-code microarray read-out (Decourty et al. 2008). The development of SGA-based approaches extended the scope of genetic interaction studies enhancing our knowledge of effects of combinatorial perturbations.

1.8 Mapping genetic interactions in other eukaryotes

Genetic interaction studies have also been conducted in other organisms revealing important insight into gene function and species evolution. Pioneering genetic interaction studies in Schizosacharomyces pombe provided insight into the extent of genetic network conservation between distantly related yeast species (Dixon et al. 2008; Roguev et al. 2008). Roguev and colleagues mapped genetic interactions in the fission yeast involving double mutant combinations among 550 genes related to chromosome biology. Comparison with the budding yeast network showed that about 17% of negative genetic interactions were conserved, which nearly doubled to 33% when the analysis was restricted to functionally co-annotated gene pairs suggesting that conservation is mostly observed for functional modules, which may restrain evolutionary divergence (Roguev et al. 2008). Another study interrogated ~220 mutants for genetic interactions involved in DNA damage checkpoint activation and repair, chromatin remodeling, intracellular trafficking and various other processes (Dixon et al. 2008). Comparing the resulting genetic interaction network to S. cerevisiae revealed species-specific differences suggesting extensive functional rewiring of genetic networks across large evolutionary distances. For example, genetic perturbations of members of the conserved prefoldin complex and mitotic spindle formation/cell polarity pathway result in SS/SL in budding yeast but not in fission yeast indicative of greater redundancy of this process in the latter species. In addition, the authors also identified a smaller but a substantial fraction of interactions involving orthologous genes (29%) that are conserved between the two species. These findings suggest that despite divergence of about one billion years of evolution (Hedges 2002), the connectivity within and between some
core biological processes is conserved. Subsequent studies that mapped more genetic interactions in fission yeast confirmed the existence of conserved functional modules (Ryan et al. 2012a) and used shared gene and network properties to accurately predict highly connected \textit{S. pombe} genes (Koch et al. 2012). These results suggest that efforts to apply model organism data to evolutionary distant organisms, should be useful in providing insight into connections that are invariant versus plastic in evolution (Frost et al. 2012). Since, for instance, worm and human are as evolutionary far apart as \textit{S. pombe} and \textit{S.cerevisiae}, then the knowledge gained from large-scale genetic interaction studies which are more feasible in simpler single-cell eukaryotes and metazoans may be directly translatable to a more complex human system.

Currently efforts are ongoing to map genetic interactions in multicellular organisms. Systematic reverse genetic analysis is possible in \textit{C. elegans} by feeding worms bacteria expressing double stranded RNA (RNAi) targeting any gene of interest (Lehner et al. 2006a; Lehner et al. 2006b). Genome-wide loss-of-function phenotypic screens in worms make use of a genome-wide feeding library which consists of dsRNAs targeting 86% of 19,427 predicted open reading frames (Kamath et al. 2003). Unlike a unicellular organism such as yeast, the phenotypes used to score genetic interactions in worm measure effects on the fitness of the entire organism and include such features as viability, fecundity or growth rate. Despite the high level of conservation of genes in yeast and \textit{C.elegans}, as evident from 61% orthologous genes with an inviable phenotype in both organisms, the conservation of individual genetic interactions is at most 5% (Lehner et al. 2006a; Tischler et al. 2006; Tischler et al. 2008), which could be due to greater functional redundancy, multicellularity, incomplete knockdown by RNAi or inappropriate phenotypic comparisons (fecundity versus single cell viability). The relatively low conservation rate seen in these studies may also reflect the focus on signaling pathways, since other cross-species studies focusing on spindle assembly checkpoint and essential genes playing a role in chromosome biology revealed a much higher genetic interaction conservation rate \textasciitilde50% - 80%, although a much smaller gene set was assessed (Tarailo et al. 2007; McLellan et al. 2009). Interestingly, another cross-species study of genes involved in maintenance of genome stability showed a nearly 75% conservation of synthetic lethal genetic interactions between \textit{S. cerevisiae} and a human colorectal cancer cell line, in spite of an even greater evolutionary distance between yeast and human than worm and human (van Pel et al. 2013). These results show that the extent of conservation of genetic interactions remains an open question and depends on the interrogated
species and gene sets. However, despite unpredictable network rearrangements, the structure and general properties of networks themselves appear conserved between eukaryotic organisms emphasizing the importance of understanding the general principles governing the outcomes of combinatorial perturbations using model organisms (Tischler et al. 2008; Koch et al. 2012).

Systematic mapping of genetic interactions in mammalian cells have also been enabled by RNAi-based approaches. Genetic interaction studies are becoming recognized as important milestones for the development of therapeutic approaches based on principles of personalized medicine related to human diseases, such as cancer, a concept that was recognized early on by Lee Hartwell (Hartwell et al. 1997). Since cancer cells accumulate genetic abnormalities, it should be possible to selectively target them by identifying additional genetic perturbations that result in cancer cell-specific death while leaving normal tissue unaffected. Indeed, genome-wide shRNA screens in a wide array of different cell lines have identified a vast number of cancer cell-line specific genetic vulnerabilities (Silva et al. 2008; Luo et al. 2009; Possemato et al. 2011; Reiling et al. 2011; Vizeacoumar et al. 2013; Deng et al. 2014; Martz et al. 2014; Pandyra et al. 2015). For example, siRNA screens were used to study epigenetic regulators of colon cancer using a broad spectrum of cell parameters (Laufer et al. 2013). Other studies have used shRNA-based genetic interaction screens to study resistance pathways to therapeutic agents in human myelogenous leukemia cells (Bassik et al. 2013), as well as generate genetic interaction profiles in isogenic colorectal cancer cell lines carrying defined mutations in specific genes enabling functional characterization of unknown genes and functional modules with targetable vulnerabilities (Vizeacoumar et al. 2013). EsiRNA genetic interactions screens in mouse fibroblasts combined with multidimensional phenotyping strategy were also used to understand protein module architecture related to chromatin machinery (Roguev et al. 2013). Recent advances in CRISPR-Cas9-based genome engineering have opened new avenues for conducting synthetic genetic interaction studies without the drawbacks of RNAi such as lack of specificity and off-target effects (Doudna and Charpentier 2014).

1.9 Higher-order genetic interaction studies

Higher-order genetic interactions offer insight into genomic robustness that goes beyond two genes and captures novel multi-gene connectivity in the cell. To date there have been very few
studies that have examined the phenotypic consequences of simultaneous perturbation of three different genes (i.e. trigenic interaction studies), as the relevant reagents and methods have not been developed for systematic analysis. Yeast, yet again, has been the organism of choice for pioneering trigenic interaction studies.

The first study to conduct genome-wide trigenic interaction screens used \textit{bni1Δ bim1Δ} and \textit{bni1Δ kre1Δ} double mutants as query strains for screening against the non-essential gene deletion collection for triple mutant synthetic lethal (SL) or synthetic sick (SS) interactions (Tong et al. 2001). The screens resulted in a total of 33 SS/SL trigenic interactions compared to 294 digenic interactions specific to all the individual mutants showing a 10-fold reduction in the rate of synthetic interactions among three genes compared to two genes. However, the authors noted that trigenic interactions are 2000-fold more numerous than two-gene interactions thus the potential for triple mutant interactions contributing to the genetic wiring of a cell may be much more significant than suggested by these initial screens.

Other studies have used trigenic interactions for functional characterization of duplicated genes. Since single mutant query strains involving duplicated genes exhibit fewer interactions than singletons, it was suggested that double mutant queries would show rich and informative triple mutant genetic interaction profiles (Tong et al. 2004; Costanzo et al. 2010). The idea was tested by analyzing triple mutant interactions in strains deleted for pairs of duplicated G1 cyclin genes (Zou et al. 2009). A strain deleted for \textit{CLN1-CLN2} showed 36 SS/SL trigenic interactions in the double deletion mutant query strain screened against the non-essential gene deletion collection and 45 synthetic dosage lethal (SDL) interactions when screened against a plasmid-based overexpression collection (Zou et al. 2009). Trigenic screens using a query strain deleted for another pair of cyclin genes \textit{pcl1Δ pcl2Δ} exhibited many fewer interactions but they nonetheless proved to be functionally informative. Another study used triple mutant genetic interactions to characterize closely related paralogs, \textit{ESL1-ESL2}, providing evidence that they contribute to the regulation of adaptive stress responses (Lai et al. 2013). In addition, \textit{SCS3} and \textit{YFT2} have also been interrogated for trigenic interactions (Moir et al. 2012). These genes encode ER-resident Fat storage-Inducing Transmembrane (FIT) proteins, which are orthologs of the mammalian \textit{FIT} gene family. Their digenic and trigenic interaction profiles illuminated their functional similarity and divergence and implicated them in a regulatory mechanism that coordinates a balance between phospholipid and protein synthesis, with \textit{SCS3} communicating changes in the ER to the
transcriptional response of genes involved in phospholipid biosynthesis. The Moir et al. study represents the first attempt to quantify trigenic interactions: a semi-quantitative model was tested with an expectation that involved the product of double mutant fitness and single mutant array fitness, but filtered out double mutant effects in a binary manner regardless of the magnitude of the digenic effect. Despite its simplifying assumptions, this model was an important first attempt at quantifying trigenic interactions.

A more fine-tuned scoring method for triple mutant genetic interactions was recently developed (Haber et al. 2013). In this study trigenic screens of asf1Δ cac1Δ, rtt109Δ cac1Δ and clb5Δ clb6Δ revealed functional redundancies that are undetected when conducting digenic interaction screens, such as those that were detected between the pair of cyclins, CLB5-CLB6, and genes involved in chromosome segregation. While previous studies used four different screens for identification of trigenic interactions (a wild-type query for single mutant fitness; two single mutant queries for double mutant fitness; and a double mutant query for triple mutant fitness), Haber and colleagues, utilized a wild-type query for single mutant fitness and a double mutant query for double and triple mutant fitness quantification. Double mutant fitness was obtained by pinning the sporulated mix resulting from a cross between a double mutant query strain and an array of single mutants onto varying media types. In the case of digenic interactions, the haploid selection medium enabled selecting first for the array strain marker and then for either one of the query strain markers; for triple mutant interactions the haploids were germinated by first selecting for both query strain markers and then for the array strain marker (Braberg et al. 2014). Despite containing a mixed haploid population when selecting for double mutants, the authors reported that the results they obtain are ‘nearly identical’ to those that begin with a single mutant query and generate a colony consisting only of double mutant haploid progeny. However, there are potential issues with this approach – for example, a sick double mutant may be outcompeted by a healthier triple mutant haploid due to a suppressive effect of a third perturbation. In this case, the observed digenic interaction would be scored as weaker than its true fitness, leading to an increased false negative rate. This effect would also artificially dampen the resulting positive interaction of a triple mutant, which would also increase the corresponding false negative rate. In addition, the scoring approach for trigenic interactions relies on the subtraction of the sickest double mutant effect, which is inconsistent with the quantitative model used to score the digenic
interactions themselves are scored; I address some of these scoring issues in this thesis (Chapter 2).

Despite some caveats several complex genetic relationships were uncovered using the approach summarized above (Braberg et al. 2014). They include: 1) true redundancy in which any of the three genes are sufficient for pathway function (Figure 1.4a); 2) co-complex membership in which a protein complex contains up to two non-essential subunits where the disruption of any combination of two components is tolerated but a third perturbation results in sickness (Figure 1.4b); 3) compensatory repurposing when two genes act in parallel pathways but a third gene that performs a related but different function can compensate for defects caused by the double perturbations to the aforementioned pathways (Figure 1.4c); 4) multiple pathway members when the perturbation of each individual gene is detrimental to cell fitness but a second or a third perturbation does not increase the disruptive effect (Figure 1.4d); 5) suppression of deleterious functions when a double mutant is sick but a perturbation in a third gene suppresses the negative effect which could be caused by a third gene whose function becomes toxic in the absence of the other two genes (Figure 1.4e). Thus, trigenic interactions offer new and functionally important insights into various biological phenomena.

Recent work suggests that higher-order genetic interactions may be more pervasive than originally thought. A PCR-mediated chromosomal deletion technology was used to delete chromosomal segments in a one-step yeast transformation (Kaboli et al. 2014). This approach resulted in successful deletion of 33 of 110 attempted regions suggesting that the remaining regions are essential for viability. Synthetic lethal digenic interactions could explain the inability to delete about half of those segments due to the presence of members of co-lethal gene pairs in those segments. However, the inability of the cell to lose the rest of the segments suggests interplay of more than two genes and potentially producing many possible combinations given that the number of genes per segment ranges from 9 to 47. Thus, higher-order genetic interaction network may be extensive and provide a new view into the genetic wiring of a eukaryotic cell.
Figure 1.4. Biological interpretations of trigenic interactions. A) Redundancy is observed when any of the three genes are sufficient for performing an essential function that is impaired in a triple mutant. B) A protein complex, which contains up to two non-essential subunits, may tolerate disruption in one or two components but complex function may be lost by mutations of a third complex member, resulting in a negative genetic interaction. C) Compensatory repurposing occurs when mutation of two genes that act in parallel pathways has a minimal phenotypic effect because of an alternative pathway, such that a triple mutant shows a negative genetic interaction. D) ‘Classical epistasis’ is detected when a loss of function perturbation in one step of a biological pathway produces a mutant phenotype, but mutation in any combination of other steps does not exaggerate the effect and leads to a positive genetic interaction. E) Suppression of deleterious functions, which leads to a positive genetic interaction, occurs when a double mutant is sick but a perturbation in a third gene suppresses the negative effect, which could be caused by a gene whose function becomes deleterious in the absence of the other two genes. Wild-type scenario is highlighted with a grey box. A deleted gene is crossed-out; a conditional loss-of-function allele is depicted with a lower case letter. Adapted from Braberg et al. 2014.
1.10 Gene duplication is important for network robustness and species evolution

Genetic network redundancy may arise from gene duplication, which may contribute to the high-reliability of biological systems, which are composed of inherently unreliable and noisy components (McAdams and Arkin 1999). Signal failure will always happen in individual links with some probability, but reliability of signal transmission in genetic networks will increase predictably with redundancy. As such, gene duplicates could be retained by offering reliability benefits and hence, a fitness advantage to the organism (McAdams and Arkin 1999).

Furthermore, gene duplication is thought to be evolutionarily important for speciation and adaptive radiation. As Susumo Ohno explained in his seminal work ‘Evolution by Gene Duplication’, “natural selection often ignores such a redundant copy and, while being ignored, it accumulates formerly forbidden mutations and is reborn as a new gene locus with a hitherto non-existent function” (Ohno 1970e). In the remainder of this chapter, I focus on the importance of studying gene duplication to understand mechanisms of duplicate gene retention during evolution.

1.11 Mechanisms of gene duplication

Several mechanisms have been proposed to explain gene duplicate events (Figure 1.5).

Segmental gene duplication reflects error prone DNA replication in both mitotic and meiotic cell division. On the other hand, simultaneous duplication of all genomic segments can be generated by a variety of polyploidy events. Henceforth the former will be referred to as small-scale duplicates (SSD) and the latter as whole-genome duplicates (WGD).
Figure 1.5. Summary of mechanisms of gene duplication. Small scale duplication can occur by: 1) tandem duplication, which results from means of unequal exchange either between sister chromatids in mitosis or homologous chromosomes in meiosis I; and 2) transposition, which carries a locus from one position to another via RNA or DNA intermediates. 3) Duplication of the entire genome happens through autoploidy or allopolyploidy (details in main text). Black arrows represent a duplication event. Grey rods represent a chromosome. Coloured blocks depict a locus. Figure was adapted from (Thornton 2006).
**1.12 Segmental gene duplication**

Unequal exchange between two chromatids on the same chromosome can produce duplicated genes (Taylor et al. 1957). During mitosis there are frequent exchanges between sister chromatids; thus, even the slightest misalignment or slippage can cause two identical loci to be placed on one chromatid resulting in a deletion of the same locus on the other sister chromatid. Upon completion of a mitotic cycle, one of the daughter cells would receive two copies of a gene and would be heterozygous for the gene duplication and the other would be hemizygous carrying only one copy of the gene. If this event occurs during mitosis of germ cells with the duplication providing a selective advantage to the offspring then a small isolated population homozygous for the duplication could arise, which may be important for speciation (Ohno 1970c). Duplication events may also occur in meiosis prophase I if the process of genetic recombination by crossing-over is uneven (Smithies 1964). This would position loci from maternal and paternal homologous chromosomes in tandem fused on the same strand. Thus, errors in chromosomal division during the cell cycle give rise to gene duplicates.

Other mechanisms that generate segmental duplicates have also been described. For example, a segment of a DNA strand may replicate during G1-phase due to regional redundant DNA replication. Then, a single break and fusion at either end of the region of interest would result in tandem duplication of that entire segment (Ohno 1970b). Horizontal (or lateral) gene transfer from one organism to another is considered another possible mechanism (Hilario and Gogarten 1993). Such transfer could occur when there is mating between cells from different species, mediated by viruses or the transformation of plasmid DNA. Duplication would be observed in cases when the transferred gene has a homolog in the genome. However, the contribution of horizontal gene transfer to evolution is disputed since it requires the exchange of genetic material between different organisms in a single generation, which is in contrast to the gradualist approach advocated by neo-Darwinians (Boto 2010). Transposition has, also, been reported to generate duplicated genes. Transposition involves the reintroduction of a copy of a gene through a DNA or RNA intermediate into a new location in the genome. RNA-mediated retrotransposition can be inferred by the absence of introns and presence of a 3’ poly-A tract in a genomic region, which can be detected in the human and other genomes (Lander et al. 2001); many of these genomic sequences are non-functional pseudogenes but some still retain their activity (Lahn and Page 1999). DNA-mediated transposition is deduced by the presence of a
second exact copy of a gene (Thornton 2006). Transposition events are usually evident by the conserved terminal sequences that flank the duplicated segments (Hughes et al. 2003).

1.13 Whole genome duplication

Duplication of an entire genome can occur through polyploidy. Such a mechanism of duplication lacks the major challenges associated with segmental duplications, as it avoids dosage imbalances for functionally related genes and members of protein complexes and structural genes that are duplicated along with their corresponding regulators (Ohno 1970a). Chromosomal instability is also less likely to be induced because the original gene and its duplicate are located on two separate chromosomes that never undergo one-to-one pairing. There are two means to achieve polyploidy: auto- and allo-polyploidy. Autopolyploidy occurs when two daughter cells, which are generated in mitotic telophase, fuse into one cell or when two successive DNA replications happen in the absence of mitosis (Ohno 1970d). During meiosis such a tetraploid produces diploid gametes that go on to produce tetraploid organisms. On the other hand, allopolyploidy originates from an interspecies cross. Most such organisms are sterile because of the absence of pairing between chromosomes since they do not show homology. However, the Barbus barbus fish (Wolf et al. 1969), a cyprinid fish twice the size of other species, has twice the number of chromosomes compared to others of the genus, forms bivalents during meiosis resulting in successful reproduction, and is the first reported allotetraploid species.

In yeast, whole genome duplication appears to have occurred approximately 100 Mya after the divergence of Kluyveromyces from Saccharomyces lineages [Figure 1.6 (Wolfe and Shields 1997; Kellis et al. 2004)]. The timing of the duplication even is inferred from protein and nucleotide sequence alignments, which show that some regions in K. waltii correspond to two regions in the S. cerevisiae genome with blocks of conserved synteny, regions where genes lie in the same order in both species (Kellis et al. 2004). A model has been proposed to explain the emergence of a novel yeast species from such a polyploidy event (Scannell et al. 2006; Conant 2014). The model posits that two haploids from different species fused to make a diploid, but because of hybrid sterility, the spores were inviable so the diploid divided mitotically. This organism would have been losing alleles from every locus unless it produced haploinsufficiency, and thus could have lost a MAT locus, becoming functionally haploid. It could have switched
mating type and undergone mother daughter mating producing a diploid that would regain fertility producing separate loci from former alleles. Continued loss of redundant copies would then lead to creation of separate lineages.

Figure 1.6. Phylogeny and relative time of whole genome duplication (WGD). In yeast WGD happened approximately 100 Mya after the divergence of Kluyveromyces from Saccharomyces lineages. Figure adapted from Kellis et al. 2004.
1.14 Prevalence of gene duplication in eukaryotes

Duplication events have been described as major determinants of evolutionary history for many eukaryotic organisms ranging from yeast to humans. Table 1.1 depicts the prevalence of duplicated genes proposed to originate from SSD and WGD for *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Homo sapiens*.

### Table 1.1 Prevalence of gene duplication in eukaryotes

<table>
<thead>
<tr>
<th>Common name</th>
<th>Scientific name</th>
<th>Total no. genes in genome</th>
<th>No. WGD rounds</th>
<th>% WGD duplicates in genome</th>
<th>% SSD duplicates in genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>6604&lt;sup&gt;(1)&lt;/sup&gt;</td>
<td>1&lt;sup&gt;(4,5)&lt;/sup&gt;</td>
<td>18&lt;sup&gt;(4,5)&lt;/sup&gt;</td>
<td>30&lt;sup&gt;(6)&lt;/sup&gt;</td>
</tr>
<tr>
<td>Plant</td>
<td><em>Arabidopsis thaliana</em></td>
<td>26028&lt;sup&gt;(2)&lt;/sup&gt;</td>
<td>Multiple&lt;sup&gt;(2)&lt;/sup&gt;</td>
<td>29-59&lt;sup&gt;(2)&lt;/sup&gt;</td>
<td>40&lt;sup&gt;(7)&lt;/sup&gt;</td>
</tr>
<tr>
<td>Human</td>
<td><em>Homo sapiens</em></td>
<td>22980&lt;sup&gt;(3)&lt;/sup&gt;</td>
<td>2&lt;sup&gt;(3)&lt;/sup&gt;</td>
<td>26&lt;sup&gt;(3)&lt;/sup&gt;</td>
<td>5&lt;sup&gt;(8)&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

(1) YeastMine downloaded April 21, 2015 (ORF count includes genes of unknown and putative function)
(2) (Bowers et al. 2003)
(3) (Dehal and Boore 2005)
(4) (Byrne and Wolfe 2005)
(5) (Kellis et al. 2004)
(6) (Guan et al. 2007)
(7) (Maere et al. 2005)
(8) (Eichler 2001)
1.15 Conceptual frameworks for evolution of duplicated genes

How duplicated genes have been retained in evolution is a long-standing question in evolutionary genetics. Most duplicated genes become non-functional by a process termed ‘non-functionalization’, when one copy acquires a null mutation in the coding region or becomes otherwise non-functional, which drifts to fixation leading to gene loss (Force et al. 1999). However, those duplicates that remain are thought to be preserved by a variety of mechanisms that eventually lead to duplicate divergence or retention of functional redundancy (Figure 1.7).

It has been proposed that gene duplication provides a molecular landscape for functional innovation as the redundant copy escapes the constraints of natural selection and is free to acquire a novel or more specialized function, a concept termed neo-functionalization (Ohno 1970f). In the event of gene duplication, the accumulation of ‘forbidden’ mutations would be tolerated by the cell and ignored by natural selection because the production of a mutated polypeptide is generally harmless to the cell if there is a redundant functional copy. Thus, a protein with a previously non-existent function, which could simply be a change in the active site, could emerge from gene duplication (Figure 1.8). Examples of neo-functionalization include enzymes belonging to the fungal maltase family that display varying substrate specificities (Voordeickers et al. 2012). Neo-functionalization by subcellular reprogramming has also been

Figure 1.7. Evolutionary fates of duplicated genes. After a duplication event (small-scale or whole genome duplication), sister duplicates may become fixed if they have 1) functionally diverged by sub- or neo-functionalization or 2) retained functional redundancy by back-up compensation or dosage amplification.
reported, which occurs when sequence changes to protein targeting regions generate new localization patterns (Marques et al. 2008); however, the contribution of this mechanism to duplicate retention has been questioned given the similarity in frequency with which it also happens in singletons (Qian and Zhang 2009).

Retention of duplicates has also been proposed to result from sub-functionalization (Figure 1.8), when duplicates degenerate in function but are retained since they each provide a component of the function previously provided by the ancestor. The duplication-degeneration-complementation model (Force et al. 1999) predicts the following: 1) accumulation of degenerative mutations in regulatory elements can increase the likelihood of duplicate gene retention; 2) duplicate gene preservation can occur by partitioning ancestral functions. Redundancy is subject to processes of mutation and random genetic drift gene inactivation; thus if duplicated genes lose different functional domains or regulatory subfunctions, they must complement each other by together performing ancestral functions to prevent deleterious phenotypic effects. This mechanism of division of labour has substantial statistical support as well as examples in extant species that clearly show evidence of functional partitioning. Some well-known examples include engrailed in zebra fish, floral homeotic genes ZAG1-ZMM2 in maize (Force et al. 1999) and SKI7-HBS1, ORC1-SIR3, RNR2-RNR4, SNF12-RSC6 in yeast (van Hoof 2005). Recently, retention of SKI7-HBS1 has been shown to result from fixation of splice variants from an ancestral alternatively-spliced multifunctional protein (Marshall et al. 2013). Interestingly, in multicellular organisms sub-functionalization of regulatory domains has also been proposed to be a common fate of duplicated genes leading to tissue-specific expression (Lynch and Force 2000). Thus, subfunctionalization has explained the retention of a wide variety of duplicated genes.

A special case of subfunctionalization has been termed ‘escape from adaptive conflict’ and refers to a scenario whereby a multifunctional ancestral gene cannot optimize all its subfunctions simultaneously, a task that can be achieved through gene duplication that positively selects for separate optimization (Piatigorsky and Wistow 1991). Minimization of paralog interference is another special case of subfunctionalization that results in fixation of duplicates, which partitioned their functions (Baker et al. 2013). It focuses on duplicates for which the ancestral gene product participated in cooperative assemblies. Thus, immediately upon duplication sisters can engage in competitive interference for protein-protein or protein-nucleic acid interactions.
and division of interacting domains would be a favoured resolution. Such special cases provide support for sub-functionalization playing an important role in duplicate gene retention.

The selection of duplicated genes for functional redundancy is more controversial, since some population genetic theories suggest that duplication is inherently genomically unstable (Brookfield 1992), whereas others propose genetic models that support active selection of redundancy (Nowak et al. 1997). Regardless of the model it has been postulated that retention of gene duplicates may confer a fitness advantage by providing back-up compensation or dosage amplification. The analysis of the yeast deletion collection showed a lower fitness cost associated with deletion of a duplicated gene compared to singleton gene deletion highlighting the important role that duplicated genes play in robustness against null mutations (Gu et al. 2003). Evidence of partially shared regulatory motifs predicted transcriptional patterns that provide backup, which may be important for growth across different conditions (Kafri et al. 2005). In fact, there have been reports of “responsive backup circuits” that exist across various species in which a redundant gene copy is up-regulated when its sister is subjected to an inactivating perturbation (Kafri et al. 2006; DeLuna et al. 2010). Buffering capacity has also been explored under alternate growth conditions providing evidence for duplicates contributing to adaptability to environmental insults (Kondrashov and Kondrashov 2006; Musso et al. 2008). Thus, redundancy conferred by gene duplication results in backup compensation and confers fitness advantages.

Increased dosage of a gene product is another potentially beneficial outcome of gene duplication. For example, since the yeast whole genome duplication occurred around the same time as the evolutionary emergence of large, glucose-rich fruit, fixation of duplicates involving glycolytic genes may have been favourable (Conant and Wolfe 2007). Increased dosage of glycolytic enzymes may have produced higher glycolytic flux resulting in a faster growth rate, which was favoured by selection despite reduced efficiency of fermentation. High enzymatic flux associated with retention of duplicated genes has also been shown by other studies (Papp et al. 2004; Vitkup et al. 2006). However, recent work suggests that gene dosage effects of duplicates may not be sufficient to explain the fermentative capacity of yeast (Solis-Escalante et al. 2015). In addition, duplication, arising from an en masse duplication event such as WGD, would likely be subject to retention to ensure proper stoichiometry of key protein complexes (Hakes et al. 2007; Presser et al. 2008), such as the ribosome and chromatin (Papp et al. 2003).
Figure 1.8. Duplicate divergence by sub- or neo-functionalization. Duplicate divergence may proceed by sub-functionalization, retention of complementary sub-functions of an ancestral gene (‘division of labor’ by sister duplicates) or by neo-functionalization whereby over time one sister duplicate evolves a novel function not performed by the ancestral gene.
1.16 Experimental approaches to study fates of duplicated genes

As noted throughout this chapter, the question of why duplicated genes have been maintained during evolution remains a longstanding question in the field. Various studies have attempted to address this question experimentally. Analysis of genome sequences allows estimates of rates of divergence of both coding and regulatory regions (Gu et al. 2003; Kellis et al. 2004; Byrne and Wolfe 2005; Kafri et al. 2005; Ihmels et al. 2007; Musso et al. 2007; Grassi et al. 2010; Li et al. 2010; Plata and Vitkup 2014) and protein domain architecture (Grassi et al. 2010) and has identified gene families with accelerated and decelerated evolution which has helped to explain compensatory ability of sister duplicates. Gene ontology (GO) semantic distance has also been utilized to study sister duplicates (Baudot et al. 2004; Hakes et al. 2007; Li et al. 2010; Fares et al. 2013). This bioinformatics approach compares the functional relatedness of gene products based on the semantic similarity of their annotated GO terms and is thought to be more functionally informative than sequence-derived information alone. Metabolic flux balance analysis has also provided insight into evolutionary constraints on gene duplicate retention by examining buffering accomplished by metabolic network structure and function as well as flux reorganization (Papp et al. 2004; Vitkup et al. 2006).

As noted earlier, the yeast deletion collection has been used to generate a wealth of functional genomics data. Single mutant fitness of duplicated genes in nutrient-rich and alternative growth conditions has been used to examine dispensability of duplicates relative to singletons (Wagner 2000; Gu et al. 2003; He and Zhang 2006; DeLuna et al. 2008; Plata and Vitkup 2014). Analysis of duplicate gene expression levels, co-expression and gene expression profiles have also been used to probe gene function and understand transcriptional circuitry associated with duplicated genes (Wagner 2000; Wagner 2002; Gu et al. 2005; He and Zhang 2005; Kafri et al. 2005; Conant and Wolfe 2006; Guan et al. 2007; Tirosh and Barkai 2007; Wapinski et al. 2007; Wapinski et al. 2010). Changes in protein abundance of one sister upon perturbation of another (DeLuna et al. 2010) and analysis of similarity of interaction partners on the protein-protein interaction network (Wagner 2002; Baudot et al. 2004; He and Zhang 2005; Makino et al. 2006; Hakes et al. 2007; Musso et al. 2007; Zhu et al. 2013) have inched closer to more realistic functional readouts for estimating functional redundancy of duplicated genes. However, analysis of protein interaction networks has been limited by the sparsity of the data given that TAP-MS experiments even when gathered from multiple sources (Gavin et al. 2006; Krogan et al. 2006)
reveal at least one shared protein for only 8% of all possible duplicate pairs. Finally, the yeast GFP collection was used to survey subcellular localization patterns and then reconstruct phylogenetic relationship of proteins in a protein family to infer cases of sub- and neo-functionalization (Marques et al. 2008; Qian and Zhang 2009). All the studies summarized above have relied on either indirect measures of functional relationships of duplicated genes or have used functional data, which has been too sparse to address the question at hand. In my thesis project, I used systematic genetic interaction profiling to capture the true functional relationship of duplicated genes and instead of assessing single mutants, analyze data collected directly from the phenotypic manifestations of double mutants.

1.17 Rationale

Analysis of genetic interactions in double mutants has produced large functional genomics datasets that illuminate gene function and reveal key features of the global genetic interaction network. However, many phenotypes in simple eukaryotes as well as in metazoans are influenced by more complex genetic interactions involving larger sets of genes and genetic variants. Thus, deciphering the genotype-phenotype relationship for an individual requires that we expand our focus beyond pair-wise genetic interactions (digenic interactions) to include complex, higher-order genetic interactions involving more than two genes (trigenic interactions). The general principles that emerge from mapping more complex genetic interaction networks should shed further light on how genes buffer one another as well as on the genetic mechanisms that underlie heritable diseases.

Duplicated genes represent a fundamental class of genes for which mapping trigenic interactions will provide critical insight into mechanisms underlying genetic buffering, robustness and phenotypic variation. They often share some functional redundancy and thus screening for trigenic interactions involving duplicated gene pairs will provide insight into their biological roles and their functional overlap. Gene duplication has long been viewed as a primary mechanism for generating functional novelty because it allows for the relaxation of selective constraints and thus provides an opportunity for functional innovation or specialization. Genome sequence data from organisms across the tree of life have confirmed the prevalence of duplicated genes, many of which retain a high degree of sequence similarity. Despite the apparent central
role of duplication in genome evolution, the processes that shape the evolution of duplicate genes remain largely uncharacterized. I reasoned that measuring the genetic interaction profiles of duplicate gene pairs should provide a novel dataset with which to address fundamental questions regarding the evolutionary trajectories of duplicated genes.

In Chapter 2, I describe the optimization of the SGA methodology to construct triple mutant strains, the development of a robust scoring strategy to identify and quantify trigenic interactions in an effort to study duplicated gene retention. I constructed a collection of 275 double mutants containing deletions of 242 WGD and 33 SSD gene pairs, which is a unique resource for the yeast community. I used my collection to assemble a quantitative fitness standard for gene duplicate mutants on nutrient-rich and environmental stress conditions to probe the question of the contribution of different classes of duplicated genes to robustness against genomic and environmental perturbations. I produced trigenic interaction profiles for duplicated genes, providing a quantitative measure of the level of their functional divergence and redundancy and attempt to explain factors that drive duplicate gene evolution.

In Chapter 3, I exploit complex trigenic interactions to survey the trigenic landscape of a cell on a more global level. I developed a novel high-throughput strategy to construct double mutant query strains that are designed to enable a survey of ‘triple mutant space’. I also generated a trigenic interaction network by screening those strains and then used the network to develop a model for predicting trigenic interactions based on features characteristic of a digenic interaction network. This approach enabled me to estimate the extent of triple mutant interaction space in a eukaryotic cell. Third, I describe trigenic interactions involving \textit{S.cerevisiae} homologs of \textit{S.pombe} synthetic lethal pairs and estimate the level of genetic interaction network conservation between distantly related species.
Chapter 2

Portions of this chapter have been adapted from:


I wrote or substantially edited all sections presented in this Chapter.

I designed and implemented the strain construction strategy, adapted SGA for studying triple mutant genetic interactions, screened query strains on standard and alternative growth conditions, conducted validation experiments and coded and implemented analyses. Yiqun Chen, Ermira Shuteriqi, Bryan-Joseph San Luis and Matej Usaj provided technical assistance during strain construction and screening processes. In close collaboration with Benjamin VanderSluis, I developed the scoring methodology for triple mutant genetic interactions. Benjamin VanderSluis wrote and implemented the scoring code, evaluated the quality of the data using precision-recall analysis and modeled paralog divergence. Raamesh Deshpande wrote and implemented the code for selecting the genes comprising the diagnostic array. With Vince Messier, I designed and implemented experiments assessing colony growth on alternative growth conditions and collaborated with Elizabeth Koch to generate single and double mutant fitness standards on standard and alternative growth conditions and quantified conditional genetic interactions. Carles Pons contributed to the reproducibility analyses. Alex Nguyen calculated the measure of asymmetric divergence.
2 Retention of duplicated genes unravelled through complex genetic interactions

2.1 Introduction

A sizable fraction of many genomes across the tree of life is duplicated (Eichler 2001; Bowers et al. 2003; Byrne and Wolfe 2005; Dehal and Boore 2005; Maere et al. 2005; Guan et al. 2007). Despite the apparent central role of duplication in genome evolution, the factors that influence duplicate retention are poorly understood. Analyses of duplicated genes have generally relied on either indirect sequence-based measures of the functional relationship of duplicated genes or have used functional data, which have been too sparse to address the question at hand. Here, I describe an approach based on genetic interactions, which I used to capture the functional relationship of duplicated genes and to directly assess the phenotypic manifestations of deletions of duplicate gene pairs. I discovered that dispensable paralogs retained a gradient of buffering capacity, which is masked by complex genetic interactions. My work suggests that the evolutionary trajectories of duplicated genes can be explained by considering multi-functionality and structural entanglement.

2.2 Materials and Methods

2.2.1 Constructing a yeast paralog double mutant collection

**Gene pair selection:** Data were extracted from Byrne et al. 2005 to assemble a list of 242 WGD duplicate gene pairs. The set of 33 SSD pairs was chosen such that the paralogs did not belong to large gene families (greater than 2 members); members of gene families were identified by a sufficient sequence similarity score (FASTA Blast, E = 10), sufficient protein alignment length (> 80% of the longer protein) and amino acid level identity of at least 30% for proteins with aligned regions longer than 150 aa (VanderSluis et al. 2010). Only strains with a double mutant fitness $\geq 0.7$ (Costanzo et al. 2010) were chosen. Ten pairs involving members of protein complexes were also included (protein complex standard from Baryshnikova et al 2010); members of each complex must have had some positive and no negative genetic interactions between them and the selected pairs were connected by a positive pairwise genetic interaction. All query gene pairs are listed in Appendix 1.
Figure 2.1. Overview of query strain construction. A haploid lab SGA query strain deleted for paralog 1 (par1Δ::natMX4), was crossed to Y7091 (MATa can1Δ::STE2pr-Sp_his5 lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 LYS2) to generate a diploid strain heterozygous for par1Δ::natMX4/PAR1 and homozygous for the SGA reporters can1Δ::STE2pr-Sp_his5/can1Δ::STE2pr-Sp_his5 lyp1Δ/lyp1Δ. Diploids were isolated by micromanipulation and the resulting strain was used for PCR-mediated gene deletion of paralog 2. Upon induction of sporulation, the desired haploid double mutant was isolated by tetrad analysis.
Query strain construction by PCR-mediated gene deletion: The main query strain construction method is described below and depicted in Figure 2.1. Strains deleted for paralog 1 (par1Δ::natMX4) were compiled from an SGA query strain collection (SN) (Costanzo et al. 2010) and crossed to Y7091 (MATa can1Δ::STE2pr-Sp_his5 lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0LYS2) to generate a diploid strain heterozygous for par1Δ::natMX4/PAR1 and homozygous for the SGA markers can1Δ::STE2pr-Sp_his5/can1Δ::STE2pr-Sp_his5 lyp1Δ/lyp1Δ. Diploids were isolated by micromanipulation and the resulting strains were used for PCR-mediated gene deletion of paralog 2 using a Kluyveromyces lactis URA3 cassette (KlURA3; GenBank accession number: D00431; par2Δ::KlURA3). This marker was chosen since: 1) it avoids the use of additional costly antibiotics in the medium and 2) it is expressed from its endogenous promoter avoiding potential metabolic effects caused by gene overexpression (Amy Caudy, personal communication). Plasmid pFA6a_GST_KlURA3 (Sung et al. 2008) was used as a template for amplification of KlURA3 flanked by 55 bp gene-specific sequences for directed integration. The PCR amplification was conducted in 100 µl volume, using the Expand High Fidelity PCR System, with the following final concentrations: 1.5 mM buffer + 15 mM MgCl₂, 200 µM dNTP, 2.1 U Expand HiFi enzyme mix, 5% DMSO, 25 ng pFA6a_GST_KlURA3, 2 µM gene specific forward primer (52 b immediately upstream of the start codon + 3 b start codon) and 2 µM gene specific reverse primer (52 b immediately downstream of the stop codon + 3 b stop codon); and under the following conditions: 95°C/5 min; 30 cycles of 95°C/30 sec, 55°C/30 sec, 68°C/2 min; 68°C/10 min; 4°C. The PCR product size (1,513 bp = 1,403 bp KlURA3 + 55 bp upstream gene specific sequence + 55 bp downstream gene specific sequence) was verified using gel electrophoresis (sample volume: 5 µl; gel: 1.5% agarose gel + 1:10,000 SYBR safe; 120V, 30 min).

Primers for the amplification of KlURA3 (5’-3’):

Forward KlURA3_F: cggagacaatcatatgggag
Reverse KlURA3_R: tctggaggaagtttgagagg

High-efficiency LiAc transformation: The LiAc transformation method was adapted from http://research.fhcrc.org/gottschling/en/protocols/yeast-protocols/transformation.html. Briefly, an overnight culture was set-up in 1 ml of YEPD (1% yeast extract, 2% peptone, 2% glucose, 0.012% adenine), 200 rpm, 30°C. The next day, 5 ml YEPD were inoculated with 100 µl of
saturated overnight culture, and incubated for 5 hr, 200 rpm, 30°C. The cells were harvested by centrifugation at 1000 x g, 5 min, then washed in 1 ml sterile water, followed by a wash in 1 ml 100 mM LiAc and incubated in fresh 500 µl 100 mM LiAc for 10 min. The cells were then microfuged at maximum speed for 15 sec, LiAc was removed by aspiration and the pellet was resuspended in the transformation mix: 240 µl 50% PEG, 36 µl 1 M LiAc, 10 µl ssDNA (9-12 mg/ml), 54 µl sterile water, 20 µl PCR product for a total volume of 360 µl. Then, the cell mixture was heatshocked at 42°C, 40 min in a heat-block. Afterwards it was microfuged at maximum speed for 15 sec and the pellet was resuspended in 1 ml YEPD and incubated for 2 hrs, 200 rpm, 30°C. Microfugation was repeated and the pellet was resuspended in 200 µl sterile water and plated on 10 cm SD-Ura agar (0.67% yeast nitrogen base without amino acids, 0.2% amino acid supplement, 2% agar, 2% glucose) plates. Twelve average sized transformants were streaked out for single colonies, patched onto SD-Ura agar plates and confirmed by colony PCR.

**Confirmation PCR:** The colony PCR protocol was adapted from Ryan et al. 2012b. Cell lysate was prepared by suspending a yellow tip full of cells in 20 µl of zymolyase (5 mg/ml in 1 M sorbitol) in a 96-well PCR plate format and incubating at 37°C, 30 min followed by 95°C, 10 min. The PCR amplification was conducted in 25 µl volume using Expand High Fidelity PCR System with the following final concentrations: 1.5 mM buffer + 15 mM MgCl₂, 200 µM dNTP, 0.7 U Expand HiFi enzyme mix, 5% DMSO, 10% v/v lysate, 2 µM common primer internal to KIURA3 and 2 µM gene-specific confirmation primer either 400 bp upstream of the start codon (Ryan et al. 2012b) or 400 bp downstream of the stop codon (Costanzo et al. 2010). Both the forward and reverse PCR confirmations were performed for each mutant. The conditions for PCR amplification were the following: 94°C/5 min; 40 cycles of 94°C/45 sec, 52°C/2 min, 68°C/4 min; 68°C/7 min; 4°C/forever. The PCR product size (forward confirmation PCR product: 770 bp + X bp upstream of start codon; reverse confirmation PCR product: 654 bp + X bp downstream of stop codon) was verified using gel electrophoresis (sample volume: 25 µl; gel: 1% agarose gel + 1:10,000 SYBR safe; 120V, 30 min).

Primers internal to KIURA3 (5’-3’):

Forward KIURA3_901_F: caagtgcgcaagagttcac (for downstream confirmations)

Reverse KIURA3_901_R: gtcacctttggcagtcttgtg (for upstream confirmations)
**Isolation of double mutants by tetrad dissections:** Tetrad analysis was employed to isolate the desired double mutant strains (Figure 2.2). For this purpose, the PCR confirmed diploid strain was patched onto medium low in carbon and nitrogen to induce sporulation (1% potassium acetate, 0.01% amino acid supplement, 2% agar) and incubated for 7-14 d, 22°C. To digest the ascus wall, the sporulated mix was resuspended in zymolyase solution (0.5 mg/ml in 1 M sorbitol), then incubated at 37°C for 10 min. Digestion was terminated by the addition of 300-500 µl cold sterile water and 20 µl of the cell suspension was plated on synthetic complete medium. Twelve tetrads per mutant were dissected and incubated at 30°C, for 3 d at which time they were replica plated on YEPD, YEPD + clonNAT (100 µg/mL), SD-Ura, SD_msg – Ura + clonNAT (0.17% yeast nitrogen base without amino acids and ammonium sulfate, 0.1% monosodium glutamic acid, 0.2% amino acid supplement, 2% agar, 2% glucose), YEPG (1% yeast extract, 2% peptone, 2% galactose, 0.012% adenine), DI (0.2% amino acid supplement (containing histidine, leucine, uracil, methionine and adenine), 2% agar, 2% glucose), SY2014 (MATa ste3Δ306::LEU2 sst2Δ; ste2Δ mfa1Δ mfa2Δ::FUS1-lacZFUS1::HIS3; ura3-52 leu2-3-112, ade1) lawn and SY2625 (MATa bar1Δ; his3::FUS1-HIS3 mfa2Δ::FUS1-lacZ; ura3-1 leu2-3,-112 trp1-1 ade2-1 can1-100) lawn. Mating type assays were performed as previously described (Sprague 1991; Adames et al. 1995). SY2014 (ste3Δ306 sst2Δ) shows hypersensitivity to a-factor such that secretion of bioactive a-factor from the replica plated spore colonies would arrest the growth of the surrounding SY2014 lawn cells, creating a clear zone or halo. Similarly, SY2625 (bar1Δ) shows hypersensitivity to α-factor such that secretion of bioactive α-factor from the replica plated spore colonies would arrest the growth of the surrounding SY2625 lawn cells, creating a clear zone or halo. The strain with the following genotype *MATa par1Δ::natMX4 par2Δ::KIURA3 can1Δ::STE2pr-Sp_his5 lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 LYS2* was preserved in 20% glycerol for future experiments.
Constructing single mutant control query collection

Single mutant control strains (550 in total) deleted for one member of a paralog pair were constructed such that each strain carried the deletion of one member of a paralog pair with the relevant control marker inserted at the benign ho locus. To construct single mutant controls, strains with the following genotype were used: 1) SN deletion strain $MATa\ par1\Delta::natMX4\ can1\Delta::STE2pr-Sp\_his5\ lyp1\Delta\ his3\Delta1\ leu2\Delta0\ ura3\Delta0\ met15\Delta0\ LYS2+$; or 2) newly constructed $MATa\ par2\Delta::KlURA3\ can1\Delta::STE2pr-Sp\_his5\ lyp1\Delta\ his3\Delta1\ leu2\Delta0\ ura3\Delta0\ met15\Delta0\ LYS2+$.

They were transformed with either a $KlURA3$ or $natMX4$ PCR product containing 55 bp of homology flanking the $HO$ ORF. Correct gene disruption was confirmed by colony PCR.

Primers for amplification of $natMX4$ (5’-3’):

Forward Ptef_F: acatggagcccagaatacc

Reverse Ttef_R: gaatgctggtcgctatactg

Primers internal to $natMX4$ (5’-3’):

Forward NATB2: cggatcagagtagaccagaga (for downstream confirmations)

Reverse NATC2: atgggtctctaccacc (for upstream confirmations)
Query strain ploidy was verified using Fluorescence-Activated Cell Sorting (FACS). Briefly, strains were cultured in YEPD overnight then diluted to OD$_{600} =$ 0.1 in fresh YEPD and grown 5-6 hr, 30°C, 200 rpm. The cell suspension was centrifuged, 1,000 x g, 3 min and resuspended in cold 1 ml 70% ethanol, then incubated overnight at 4°C with rotation. The cells were washed in 1 ml sterile water and 500 µl were discarded. The remaining 500 µl were pelleted and resuspended in 200 µl of 50 mM Tris-Cl pH 8.0 + 0.2 mg/ml RNase A, then incubated 3 hr, 37°C. The cell suspension was centrifuged, 1,000 x g, 3 min and resuspended in 200 µl of 50 mM Tris-Cl pH 7.5 + 2 mg/ml proteinase K, incubated 30-60 min, 50°C. The cell suspension was centrifuged, 1,000 x g, 3 min and resuspended in 0.5 ml FACS buffer. Then, on average 14 µl were transferred to a round bottom 96-shallow well plate with the addition of 200 µl SYBRgreen (1:5,000, 50 mM Tris-Cl pH 7.5). A benchtop flow cytometer running the GUAVA program was used to obtain a FACS profile, which was analyzed using FLOWJO. Another quality control step consisted of verifying the mating type of the strain by performing crosses with Y12978 (MAT$^a$ lys1Δ) and Y12979 (MAT$^\alpha$ lys1Δ) and testing for complementation and growth on SD minimal media.

2.2.3 Synthetic genetic array (SGA) analysis technique for triple mutants

The BioMatrix Colony Arrayer was employed to screen the double mutant and control query strains against an array of single mutants. First, query strain lawns were prepared by inoculating 5 ml of YEPD medium with a single colony of a mutant of interest and incubating at 26°C, 2 days, at 200 rpm shaking incubator. Then, 800 µl of the query strain culture were spread onto YEPD agar plate (OmniTray). For genome-wide screens of the deletion mutant array (DMA; 14 plates), 4 query lawns were used, while 1 query strain lawn was needed for screens against the 4-plate array of temperature sensitive alleles of essential genes (TSA) and 1 lawn for a screen against a 4-plate array of a diagnostic array triple mutant array. Lawns were dried and then incubated at 26°C, 2 days. Query strain lawns were pinned onto a fresh YEPD plate and then the DMA on top of the query. The mated mix was incubated at room temperature for a day and then diploids were selected by pinning the resulting MAT$^a$/α diploid zygotes to YEPD + G418/clonNAT and incubated at 26°C, 2 days. The resulting natR/Ura+/kanR MAT$^a$/α diploids were sporulated by pinning them onto enriched sporulation agar plates and incubated at 22°C, 7
days. To select for $MAT_a$ meiotic haploid progeny, the spores were pinned onto SD – His/Arg/Lys + canavanine/thialysine (50 µg/ml of each analog) and incubated at 26°C, 2 days. Following haploid selection step, kanR $MAT_a$ meiotic haploid progeny were selected by pinning the haploid mix onto SD$_{MSG}$ – His/Arg/Lys/Ura + canavanine/thialysine/G418 and incubating the plates at 26°C, 2 days. Finally, natR/Ura+/kanR $MAT_a$ meiotic haploid progeny were selected by pinning the double/triple mutant haploid mix onto SD$_{MSG}$ – His/Arg/Lys/Ura + canavanine/thialysine/G418/clonNAT and incubating at 26°C, 2 days. The resulting triple mutant arrays were imaged using a high-resolution digital imaging system developed by S&P Robotics, Inc. (Toronto, ON). The images were processed using custom-developed image-processing software that distinguished colonies from the background and measured their area in pixels. The raw colony sizes were corrected for systematic effects and genetic interactions were quantified using the single mutant control screens as reference (Baryshnikova et al. 2010b).

### 2.2.4 Estimating single and double mutant fitness on standard growth conditions

To establish a double mutant fitness standard for duplicated genes, a high-density array (TMQv1) composed of all double mutant query strains was assembled (Appendix 4). The array was organized such that strains on the same chromosome or with a fitness defect were not close together. Gaps on the array were filled (including the border) with either Y13096 ($MAT_a$ $ura3Δ::natMX4$ $hoΔ::KiURA3$ $can1Δ::STE2pr-Sp$ $his5$ $lyp1Δ$ $his3Δ1$ $leu2Δ0$ $ura3Δ0$ $met15Δ0$ $LYS2+)$ or Y14412 ($MAT_a$ $his3Δ1::natMX4$ $hoΔ::KiURA3$ $can1Δ::STE2pr-Sp$ $his5$ $lyp1Δ$ $his3Δ1$ $leu2Δ0$ $ura3Δ0$ $met15Δ0$ $LYS2+)$ to enable the arrays to be crossed to two control strains carrying deletions in two different benign loci ($HIS3$ and $URA3$: DMA1 ($MAT_a$ $his3Δ1::kanMX4$ $leu2Δ0$ $ura3Δ0$ $met15Δ0$) and Y14420 ($MAT_a$ $his3Δ1$ $leu2Δ0$ $ura3Δ0$ $::kanMX4$ $met15Δ0$). Two loci were chosen such that most query strains would not fall into the linkage group of either locus (estimated to affect ~10% of the paralogs). Once the arrays were crossed to these control strains, they were subjected to the steps of triple mutant SGA and scored for colony size to estimate fitness (Baryshnikova et al. 2010b). Each high-density array was screened in triplicate for a total of 6 replicates. Since the arrays were in 1536-colony format, there are 4 technical replicates of mutants on the array resulting in a total of 12-24 colony measurements for each fitness estimate. The evaluation of fitness measures was done by comparison with Costanzo et al.
2010: single mutant fitness was obtained from ‘smf_t26_130417.txt’ and double mutant fitness was estimated from most up to date genetic interaction screens extracting from the first available datasets in this order: scored_sga_fg_t26_131130_scored_140103.txt, scored_sga_ts_t26_131130_scored_140103.txt, scored_sga_fg_t30_131130_scored_140103.txt, scored_sga_ts_t30_131130_scored_140103.txt).

2.2.5 Quantitative fitness-based model of genetic interactions

Digenic interactions were scored as previously described in Baryshnikova et al. 2010 and trigenic interactions were scored using the following models:

\[(\text{digenic}) \quad \epsilon_{ij} = f_{ij} - f_i f_j \quad \text{(Eq. 2.1)}\]

\[(\text{trigenic}) \quad \tau_{ijk} = f_{ijk} - f_i f_j f_k - \epsilon_{jk} f_i - \epsilon_{ik} f_j - \epsilon_{ij} f_k \quad \text{(Eq. 2.2)}\]

where \(\epsilon\) is the digenic interaction score, \(\tau\) is the trigenic interaction score, \(f\) is fitness and \(i, j, \) and \(k\) are individual mutations.

2.2.6 Selecting strains for array miniaturization

A greedy algorithm was employed to select mutants for construction of a miniaturized array with a reduced representation of the entire genome. The process was based on genetic interaction profiles gathered from screens against arrays consisting of non-essential gene deletion mutants (NES) and temperature sensitive alleles of essential genes (ES) from the large SGA project in the lab (Costanzo et al., in preparation). Each essential gene was represented by one allele with an average number of genetic interactions that behaves well at 26°C - the intended screening temperature. A total of 1,232 mutants (1,032 NES : 200 ES) were selected and positioned such that strains with fitness defect and chromosomal proximity were as far apart from each other as possible.
2.2.7 Confirmation of trigenic interactions by tetrad and random spore analyses

Negative trigenic interaction scores were confirmed by tetrad and random spore analyses using standard techniques. Briefly, a double mutant query strain was crossed to single mutant array strains of interest and subjected to tetrad analysis described in section 2.2.1, except YEPD+G418 (200 µg/ml) agar was added during the replica-plating step to identify kanR spore colonies that carry the array mutation. Random spore analysis was performed as previously described (Kuzmin et al. 2014) with additional steps for triple mutants. A small amount of spores (~ the size of a pinprick) was resuspended in 1 ml of sterile water and mixed well. Then, 20 µl of the suspension was plated on SD – His/Arg/Lys + canavanine/thialysine; 40 µl on SD<sub>MSG</sub> – His/Arg/Lys + canavanine/thialysine/G418; 40 µl on SD<sub>MSG</sub> – His/Arg/Lys + canavanine/thialysine/clonNAT; 40 µl on SD<sub>MSG</sub> – His/Arg/Lys/UrA + canavanine/thialysine; 80 µl on SD<sub>MSG</sub> – His/Arg/Lys/UrA + canavanine/thialysine/G418, 80 µl on SD<sub>MSG</sub> – His/Arg/Lys/UrA + canavanine/thialysine/clonNAT, 80 µl on SD<sub>MSG</sub> – His/Arg/Lys/UrA + canavanine/thialysine/G418/clonNAT, and 160 µl on SD<sub>MSG</sub> – His/Arg/Lys/UrA + canavanine/thialysine/G418/clonNAT. The plates were incubated at RT for 4 days and scored by comparing growth on the single, double, and triple drug selection plates.

2.2.8 Estimating the effect of environmental stress conditions on single and double mutant fitness

To assay growth in different stress conditions, the final array plate following SGA (Section 2.2.4) was replica pinned onto two additional plates containing stressors in addition to a standard growth condition as depicted in Figure 2.3. Stress conditions tested are summarized in Table 2.1. Fitness was assessed on three concentrations of stressor and in two replicates for all conditions and matched with no-drug/vehicle control conditions (Appendix 4). The vehicles were 0.5% DMSO for all concentrations of tunicamycin, trichostatin, bortezomib, fluconazole and 1 µg/ml brefeldin A; 2.5% DMSO for 10, 100 µg/ml brefeldin A and 0.01% ethanol for all concentrations of rapamycin. All concentrations of calcofluor white, cycloheximide, hydroxyurea, caspofungin, bleomycin, tunicamycin, brefeldin A, trichostatin and bortezomib were buffered with 100 mM Tris-HCl. Images of colonies were captured across days 2-7 of incubation at 26°C.
Fitness defects caused by stress conditions were quantified using a differential score that utilizes an experimental framework that matches the control and condition environments as closely as possible. Each conditional screen was performed alongside a paired control screen so that both screens were done on the same machines, at the same time, and by the same person, to minimize systematic experimental effects. We assumed that the conditional fitness of a yeast strain follows the model

$$f_{\text{condition}} = f_{\text{ctrl}} + f_{\text{diff}}$$  (Eq. 2.3)

where \( f_{\text{ctrl}} \) is the strain’s fitness under a standard condition and \( f_{\text{diff}} \) is the effect the condition has on the strain’s growth, i.e. the differential score was calculated from experimental data as

$$f_{\text{diff}} = f_{\text{condition}} - f_{\text{ctrl}}$$  (Eq. 2.4)

The method is similar to previous methods (Bandyopadhyay et al. 2010), except that subtraction of the control from the case values occurs near the beginning of the process whereas corrections and scoring is performed in the differential space. First, raw colony sizes are scaled so that each physical plate has the same median colony size. The scaled colony sizes of strains grown in case conditions are shifted into conditional space by subtracting the matched control-condition strains, producing differential interaction scores. Spatial normalization as well as some quality control filtering (described in Baryshnikova et al. 2010) is performed on the differential scores and p-values are estimated from variance of biological replicates.

Functional evaluation involved generation of a gold standard list of literature-curated fitness defects using YeastMine for precision-recall analysis (download date - January 27, 2015). The following series of steps were followed for each condition to generate the list: under ‘Templates’ \( \Rightarrow \) ‘Phenotype \( \Rightarrow \) Genes’ \( \Rightarrow \) ‘Edit Query’. The ‘Model Browser’ was constrained to 1) a condition of interest under ‘Chemical’; 2) ‘null’ under ‘Mutant type’; 3) ‘haploid’ and ‘homozygous’ mutants under ‘Experiment type’ and 4) ‘S288c’ under ‘Strain background’. First, precision-recall analysis was performed as previously described in Baryshnikova et al. 2010. Second, enrichment analysis was done using GO biological process annotations and significance was assessed using the hypergeometric test. Differential scores < -40 were used for both precision-recall and GO enrichment analyses. Third, growth curve analysis used normalized colony sizes to select a time-point at which no more growth was observed. Finally, an attempt
was made to maximize replicate reproducibility, which was assessed using Pearson correlation coefficient, and variance to capture a spectrum of fitness defects. The chosen concentrations are highlighted in bold in the Table 2.1.

**Figure 2.3. Conditional SGA for estimating fitness in environmental stress conditions.**
A control strain carrying a deletion in a benign locus was crossed to an array of double mutants lacking paralog pairs, taken through all the conventional selection steps of SGA. In the last pinning step, the array was copied three times onto a no-drug/vehicle condition and two conditions of interest. The resulting arrays were incubated 2-7 days, imaged and quantified.
To quantify conditional genetic interactions, observed in the double mutant, double mutants with significantly negative differential scores were identified as described above (Appendix 5). Then, an expected double mutant fitness was calculated based on the multiplicative model using single mutant fitness on a particular condition and compared to the observed double mutant fitness on that condition to identify cases of deviation. A genetic interaction was called if its score was one standard deviation away from the mean.

\[ f_{\text{diff}} = f_{\text{condition}} - f_{\text{ctrl}} \]  
\[ e_{1,2} = f_{1,2} - (f_1 f_2) \]  

(Eq. 2.5)  
(Eq. 2.6)

Table 2.1. Environmental stress conditions for fitness assessment.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>brefeldin A</td>
<td>1/10/100 µg/ml</td>
</tr>
<tr>
<td>caspofungin</td>
<td>0.005/0.03/0.1 µg/ml</td>
</tr>
<tr>
<td>calcofluor white</td>
<td>0.5/5/50 µg/ml</td>
</tr>
<tr>
<td>bleomycin</td>
<td>0.5/2/7.5 µg/ml</td>
</tr>
<tr>
<td>tunicamycin</td>
<td>0.1/0.5/1 µg/ml</td>
</tr>
<tr>
<td>fluconazole</td>
<td>0.2/2/16 µg/ml</td>
</tr>
<tr>
<td>rapamycin</td>
<td>0.5/2/10/100 nM</td>
</tr>
<tr>
<td>MMS</td>
<td>0.0005/0.001/0.01%</td>
</tr>
<tr>
<td>hydroxyurea</td>
<td>2/20/200 mM</td>
</tr>
<tr>
<td>bortezomib</td>
<td>100/250/500 μM</td>
</tr>
<tr>
<td>galactose</td>
<td>2%</td>
</tr>
<tr>
<td>cycloheximide</td>
<td>0.01/0.05/0.5 µg/ml</td>
</tr>
<tr>
<td>trichostatin A</td>
<td>12.5/25/50 µg/ml</td>
</tr>
<tr>
<td>sorbitol</td>
<td>0.1/0.5/1 M</td>
</tr>
</tbody>
</table>

2.2.9 Correlation with evolutionary and physiological features

Co-localization data were obtained from (Marques et al. 2008), expression difference from (Holstege et al. 1998) calculated in the following manner |PAR1-PAR2|/(PAR1+PAR2), genetic interaction network path length and degree asymmetry from (Costanzo et al. 2010), sequence
divergence asymmetry I using rate of divergence ‘2div1’ (Kellis et al. 2004) and sequence divergence asymmetry II using the measure described below. Previously, we have shown that we can computationally predict large conserved protein segments, which are very likely to be functional domains of proteins and are efficient filters for studying disordered protein regions (Nguyen Ba et al. 2014). These large conserved protein segments can be predicted in species prior to the whole-genome duplication and analyzed for changes in rates of evolution post-genome duplication. Focusing only on these domains (which we concatenated to prevent mispredictions of a single domain into multiple smaller domains), we used PAML (Yang 2007) to estimate the number of amino acid substitutions per site for both post-WGD clade and the pre-WGD clade (PAML program: AAML, clock=0, cleandata=0, fix_omega=0, ncatG=8, WAG rate matrix). The rates of substitutions are normalized against a pruned species tree to account for differing species composition between clades and overall different rates of evolution, and we then obtain a fold-change in substitutions per site post-WGD. The raw difference between the fold-change between the two paralogous clades is used as our measure of sequence evolution asymmetry. Formally, we calculate a species tree on the whole dataset to obtain an expected rate of evolution ($\alpha_{\text{pre-WGD}}$, $\alpha_{\text{post-WGD}}$) and estimate the rates of evolution for a particular paralog pair ($\beta_{\text{pre-WGD}}$, $\beta_{\text{post-WGD}_1}$, $\beta_{\text{post-WGD}_2}$). Because some proteins evolve faster than others, we scale the calculated rates of evolution such that $\alpha_{\text{pre-WGD}} = \lambda \beta_{\text{pre-WGD}}$. Our final score for the rate asymmetry is given by $\text{abs}(\lambda \beta_{\text{post-WGD}_1} / \alpha_{\text{post-WGD}} - \lambda \beta_{\text{post-WGD}_2} / \alpha_{\text{post-WGD}})$.

2.2.10 Simulating paralog divergence

The computational framework was based on the Duplication-Mutation-Complementation model with elements of Escape from Adaptive Conflict model to explain evolutionary outcomes stemming from gene duplication.

The model created a gene with a fixed length; length represented the number of positions that can be subjected to a degenerative mutation and thus simulated the sequence length of a real gene. The model assigned hypothetical functions to contiguous regions of a gene. Functional regions were defined as those that act independently of their neighbouring regions, can be overlapping in sequence space and can be composed of one or more sequence-based domains with a varying distance between them. The number of functions was a parameter, whereas the
length of each functional region and its position were randomly generated. Once the gene was created, it was duplicated and the functions immediately after duplication and at steady-state after evolutionary divergence remain intact. Paralog evolution was simulated by introducing random degenerative mutations at a constant rate. Initially, paralog A or paralog B were selected with equal probability. Then, a position was randomly chosen for a mutation. By falling within a functional region, a mutation disabled the function of that paralog and the region was removed. Further mutations had three possible outcomes, [1] ‘non-functional’ when a mutation landed in a region devoid of functions, not resulting in functional perturbation, [2] ‘divergent’ when a mutation perturbed a function that can be carried out by the remaining sister paralog, and [3] ‘deleterious’ when a mutation perturbed a function which has also been inactivated in the sister paralog. A paralog pair was deemed to have reached steady-state when no more divergent mutations could take place. For each parameter that described the number of functions simulations included 50,000 duplicate pairs for which functions were allowed to overlap and 50,000 control pairs for which they were forbidden to overlap. Each group was evolved to steady-state and the number of functions partitioned to each paralog was counted.

2.2.11 Testing hypotheses about asymmetry, structural and functional overlap by modeling duplicate divergence
The evolution of 5,000 paralog pairs was simulated with a range of multi-functionality parameters (3-30) and binned by their initial structural entanglement. The number of pairs that reverted to singleton state were counted and their asymmetry measured with respect to the number of functions at steady-state and the number of retained functions per bin.
2.3 Results

2.3.1 Strain construction and estimation of single and double mutant query fitness on standard growth conditions

To uncover genes that buffer the loss of both paralogs, I systematically assessed 275 double mutants for triple mutant genetic interactions using an optimized triple mutant SGA method which I describe in the next few sections. Genes chosen for analysis represent *Saccharomyces cerevisiae* dispensable paralogs originating from the whole genome duplication (WGD) event or from small-scale duplications (SSD), 242 and 33 gene pairs, respectively. In total, I constructed 550 single and 275 double mutant strains (Appendix 1) and assessed their growth using a highly quantitative growth assay conducted in 6 replicates. Analysis of the fitness distributions for single and double mutants [Figure 2.4 (top panels)] showed that most query strains were characterized by fitness value between 0.9 and 1. The measurements correlated with measurements of single ($r = 0.54, p = 0$) and double ($r = 0.63, p = 0$) mutants of a previous large-scale study [(Costanzo et al. 2010); Figure 2.4 (bottom panels)]. Differences between fitness measurements may reflect variation in experimental procedure and the fact that most of the strains had near wild-type growth rates. The collection of double mutants strains deleted for non-essential paralogs represents an important resource for the yeast community for further phenotypic characterization and evolutionary analyses.

![Image of fitness distributions](image.png)

*Figure 2.4. Quantification of fitness of query strains using high-density arrays.* Top panels: (left) histogram of double mutant fitness, n=237; (right) histogram of single mutant fitness, n=554. Bottom panels: (left) scatter plot comparing double mutant fitness derived from this study and Costanzo et al. 2010, n=194; (right) scatter plot comparing single mutant fitness derived from this study and Costanzo et al. 2010, n=526.
2.3.2 Optimization of Synthetic Genetic Array (SGA) for triple mutants

Previously our lab has employed SGA to study double mutant interactions, but the approach required optimization for screening for triple mutant interactions. I tested three approaches (see Figure 2.5) for selection of triple mutants during the final steps of SGA. In a standard SGA experiment, each gene deletion is selected sequentially and so I first extended this approach to triple mutant selection by pinning to media selecting first for one single mutant, followed by double and triple mutant progeny: kanR → Ura+/kanR → natR/Ura+/kanR reducing each time the number of progeny by 2-fold. I also tested the efficacy of triple mutant selection when one of the selection steps was eliminated. Single mutant selection followed by direct selection for triple mutants (kanR → natR/Ura+/kanR) reduced the progeny by 2-fold followed by a 4-fold reduction, which produced uneven colonies that were difficult to score. However, selection for double, then triple mutants (Ura+/kanR → natR/Ura+/kanR), which reduced the progeny by 4-fold and then by 2-fold, produced round colonies that were easily scored. I saw a high correlation of normalized colony sizes (Pearson’s correlation coefficient $r = 0.84$) and binned colony size variability was not significantly different when the single mutant selection step was eliminated (Student’s t test, $p = 0.66$). Elimination of one selection step proved to be more time and cost effective and thus I used the third approach for all triple mutant screens described in this thesis.
Triple mutants have not been extensively studied in a high-throughput manner. To date, a total of 8 double mutant queries have been screened against single mutant arrays to identify trigenic interactions using scoring methods that were either qualitative or semi-quantitative (Tong et al. 2004; Zou et al. 2009; Moir et al. 2012; Haber et al. 2013; Lai et al. 2013). For example, Haber

2.3.3 Quantitative fitness-based model for trigenic interactions

Figure 2.5. Optimization of selection steps for triple mutant SGA. Approach #1 sequentially selects for one marker at a time. Approaches #2 and #3 eliminate one of the steps. G418 – geneticin (selection for kanMX4); NAT – nourseothricin (selection for natMX4); Arg – arginine; Lys – lysine; Can – canavanine; Lyp – thialysine; Ura – uracil (selection for URA3). Haploid selection – SDmsg – Arg/Lys + Can/Lyp.
et al. 2013 used E-Map derived S-scores to subtract the stronger of the two double mutant interactions from the triple mutant effect to generate a trigenic interaction score:

\[ \text{MinDC} = S_{ij,k} - \min(S_{i,k}, S_{j,k}) \]  
\[ \text{(Eq. 3.1)} \]

However, for the trigenic model to quantify differences in the similar manner as the digenic interaction model it needs to be refined under an equivalent framework. The digenic interaction score under the multiplicative model is

\[ \varepsilon_{ij} = f_{ij} - (f_i f_j) \]  
\[ \text{(Eq. 3.2)} \]

where \( \varepsilon \) is the digenic interaction score, \( f_{ij} \) is the observed fitness of the double mutant and \( f_i f_j \) is the expected fitness based on the product of single mutants. In this model, the expected fitness of a triple mutant is \( f_{ijk \text{ expected}} = f_i f_j f_k \), but to remove influence from cases where two of the genes are not independent, the expected fitness would be the product of the interacting double mutant and the unrelated single mutant \( f_{ijk \text{ expected}} = f_{ij} f_k \). Expressing this equation in terms of single mutant fitness terms only provides the following:

\[ f_{ijk \text{ expected}} = (f_i f_j + \varepsilon_{ij}) f_k \]  
\[ \text{(Eq. 3.3)} \]

It is evident that the expectation contains the digenic interaction effects that are scaled by the fitness of non-interacting genes. As a result a trigenic interaction score would be:

\[ \tau_{ij,k} = \text{observed} - \text{expected} = f_{ijk} - (f_i f_j + \varepsilon_{ij}) f_k \]  
\[ \text{(Eq. 3.4)} \]

By accounting for the remaining two possible digenic interactions and after rearranging the terms for clarity, the expression for trigenic interactions becomes:

\[ \tau_{ij,k} = f_{ijk} - f_i f_j f_k - \varepsilon_{j,k} f_i - \varepsilon_{i,k} f_j - \varepsilon_{i,j} f_k \]  
\[ \text{(Eq. 3.5)} \]

Furthermore, we wanted to include terms expressing either single mutant fitness values or genetic interaction scores obtained from the available single, double and triple mutant screens, in which case

\[ \varepsilon_{ij,k} = f_{ijk} - f_i f_j \]  
\[ \text{(Eq. 3.6)} \]
By solving the latter two equations (3.5 and 3.6) for \( f_{ijk} \) and setting them equal to each other, we can solve for the trigenic interaction term from known quantities:

\[
\tau_{i,j,k} = \epsilon_{ij,k} - \epsilon_{ik} - \epsilon_{j,k}f_i \quad \text{(Eq. 3.7)}
\]

The resulting equation (3.7) is similar to Haber et al. 2013 in cases where single mutant fitness estimates are 1 as is the case for most duplicates, however models diverge in cases where such estimates are not 1. The error term can be expressed in the following manner:

\[
\text{error} = S - \epsilon = (\epsilon_{1,2} - \epsilon_2) - (\epsilon_{1,2} - f_1\epsilon_2 - f_2\epsilon_1) = \epsilon_2(f_1-1) + f_2\epsilon_1 \quad \text{(Eq. 3.8)}
\]

In cases where digenic interactions of both query genes have significant negative terms (see Figure 2.6a lower left corner), the S score will consistently produce a lower quantity than \( \tau \) because it subtracts the most negative term rather than both terms, causing the trigenic interaction to appear more negative than it should be. On the other hand, when one of the digenic interactions is positive (Figure 2.6a right side of the graph), the S score will consistently produce a higher quantity than \( \tau \) causing the trigenic interaction to appear more positive than it should be. Figure 2.6b shows simulated data from a hypothetical screen (scatter plot) and the magnitude of the error term when single mutant fitness estimates are 1 (top heatmap) compared to when they differ from 1 (bottom heatmap), resulting in a higher error term which is shown by the larger area covered with dark colours on either extreme of the spectrum.
2.3.4 Evaluation of control screens

To quantify the expected fitness of the triple mutants using the multiplicative model, I needed accurate estimates of the fitness of the array mutants. Since the double mutant query strains carry deletions in two genes marked with natMX4 and URA3, the wild type screens that were done to obtain the array fitness needed to have a matching set of benign loci also similarly marked. A previous high-throughput study in our lab (Baryshnikova et al. 2010b) determined array fitness by screening a wild-type control query strain in which the natMX4 marker replaced the URA3 gene (ura3Δ::natMX4) against the array of gene deletion mutants. I compared the published results to a comparable screen done using the wild-type control strain constructed for comparison to triple mutant screens (ura3Δ::natMX4 hoΔ::URA3) and found that array strain fitness measurements were highly correlated ($r = 0.97$, $r^2 = 0.94$, Figure 2.7), suggesting that a deletion in the second locus was benign as expected and exerted little/no effect on array fitness. However, since we are utilizing a quantitative model for scoring trigenic interactions, we chose to be conservative and conduct digenic and trigenic screens in the same context (selecting for natR, Ura+, kanR phenotypes) to minimize the introduction of noise into the system.

---

**Figure 2.6. Comparison of $\tau$ to S-score approaches for quantifying trigenic interactions.** A) Cartoon depiction of the behavior of $\tau$ and the S-score when that both digenic interactions corresponding to the double mutant query genes are significantly negative (lower left) or when one of the digenic interactions is significantly positive (right). B) Simulated data from a hypothetical screen (scatter plot) and the magnitude of the error term when single mutant fitness estimates are 1 (top heatmap) compared to when they differ from 1 (bottom heatmap).
2.3.5 Array miniaturization for high-throughput screening of triple mutants

Generating a trigenic interaction score requires four independent screens: a wild-type query crossed to mutant arrays for measuring single mutant fitness (see above); both single mutant query strains crossed to mutant arrays to measure specific digenic interactions; and the corresponding double mutant query crossed to single mutant arrays for measuring triple mutant fitness. The cost and logistics of doing such extensive screening prompted us to select a subset of strains for our screens – a ‘diagnostic array’. This subset of genes was selected computationally to cover a range of single mutant fitness, genetic interaction degree, and represent major biological processes in the cell (Figure 2.8a-c). It comprised of 1,032 non-essential gene deletion mutants and 200 strains carrying temperature-sensitive (ts) alleles of essential genes covering ~20% of the genome. The digenic interactions associated with the diagnostic array mutant strains re-capitulated functional profiles seen in our genome-wide reference map (Costanzo et al. 2010) indicating that the selected strains provide a representative view of the global digenic interaction network (Figure 2.8d). Perhaps most importantly, these strains exhibited highly reliable and
robust behaviour in our SGA methodology. That is, these strains typically showed strong and reproducible genetic interactions, thus reducing both our false positive and our false negative rates, statistically sound identification of trigenic interactions. While the use of a subset of array strains does not provide a complete view of trigenic interactions, it does provide a feasible means of surveying trigenic interactions in an efficient and quantitative manner. Ultimately, the results obtained from this representative study can be extrapolated to the entire genome.
Figure 2.8. Diagnostic array characteristics. Diagnostic array captures the distribution of single mutant fitness (A) and digenic interaction degree (B) of array strains that exhibit trigenic interactions in which ‘Diagnostic array’ includes only strains chosen for inclusion on the diagnostic array; ‘All strains’ represents all essential and non-essential strains on the full arrays; ‘Trigenic interactions’ includes strains that are involved in trigenic interactions. C) Coverage of biological processes based on in-house terms (Costanzo et al. 2010) across the genome and on the diagnostic array. D) The ability of the selected diagnostic strains (left panel: 1032 non-essential gene deletion mutants; right panel: 200 strains carrying temperature sensitive alleles of essential genes) to re-cluster the digenic interaction profile correlation matrix. A true positive (TP) was defined as a gene pair, which has a high profile similarity and is co-annotated.
To assess the suitability of array miniaturization for functional characterization of duplicated genes, I screened 10 double mutant query strains including their cognate control strains against three types of arrays: [1] an existing array of 1216 strains carrying 1016 temperature sensitive alleles of essential genes, covering 686 unique essential genes (Li et al. 2011) as well as 200 non-essential gene deletion strains (TSv6); [2] the miniaturized 4-plate array described in the previous section (Diagnostic–1200); and [3] an array, which is half its size (Diagnostic–600). The strains composing these arrays are listed in Appendix 2. In general the genetic interaction profiles obtained from the reduced arrays were of similar quality to those obtained from screening the full genome deletion and ts arrays (Figure 2.9a). Also, the discovery of genetic interactions is not drastically compromised across array types since the median number of genetic interactions was similar (Figure 2.9b).

I next tested the frequencies of digenic and trigenic interactions detected in SGA screens using the three arrays (TSv6, Diagnostic-1200, Diagnostic-600) compared to those seen by screening the full deletion array (SGAv2) (Figure 2.10). I observed similar yet discreet patterns of frequency of trigenic and digenic interactions for both highly functionally similar (e.g. $CLN1$-$CLN2$, $SBE22$-$SBE2$ and $MLP1$-$MLP2$) and more divergent pairs ($SKI7$-$HBS1$, $DPB3$-$DLS1$ and $BCH2$-$CHS6$). Therefore, reducing the size of the array up to 4 times continues to enable the classification of paralogs based on their level of functional divergence (see Section 2.3.8 for more details).

However, there was a difference among the array types with regards to the functional information that they provide. GO enrichment analysis (Figure 2.11) suggests that screens done using the Diagnostic-1200 were as functionally informative as SGAv2, and outperformed Diagnostic-600 and TSv6. TSv6 contains mostly essential genes, which are known to be pleiotropic and may contribute to the loss of specific genetic interaction profiles important to dissect the functional relationship of sister duplicates. On the other hand, Diagnostic-600 may not contain a sufficient number of genes to observe biological patterns in genetic interaction profiles and achieve a significant GO enrichment signal. Therefore, Diagnostic-1200 array is the array of choice for the future experiments.
Figure 2.9. Technical comparison of array types. A) Genetic interaction profile correlations across array types as measured by Pearson correlation coefficient (r). The array types are described in the text. B) Genetic interaction discovery across array types, score < -0.08, pval < 0.05 for trigenic (left panel) and digenic (right panel) interactions.

Figure 2.10. Triple and double mutant interaction fraction of preliminary paralog pairs across different array types. Genetic interaction (GI) degree is based on an intermediate cut-off (SGA score < -0.08). SGAv2 refers to the full-genome non-essential gene deletion mutant set; TSv6 is the lab’s largest collection of temperature sensitive gene deletion mutants; Diagnostic-1200 is a reduced array of 1000 non-essential gene deletion mutants and 200 temperature sensitive (ts) alleles of essential genes. Diagnostic-600 array consists of 500 non-essential gene deletion mutants and 100 temperature sensitive (ts) alleles of essential genes. Shared double mutant interactions are shown in red. Functionally similar pairs are in the top row, divergent pairs are in the bottom row sorted by their fraction of triple mutant interactions relative to total number of genetic interactions of a paralog pair. \( \text{par1} \Delta \text{par2} \Delta \) – denotes a double mutant query strain; \( \text{par1} \Delta \) – paralog-1 single mutant query strain; \( \text{par2} \Delta \) – paralog-2 single mutant query strain.
Figure 2.11. Functional analysis across array types. Functional information using enrichment analysis of genetic interactions of ten paralog pairs across various array types. p-value is denoted in yellow. In-house biological process terms are from Costanzo et al. 2010.
2.3.6 Evaluation of reproducibility of genetic interactions

Scoring trigenic interactions presents a multiple observations challenge. As mentioned earlier each trigenic score is influenced by four observations: single mutant fitness, digenic interactions of individual mutants and the triple mutant effect on fitness itself. To confidently identify trigenic interactions, we needed to ensure that the triple mutant effect on fitness was not a false positive and similarly that digenic interaction screens did not produce false negatives. To address these issues, we need to consider that interactions discovered in each triple mutant screen (before the quantitative adjustment for digenic interactions) would include: a) real trigenic interactions, b) real digenic interactions that were true positives in the control double mutant screens; c) real digenic interactions that were false negatives in the control double mutant screens; d) false positives. To calculate each of these parameters, we reasoned that the number of true trigenic interactions (a) depends on the number of observed trigenic interactions (T) and the recall of a screen (R), while the number of real digenic interactions that were detected in triple mutant screens and were true positives in the control double mutant screens relates to $R^2$. The number of real digenic interactions that were detected in triple mutant screens but were missed in the control double mutant screens depends on the number of the observed digenic interactions (D) and recall (R).

\[
(\text{Real trigenic interactions}) \ a = R \times T \quad \text{(Eq. 3.9)}
\]

\[
(\text{Real digenic interactions, observed in controls}) \ b = R^2 \times D
\]

\[
(\text{Real digenic interactions, missed in controls}) \ c = R \times D \times (1-R)
\]

\[
(\text{False positives}) \ d = (a + b + c) \times (1-P)/P
\]

Since a, b, c encompass real genetic interactions then technical precision of a screen can be expressed as:

\[
\text{Technical screen precision (P)} = \frac{(a + b + c)}{(a + b + c + d)} \quad \text{(Eq. 3.10)}
\]

After removing digenic interactions that were observed in double mutant screens, the precision of a trigenic interaction screen is:

\[
\text{Trigenic interaction precision} = \frac{(a)}{(a + c + d)} \quad \text{(Eq. 3.11)}
\]
Assuming that the number of trigenic interactions is on the same order as the number of digenic interactions then we can use equation 3.9 to re-write equation 3.11 in terms of precision and recall of digenic screens of individual mutants.

\[
\text{Trigenic interaction precision} = \frac{P}{P(2-R) + 2(1-P)} \quad (\text{Eq. 3.12})
\]

Empirical estimates of technical precision and recall of SGA screens are affected by the number of replicates (Figure 2.12a,b). Single replicate precision and recall of a digenic interaction screen of 50% and 40%, respectively, resulted in an estimated 28% precision for identifying trigenic interactions. However, performing screens in multiple replicates increased precision at little cost to recall. As such, conducting two replicates results in precision and recall of a digenic interaction screen of 80% and 38%, respectively, elevating trigenic screen precision to ~50%, similar to the precision achieved with one digenic interaction screen replicate reported in another large-scale SGA study (Costanzo et al. 2010). Figure 2.12c depicts the recovery of negative digenic interactions in unadjusted triple mutant genetic interaction profiles behaves similarly to replicates. It shows that profiles with poor recovery of digenic interactions also have very few interactions and for the majority of the screens that result in high degree recover the most number of interactions. Given these considerations, all trigenic interaction screens have been performed twice.

**Figure 2.12. Replicate analysis of SGA screens.** A) Precision and B) recall are calculated based on the ‘Gold Standard’ of true positives that consists of any interaction called significant at the intermediate threshold, \( \epsilon < -0.08 \). Queries (n=31) are scored with varying number of replicates using the established SGA score method (Baryshnikova et al. 2010a). C) Detection of digenic interactions in unadjusted triple mutant genetic interaction profiles. Black – fraction of digenic screens showing the fraction of negative digenic interactions recovered in triple mutant screens. Grey – average negative digenic interaction degree plotted against the fraction negative digenic interactions recovered in triple mutant screens.
The reproducibility of genetic interactions correlates with their magnitude such that the strongest genetic interactions are more likely to be detected in multiple replicates (Figure 2.13, left panel). To ask if this effect may explain the ability to detect expected digenic interactions in triple mutant screens, I compared the distribution of their digenic and trigenic interaction scores. The distribution of digenic interactions that were detected in triple mutant screens was skewed towards stronger negative genetic interactions compared to those digenic interactions that are undetected in triple mutant screens (Figure 2.13, right panel). These results suggest that stronger digenic interactions are more likely to re-appear in triple mutant screens and may explains why not all significant genetic interactions of individual paralogs are re-discovered in the triple mutant screens.

![Figure 2.13. Reproducibility of genetic interactions depends on their magnitude.](image)

**Figure 2.13. Reproducibility of genetic interactions depends on their magnitude.** Left) Control digenic interaction screens (n=19) with 4 replicates were combined in random pairing of 2 replicates and scored. Average fold enrichment for overlapping negative genetic interactions across 3 combinations of data sets is shown. Right) Distribution of digenic interactions that are detected and undetected in triple mutant screens. The entire digenic duplicate dataset is plotted (275 triple mutant screens and 550 double mutant screens were used for the analysis).

### 2.3.7 Validation of trigenic interactions

I validated trigenic interactions using a set of manually confirmed synthetic sick or lethal interactions discovered through independent screening of *cln1Δ cln2Δ* double mutant for trigenic interactions (Zou et al. 2009). I retested the 19 overlapping interactions discovered in the
published screen and in my screens by tetrad analysis to obtain a high confidence set of 14 synthetic sick or lethal interactions (3 failed to confirm and 2 were inconclusive, Table 2.2). Thus, I restricted my analysis to 14 interactions of which 10 showed a significant negative trigenic interaction at an intermediate cut-off ($\tau \leq -0.08$) resulting in a true positive rate of 71% and a false negative rate of 29% (see Table 2.2), consistent with other high-throughput SGA experiments (Costanzo et al. 2010). I used random spore analysis to further validate resulting trigenic interactions. Screening $cln1\Delta \ cln2\Delta$ against the reduced array produced 73 negative trigenic interactions at an intermediate cut-off, $\tau \leq -0.08$. I selected an arbitrary set for confirmations that have not already been confirmed by tetrad analysis and found that 30 of 38 interactions confirmed, generating a true positive rate of 79% and false positive rate of 21% (see Table 2.2).

### Table 2.2. Validation of $cln1\Delta \ cln2\Delta$ trigenic interactions by tetrad and random spore analyses (TA and RS, respectively).

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Systematic name</th>
<th>$\tau$</th>
<th>p-value</th>
<th>Confirmation test</th>
<th>Confirmation result</th>
</tr>
</thead>
<tbody>
<tr>
<td>APL6</td>
<td>YGR261C</td>
<td>-0.25</td>
<td>6.55x10$^{-3}$</td>
<td>RSA</td>
<td>inconclusive</td>
</tr>
<tr>
<td>APS3</td>
<td>YJL024C</td>
<td>-0.13</td>
<td>1.11x10$^{-2}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>BEM2</td>
<td>YER155C</td>
<td>-0.42</td>
<td>1.96x10$^{-4}$</td>
<td>TA</td>
<td>SL</td>
</tr>
<tr>
<td>BUD2</td>
<td>YKL092C</td>
<td>-0.69</td>
<td>1.60x10$^{-9}$</td>
<td>RSA, TA</td>
<td>SS/SL</td>
</tr>
<tr>
<td>BUD6</td>
<td>YLR319C</td>
<td>-0.30</td>
<td>1.75x10$^{-2}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>CLB5</td>
<td>YPR120C</td>
<td>-0.45</td>
<td>1.48x10$^{-2}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>CLN3</td>
<td>YAL040C</td>
<td>-0.61</td>
<td>5.34x10$^{-13}$</td>
<td>RSA, TA</td>
<td>SS/SL</td>
</tr>
<tr>
<td>CSF1</td>
<td>YLR087C</td>
<td>-0.21</td>
<td>7.03x10$^{-4}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>CYK3</td>
<td>YDL117W</td>
<td>-0.20</td>
<td>3.81x10$^{-6}$</td>
<td>RSA, TA</td>
<td>SS/SL</td>
</tr>
<tr>
<td>DBF2</td>
<td>YGR092W</td>
<td>-0.62</td>
<td>1.51x10$^{-6}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>EAF7</td>
<td>YNL136W</td>
<td>-0.20</td>
<td>3.57x10$^{-4}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>EDE1</td>
<td>YBL047C</td>
<td>-0.20</td>
<td>4.91x10$^{-2}$</td>
<td>RSA</td>
<td>inconclusive</td>
</tr>
<tr>
<td>FYV10</td>
<td>YIL097W</td>
<td>-0.14</td>
<td>1.67x10$^{-2}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>GAS1</td>
<td>YMR307W</td>
<td>-0.29</td>
<td>1.67x10$^{-6}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>GBP2</td>
<td>YCL011C</td>
<td>-0.24</td>
<td>4.02x10$^{-2}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>GET2</td>
<td>YER083C</td>
<td>-0.23</td>
<td>2.14x10$^{-4}$</td>
<td>TA</td>
<td>SL</td>
</tr>
<tr>
<td>GUP1</td>
<td>YGL084C</td>
<td>-0.27</td>
<td>4.33x10$^{-2}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>HSL1</td>
<td>YKL101W</td>
<td>-0.56</td>
<td>9.57x10$^{-10}$</td>
<td>RSA</td>
<td>SS</td>
</tr>
</tbody>
</table>
| Gene   | Protein | Score | Interaction | Function | PAR
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LGE1</td>
<td>YPL055C</td>
<td>-0.32</td>
<td>2.72x10^-4</td>
<td>TA</td>
<td>SS</td>
</tr>
<tr>
<td>LSM1</td>
<td>YJR124C</td>
<td>-0.55</td>
<td>4.24x10^-6</td>
<td>RSA,TA</td>
<td>SS/SL</td>
</tr>
<tr>
<td>LSM7</td>
<td>YNL147W</td>
<td>-0.28</td>
<td>1.71x10^-3</td>
<td>RSA</td>
<td>not confirmed</td>
</tr>
<tr>
<td>NBP2</td>
<td>YJR162C</td>
<td>-0.38</td>
<td>1.08x10^-7</td>
<td>TA</td>
<td>inconclusive</td>
</tr>
<tr>
<td>NUP188</td>
<td>YML103C</td>
<td>-0.23</td>
<td>1.21x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>PAT1</td>
<td>YCR077C</td>
<td>-0.40</td>
<td>2.65x10^-3</td>
<td>RSA</td>
<td>inconclusive</td>
</tr>
<tr>
<td>PEP8</td>
<td>YJR053C</td>
<td>-0.38</td>
<td>3.66x10^-4</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>RTF1</td>
<td>YGL244W</td>
<td>-0.24</td>
<td>1.33x10^-2</td>
<td>TA</td>
<td>SS</td>
</tr>
<tr>
<td>RUD2</td>
<td>YOR216C</td>
<td>-0.32</td>
<td>2.45x10^-4</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>RXT2</td>
<td>YBR095C</td>
<td>-0.12</td>
<td>4.61x10^-2</td>
<td>RSA</td>
<td>inconclusive</td>
</tr>
<tr>
<td>SAC7</td>
<td>YDR389W</td>
<td>-0.17</td>
<td>4.91x10^-2</td>
<td>RSA</td>
<td>inconclusive</td>
</tr>
<tr>
<td>SAP190</td>
<td>YKR028W</td>
<td>-0.25</td>
<td>8.76x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>SDC1</td>
<td>YDR469W</td>
<td>-0.40</td>
<td>1.74x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>SDS3</td>
<td>YIL084C</td>
<td>-0.41</td>
<td>2.03x10^-2</td>
<td>RSA</td>
<td>inconclusive</td>
</tr>
<tr>
<td>SM11</td>
<td>YGR229C</td>
<td>-0.37</td>
<td>7.92x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>SNX4</td>
<td>YJL036W</td>
<td>-0.25</td>
<td>1.15x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>SRN2</td>
<td>YLR119W</td>
<td>-0.38</td>
<td>7.84x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>SWD1</td>
<td>YAR003W</td>
<td>-0.25</td>
<td>1.88x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>SWI4</td>
<td>YER111C</td>
<td>-0.24</td>
<td>7.76x10^-9</td>
<td>TA</td>
<td>not confirmed</td>
</tr>
<tr>
<td>TLG2</td>
<td>YOL018C</td>
<td>-0.16</td>
<td>2.73x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>VAC14</td>
<td>YLR386W</td>
<td>-0.38</td>
<td>1.64x10^-2</td>
<td>RSA</td>
<td>inconclusive</td>
</tr>
<tr>
<td>VAM6</td>
<td>YDL077C</td>
<td>-0.49</td>
<td>8.21x10^-8</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>VAM7</td>
<td>YGL212W</td>
<td>-0.38</td>
<td>1.55x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>VPS29</td>
<td>YHR012W</td>
<td>-0.32</td>
<td>3.21x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>VPS35</td>
<td>YJL154C</td>
<td>-0.48</td>
<td>7.55x10^-6</td>
<td>RSA</td>
<td>SS</td>
</tr>
<tr>
<td>VPS60</td>
<td>YDR486C</td>
<td>-0.29</td>
<td>1.94x10^-2</td>
<td>RSA</td>
<td>SS</td>
</tr>
</tbody>
</table>

### 2.3.8 Triple mutant interaction fraction: functional divergence metric

Having optimized the scoring protocol and procedure for triple mutant screens, I next assessed the data produced by screening the 242 WGD and 33 SSD paralog pairs that I chose as described in Section 2.2.1 against the Diagnostic – 1200 array (Appendix 3). In total I interrogated 660,000 double and 330,000 triple mutants to map 5,585 digenic and 2,766 trigenic interactions. I first decided to use my dataset to create a metric to quantify the spectrum of functional overlap between members of paralog pairs. In particular, I reasoned that the number of triple mutant interactions relative to double mutant interactions for a given duplicate gene pair might provide a
direct measure of functional overlap (Figure 2.14). I predicted that the degree of functional overlap shared between members of duplicate gene pairs should range from complete to little or no functional overlap, a relationship that should be captured in their genetic interactions. For example, highly divergent gene pairs should exhibit a relatively high number of paralog-specific interactions and since the sisters do not share functions or a buffering relationship, there would be very few triple mutant interactions (Figure 2.14 left panel). In contrast, duplicate genes with highly overlapping functions should largely exhibit triple mutant interactions found only in the mutant deleted for both paralogs, with few paralog-specific interactions (Figure 2.14 right panel). Thus, we used a metric to quantify the complete spectrum of functional overlap based on computing the ratio of triple mutant interactions to the total number of all interactions of the paralog pair.

Indeed the resulting genetic interaction profiles revealed a spectrum of functional redundancy detected in the screened paralog pairs (Figure 2.15a). For the 242 WGD paralog pairs that were assessed, 175 pairs showed more than six triple or double mutant interactions, with a median fraction of triple mutant interactions of 0.2, placing 45 pairs in the fourth quartile as they exhibited a triple mutant interaction fraction equal to or exceeding 0.4 indicating retention of a varying degree of functional redundancy (Appendix 7). For the 33 SSD paralog pairs that were assessed, 23 showed more than six triple or double mutant interactions, with 3 pairs exhibiting a triple mutant interaction fraction equal to or exceeding 0.4 (Appendix 7). For WGD and SSD paralogs exhibiting such a gradient of triple mutant interactions, approximately 65% of profiles correlated highly with other double mutant profiles belonging to specific functional clusters which was also consistent with evidence reported in the literature regarding their biological roles as described below, whereas the remaining triple mutant profiles correlated either to unknown genes or functionally incoherent gene sets belonging to diverse gene sets probably due to poorly characterized or pleiotropic functions (see Appendix 6). The remaining 67 WGD and 10 SSD paralog pairs, representing about a third of all screened pairs, were characterized by sparse genetic interaction profiles. Interestingly, they belonged to diverse set of biological processes rather than a specific process and showed a tendency to belong to larger gene families ($p = 0.0014$, unpaired t-test), probably indicative of higher order redundancy. Since WGD pairs represent a body of paralogs of identical age, I focused on their comparative analysis with regard to their patterns of genetic interactions.
To test the relationship between triple mutant interaction fraction and functional overlap of paralog pairs, I first focused on well-characterized paralogs (Figure 2.15b). For example, *SKI7-HBS1* is an example of a divergent paralog pair (van Hoof 2005). *SKI7* encodes a protein, which is involved in 3’ to 5’ mRNA degradation (Araki et al. 2001) while *HBS1* encodes a protein, which is involved in recycling of ribosomes when translation stalls occur (Shoemaker et al. 2010). As predicted by my model, this divergent pair exhibited a triple mutant interaction fraction of less than 0.2 with paralog-specific interactions greater than 0.8. Consistent with their known functions, *SKI7* interacted with genes involved in mRNA decay pathways such as *LSM1* and *PAT1*, while *HBS1* interacted with genes controlling ribosome biogenesis (Figure 2.15b).

The *SWI5-ACE2* pair encodes cell cycle-regulated transcription factors that are transcribed during G2 phase of the cell cycle and accumulate in the nucleus at similar times in early M phase to turn on target genes involved in G1-specific gene expression (Swi5p) and daughter cell-specific expression of cell separation genes (Ace2p), among others (Nelson et al. 2003). Consistent with partially overlapping functions *SWI5-ACE2* exhibited a medium fraction of triple mutant interactions (~0.4) and a higher fraction of paralog-specific interactions (~0.6). Their triple mutant interactions occurred with genes involved in exit from mitosis, such as *CMD1*, *DBF2*, *BUD2*, *BUD3*, *BUD4*, and DNA replication, such as *RFA1*, *MRC1*, *CSM3*, with specific interactions most likely reflecting their distinct regulation and promoter specificities. For instance, *SWI5*-specific interacting partners included genes related to DNA replication including members of origin recognition complex (*ORC1*, *ORC3*, *ORC4*, *ORC6*) and the Mcm2-7 complex (*MCM2*, *MCM5*, *MCM7*), whereas *ACE2*-specific interactions included a septin complex component (*CDC11*), the cell cycle kinase Cdc28/Cln3 complex (*CLN3*), the cell cycle phosphatase Sap190/Sit4 complex (*SAP190*), and the Bud14/Glc7 complex (*BUD14*) (Figure 2.15b) (Dohrmann et al. 1992; McBride et al. 1999; Doolin et al. 2001).

As a final example, the *SBE2-SBE22* paralog pair encodes proteins with highly similar roles in the transport of cell wall components to growth sites (Santos and Snyder 2000). The *sbe2Δ sbe22Δ* double mutant showed few paralog specific and many triple mutant interactions, with their triple mutant interaction fraction being as high as ~0.8 enriching for GO fringe terms related to vesicle mediated transport ($p = 2.18 \times 10^{-8}$) and cell wall organization ($p = 3.137 \times 10^{-3}$) (Figure 2.15b), which is consistent with their shared biological roles. These findings suggest that triple mutant interaction fraction reflects the functional relationship of paralogs.
Interestingly, a poorly characterized pair of paralogs, *ECM13-YJR115W*, was among the most highly interacting double mutant queries. This pair was characterized by a triple mutant interaction fraction of 0.82, suggesting high functional redundancy (Figure 2.15b). Both proteins contain Pfam regions of low complexity and show low expression (Holstege et al. 1998). Their triple mutant profile showed high similarity with genetic interaction profiles of *TUB3, TCP1* and *CCT4* (Figure 2.16), which are involved in tubulin assembly consistent with reported protein-protein interactions of Ecm13p with outer kinetochore proteins such as Ctf19p and Mcm22p (Wong et al. 2007), suggesting a microtubule-related role for this paralog pair.
Figure 2.14. Mapping functional relationship of paralogs through their double and triple mutant genetic interactions. Highly divergent paralogs are expected to display fewer trigenic interactions compared to the number of paralog-specific genetic interactions. However, as the functional redundancy of paralogs increases, then so should the fraction of trigenic interactions with a corresponding drop in the fraction of paralog-specific genetic interactions. Hypothetical biological functions are depicted as circles and the fraction of different types of genetic interactions is illustrated using bar graphs. The colours of the hypothetical function specific to each paralog (yellow, blue) and a shared function (green) correspond to the genetic interactions that should be uncovered in each mutant.

Figure 2.15. Functional divergence of duplicated genes represented using triple mutant interaction fraction. A) Triple mutant interaction fraction distribution of screened paralogs. B) Representative examples of genetic interaction profiles for paralogs. SKI7-HBS1 represents a highly divergent pair with SKI7 involved in 3'-5' mRNA degradation and HBS1 in recycling ribosomes when translation stalls occur. SWI5-ACE2 encode cell cycle transcription factors that are functionally redundant with some divergent characteristics. SBE22-SBE2 represent a functionally redundant pair involved in transport of cell wall components from Golgi to the cell surface. ECM13-YJR115W represent a poorly characterized functionally redundant pair. Analysis is restricted to paralogs with at least 6 triple or double mutant interactions.
Figure 2.16. *ECM13-YJR115W* triple mutant interaction profile correlations against the genome. Profile similarity was measured by Pearson correlation coefficient. Double mutant interaction profiles of non-essential gene deletion mutants and temperature sensitive alleles of essential genes were used in the analysis (Costanzo et al. 2010). Genes with profile similarity > 0.2 are listed and represent functionally informative predictions.
2.3.9 SSD and WGD paralogs contribute differently to mutational and environmental robustness

Having interrogated paralogs that originated by different duplication mechanisms, I next explored the evolutionary fates of SSD and WGD paralogs. The lower triple mutant interaction degree for SSD compared to WGD paralogs ($p = 0.048$) (Figure 2.17a) suggests a greater level of functional redundancy of WGD paralogs. This result is consistent with earlier findings from integration of diverse genomic data showing that WGD paralogs are more likely than SSD to share physical protein interaction partners and functional relationships (Guan et al. 2007). To study the contribution of WGD and SSD genes to mutational and environmental robustness, we measured fitness of the entire mutant set of paralogs across alternative growth conditions. We surveyed responses to changes in carbon source, simulated starvation conditions by treatment with rapamycin and challenged with DNA damaging agents, cell wall disrupting agents and inhibitors of intracellular trafficking/ergosterol production/histone deacetylases/protein glycosylation/translation/proteasome (Table 2.1). These phenotypic tests revealed that SSD paralogs participated in significantly fewer chemical-genetic interactions than WGD paralog pairs ($p = 0.022$) (Figure 2.17b, Appendix 5) mirroring the trend towards fewer triple mutant interactions for SSD paralogs. These findings lend support to an earlier report involving the yeast metabolic network (Deutscher et al. 2006) that the evolution of mutational and environmental robustness are correlated and they do not evolve independently from one another.

![Figure 2.17](image)

**Figure 2.17. Small scale and whole genome duplicates contribute differently to mutational and environmental robustness.** SSD denotes small scale duplicates and WGD denotes whole genome duplicates. Negative genetic interactions, score $<-0.08$, pval $<0.05$, and negative chemical genetic interactions, score $<-40$, pval $<0.05$ are depicted. T bars, SEM.
2.3.10 Conditional triple mutant interactions reveal functional relationships of duplicated genes

Alternative growth conditions are also useful in providing insight into the biological roles of paralogs that are difficult to interrogate under standard growth conditions. Duplicated genes are significantly overrepresented among the ‘dark net’ (p < 0.05), which we define as a gene set that produces genetic interaction profiles which cannot be differentiated from background noise, posing a challenge in their functional characterization. PRM5-YNL058C is an example of a ‘dark net’ pair. They are poorly characterized, contain one transmembrane domain, localize to the vacuole and have been shown to be induced in cell integrity signaling (Jung and Levin 1999; Arias et al. 2011). We observed sensitivity of prm5Δ ynl058cΔ double mutant to caspofungin, an inhibitor of beta-1,3-D-glucan synthase, which is required for cell wall synthesis and postulated that since it was shown that genetic interaction degree correlates with strain fitness (Costanzo et al. 2010), it would be informative to conduct a triple mutant interaction screen on media containing this drug. In doing so we uncovered 22 conditional triple mutant interactions, which is an 11-fold increase over standard conditions (Figure 2.18a, Appendix 8). Conditional triple mutant interactions with members of the GET complex (GET2, GET3), prefoldin complex (YKE2) and ESCRT-I (SRN2) suggest a possible role of this pair in a process affecting protein sorting and vesicle trafficking (Figure 2.18b).

Another example is ESL1-ESL2, which represent another poorly characterized paralog pair. They contain PilT N-terminus ribonuclease domain but a direct role in RNA processing has not been shown (Bleichert et al. 2006; Lai et al. 2013). We observed sensitivity of esl1Δ esl2Δ double mutant to a translation inhibitor, cycloheximide, and screening this double mutant on such a condition enriched their triple mutant interaction profile by 3.5-fold producing 28 conditional triple mutant interactions (Figure 2.18a, Appendix 8). Their interaction with members of the spliceosome (CUS1) and the nucleotide excision repair machinery (RAD10, SSL1) depicted in Figure 2.18b are consistent with previous observations that implicated this pair in RNA degradation and possible contribution to the maintenance of genomic stability that is dependent on their nuclease domain (Bleichert et al. 2006; Lai et al. 2013). Thus, conditional triple mutant interactions offer important insight into the biological roles of paralogs.
2.3.11 Paralog retention varies by cellular process

I also assessed the distribution of triple and double mutant interactions across different cellular processes. The heatmap identified biological processes that displayed distinct preferences for retention of functionally divergent or redundant paralogs (Figure 2.19a). Pairs involved in RNA processing exhibited the highest triple mutant interaction fraction and degree (1.4 and 1.7 fold change over background, respectively), whereas DNA replication, repair and cohesion had the lowest (0.55 and 0.46 fold change over background, respectively). It is possible that the retention of RNA processing genes in duplicate contributed to increased robustness of the process. Alternatively, their retention may have been selected for dosage amplification conferring an evolutionary advantage resulting in their fixation, which is supported by their higher than average correlation of double mutant interaction profiles, which has been previously employed as a metric to distinguish duplicates under dosage selection (Ihmels et al. 2007; VanderSluis et al. 2010). For example, \textit{MLP1-MLP2}, a pair involved in nuclear retention of unspliced mRNA and \textit{JRN1-PUF2}, that encode RNA binding proteins that preferentially interact with mRNA encoding membrane-associated proteins both show a high triple mutant interaction fraction and appear to be under strong dosage selection. It is possible that increased RNA processing ability was evolutionary advantageous to sustain fast growth rate, which is consistent with pairs that

Figure 2.18. Conditional triple mutant interactions reveal insight into functions of paralogs with sparse profiles on standard growth conditions. A) \(\tau\) – Standard condition refers to the triple mutant interaction scores on standard SGA condition. Differential score refers to the strength of the growth defect of the triple mutant on standard SGA condition supplemented with 0.1 \(\mu\text{g/ml}\) caspofungin or 0.05 \(\mu\text{g/ml}\) cycloheximide. Growth defects due to conditional single mutants are depicted in blue, double mutants are in green and triple mutants are in red. B) Conditional triple mutant genetic interactions with members of protein complexes, whereby 50% or more of complex members genetically interact are displayed.
belong to ribosome/translation which also show strong dosage selection and average triple mutant interaction fraction. On the other hand, duplicate retention in the other processes, such as DNA replication and repair, evidently, allowed for functional specialization as evident from a relatively low triple mutant interaction fraction. For instance, $MCK1$-$YGK3$, which encode dual-specificity ser/thr and tyrosine protein kinases with roles in chromosome segregation, genome stability, phosphorylation-dependent protein degradation and transcriptional regulation have evolved differences related to catalytically important residues and active or co-factor binding sites (Turunen et al. 2009). Thus, maintenance of paralogous genes diversified the regulatory repertoire of $S.cerevisiae$ genes involved in certain processes, while simultaneously enhancing the robustness in others.

2.3.12 Physiological and evolutionary features correlate with triple mutant interaction fraction

The choice of triple mutant interaction fraction as a metric for functional divergence of paralogs was supported by associations with other fundamental physiological and evolutionary properties (Figure 2.19b). In particular, we observed that subcellular localization pattern similarity and low expression difference correlated with triple mutant interaction fraction, consistent with functional compensation being achieved by paralogs that are found in the same compartment with similar expression levels. Also, as expected from previous findings associated with single mutant query genes and double mutant genetic interactions (Costanzo et al. 2010), double mutants query strains characterized by a greater fitness defect and shorter path length on the genetic interaction network exhibited a relatively high triple mutant interaction fraction, indicative of their functional similarity. These findings support the prediction that triple mutant interaction fraction reflects the functional overlap of paralogs.
Figure 2.19. Physiological and evolutionary characteristics that correlate with fraction of triple mutant interactions. A) Average triple mutant interaction fraction, triple and double mutant interaction density across biological processes. The fraction of screened query-array pairs exhibiting negative interactions was measured for 12 broadly defined functional gene sets. A colour was assigned to each element reflecting the fold-increase over the background fraction of interactions (0.287), triple mutant interaction density (0.0125) and double mutant interaction density (0.0017) (blue, below the frequency of random pairs; black, statistically indistinguishable from the random background of interactions; and yellow, above the frequency of random pairs). A pair was annotated to a biological process if either one of the paralogs was annotated to the corresponding processes. Processes with at least 6 paralog pairs are displayed. B) Pearson correlation coefficient between fraction of triple mutant interactions and physiological and evolutionary properties was measured for negative interactions $< -0.08$, pval $< 0.05$. Significance was assessed using a permutation test. Analysis is restricted to paralogs with at least 6 triple or double mutant interactions.
2.3.13 Asymmetric evolution and retention of functional redundancy

To explore the factors that drive paralog retention leading to maintenance of a spectrum of functional redundancy, we considered the relation between asymmetry, structure and function. A negative correlation of triple mutant interaction fraction with measures of asymmetric divergence that is either sequence- or genetic interaction degree-based (Figure 2.19b) may indicate evolutionary constraints on paralogs divergence. In an attempt to explain asymmetric divergence of paralogs we developed a computational framework for simulating their evolutionary trajectories and explored the requirement for various fates of duplicated genes. The model created genes of fixed lengths with hypothetical functions assigned to contiguous regions and evolved them by introducing random degenerative mutations (Figure 2.20a). Functional regions were defined as acting independently of their neighbouring regions, allowed to be overlapping in sequence space and composed of one or more sequence-based domains with a varying distance between them. Results of the simulations showed that a model, which allows functions to overlap led to a broader bias in the evolution of asymmetric pairs at steady–state (Figure 2.20b), suggesting that initial functional/structural entanglement is required for asymmetric divergence of paralog pairs. For example, the \textit{SCO2-SCO1} pair encodes proteins that are required for copper delivery to cytochrome c oxidase. However, their divergence is reflected in a greater effect on mitochondrial function and stability of Sco1p than Sco2p (Gamberi et al. 2009), which is consistent with \textit{SCO1} exhibiting more paralog specific interactions than \textit{SCO2}. The pair also has a shared role revealed by triple mutant interactions with genes involved in mitochondria and metabolism, such as \textit{TSA1}, \textit{GLO4} and \textit{RGI2} (Figure 2.20 inset). Thus, evolution of paralog specialization is not independent of retention of some common functionalities.
Figure 2.20. The evolution of retained overlap and asymmetry. A) Examples of duplicate divergence modeling. Duplicates with 5 functions (rows) carried out by variously sized, potentially overlapping, regions of sequence (dark bands). Immediately after duplication both paralogs carry out all functions. The pair evolves through random mutations until it reaches an evolutionarily stable-state which can sustain no further mutations without a loss of function. Top panel shows a pair that has reverted to singleton state and bottom panel shows a pair that achieves a division of labour with a retention of a common function (white blocks), the loss of which is prevented because it would compromise the unique functions of each paralog. B) Divergence model confirms asymmetry. Results of 50,000 simulated duplicate pairs, each of which began with 20 functions that were allowed to overlap and 50,000 control pairs that were forbidden to overlap. Left panel: Distribution of functions between each paralog at steady-state. The peak at 0.5 indicates that each function is as likely to sub-functionalize to either paralog, while the breadth of the distribution describes the frequency of asymmetries. Black line indicates results from a model, which allowed functional regions to overlap one another, grey line describes a model in which such overlaps are explicitly forbidden. Right panel: A log-log histogram of ratios for the number of functions for the paralog set depicted in the left panel. The grey bars form a straight line consistent with the corresponding distribution (grey histogram) being a simple binomial, whereas the black bars are skewed towards larger asymmetric ratios. Inset: SCO2-SCO1 represents a representative pair exhibiting partial/asymmetric compensation.
2.3.14 Modeling paralog divergence

Our modeling revealed an apparent contradiction that structural entanglement is positively correlated with retained overlap but is necessary for asymmetric divergence to occur which is itself negatively correlated with functional overlap. In addition, multifunctionality is related to structural entanglement in a straightforward manner as adding more functions to a fixed gene length can only increase the structural/functional entanglement. To reconcile these observations, we propose the following model in which the evolutionary fate of a duplicated gene is governed by an interplay of multi-functionality and structural entanglement (Figure 2.21a). If a pair contains easily partitioned functions, then it will most likely sub-functionalize; on the other hand, a highly structurally/functionally entangled pair would have a tendency to revert to a singleton state as one of the sisters will quickly become non-functional. However, for a medium level of entanglement, the pair will tend to partition some and retain some overlapping functions, which would allow specialization of a common activity. We tested our model by simulating the evolution of paralogs with a range of multi-functionality and structural entanglement parameters and evaluated their fates (Figure 2.21 b-d). We confirmed our predictions in which paralogs with a high score of entanglement more commonly reverted to singletons or retained a functional overlap, whereas those with a low score were more likely to be associated with greater asymmetry. Our modeling suggests that retention of functional redundancy in paralogs reflected in their triple mutant interactions is related to their structural and functional entanglement, which drives their evolutionary trajectories.
Figure 2.21. Testing the axis of divergence. A) Duplicate divergence map showing that multi-functionality and structural entanglement directly affects the evolutionary fates of duplicated genes. B-D) Paralogs were generated to represent the axis of divergence. The propensity to (B) evolve asymmetry, (C) retain overlap and (D) revert to singleton state.
2.4 Discussion

In this chapter, I described the first large-scale triple mutant interaction study involving double mutants of 275 duplicated genes in yeast. I modified the SGA method, which was previously used to interrogate pairwise interactions, to study higher-order interactions. I then developed and applied a conceptually simple quantitative model to measure triple mutant genetic interactions and distinguish them from double mutant effects. This study significantly expands our view of higher-order genetic interactions, which have so far been studied genome-wide for only eight double mutant queries relying on qualitative or semi-quantitative scoring approaches (Tong et al. 2004; Zou et al. 2009; Moir et al. 2012; Haber et al. 2013; Lai et al. 2013).

In order to conduct an extensive genome-scale analysis of paralog pairs, and to generate reproducible data, I developed a diagnostic array for triple mutant screens. Although the ~1200 strains included on the array cover ~1/5 of the genome, they were selected based on several criteria that enabled effective scrutiny of the major biological processes in the cell, in an efficient and cost-effective manner. Use of a miniaturized array allowed screens of double and corresponding single mutant query strains in parallel in two replicates, enhancing the precision of the screens relative to other low-throughput triple mutant interaction studies that either did not assess interactions in the same media or relied on a mixed haploid population of triple and double mutants to estimate double mutant interactions (Moir et al. 2012; Haber et al. 2013). The diagnostic array will also be useful in other array-based interaction mapping studies, such as assessing genetic interactions in various environmental contexts. Previously, arrays composed of a subset of yeast genes have been functionally biased probing specific biological processes (Schuldiner et al. 2005; Collins et al. 2007; Wilmes et al. 2008; Fiedler et al. 2009; Aguilar et al. 2010; Zheng et al. 2010), but unbiased rational array design is important for improving feasibility of high-throughput experiments, which necessitate multiple comparative analyses.

One of my major goals was to use genetic interaction data produced by studying pairwise and complex genetic interactions of 275 double gene deletion mutants of dispensable paralogs to address the long-standing question of why paralogs are retained during evolution, following WGD and SSD events. To generate this dataset, I constructed a collection of double mutant query strains marked with \textit{natMX4} and \textit{URA3} in an SGA-compatible strain background. As such these strains are readily used in SGA-based automated genetic analysis, which enables
construction of arrays not only for assessing colony size but a multitude of other applications, such as high-content screening (Chong et al. 2015). All the strains have been assessed for growth on standard and fourteen alternative growth conditions providing precise fitness estimates, which paves the way for further functional characterization of these mutant query strains by conditional genetic interaction studies and other phenotypic assays. I focused primarily on producing a reference dataset of double and triple mutant interactions involving paralogs grown in standard medium, and an overview of the sensitivity of strains lacking paralogs to growth in the presence of a variety of stressors. I observed that SSD and WGD double mutants significantly differed in their triple mutant interaction and chemical-genetic degrees, providing evidence for greater divergence of SSD relative to WGD, consistent with earlier findings from integration of diverse genomic data (Guan et al. 2007). These results contrast those described by Fares et al. 2013, who argued that SSD are more buffered that WGD because they have more shared paralog-specific interactions. However, this pattern of interactions has been used to define a dosage class of paralogs (Ihmels et al. 2007; VanderSluis et al. 2010) in which shared paralog-specific interactions indicate the inability of sister paralogs to rescue each other’s loss. In fact, triple mutant interactions provide a more accurate estimation of paralogs’ ability for back-up compensation because they uncover instances when synthetic sickness/lethality occurs only in the absence of both paralogs. Showing that SSD and WGD display a similar difference for genetic and chemical-genetic degree lends support for correlation between mutational and environmental robustness indicating that their evolution does not occur independently from one another (Deutscher et al. 2006) a concept that has been somewhat controversial (reviewed in Fares 2015).

In summary, we combined paralog-specific and triple mutant interactions of the pair in a single metric, termed triple mutant interaction fraction, which we validated using other fundamental physiological and evolutionary properties, allowing us to quantify the spectrum of the retained functional redundancy of dispensable paralogs. This approach provided insight into the functional relationships of paralogous genes and shed light on cellular roles of poorly characterized pairs. We also developed a framework to explain how paralog evolution relates to the evolutionary stability of retained common functionalities and asymmetric divergence. Our model and simulations indicate that sub-functionalization will tend to partition ancestral functions asymmetrically in sister paralogs, unless there is a complex sequence-function
relationship, which will result in fixation of functional redundancy. Genetic interaction studies offer a nuanced functional read-out for understanding the factors involved in shaping fates of duplicated genes, which may be key in understanding the evolution of genetic robustness and speciation.
I designed and implemented the strain construction strategy, adapted SGA for studying triple mutant genetic interactions, conducted SGA screens and analyzed the resulting interactions. Guihong Tan constructed the haploid selection plasmid, which was used during strain construction. Yiqun Chen, Ermira Shuteriqi, Bryan-Joseph San Luis and Matej Usaj provided technical assistance during strain construction and screening processes. In close collaboration with Benjamin VanderSluis, I was involved in selecting the gene set for screens. Benjamin VanderSluis wrote and implemented the scoring code for triple mutant genetic interactions and evaluated the quality of the data using precision-recall analysis. I collaborated with Elizabeth Koch to generate single and double mutant fitness standard.
3 Surveying triple mutant interaction space in yeast

3.1 Introduction

Genetic interactions have long been recognized as playing a fundamental role in shaping phenotypes of living organisms (Hartman et al. 2001). A ‘genetic interaction’ denotes any phenotype that is unexpected given the known combined effects of individual mutations (Phillips 1998). The identification of genetic interactions relies on determining the deviation of the observed from expected mutant phenotype using a multiplicative model, which formulates the expectation using the product of the effects of individual mutant loci (Mani et al. 2008). As discussed in Chapter 1, negative genetic interactions, then, refer to a combination of genetic perturbations that result in a more severe phenotype than expected (Costanzo 2012). Synthetic lethality is an extreme negative interaction where two mutations, each causing little growth defect on their own, result in a lethal phenotype when combined as double mutants. Synthetic lethal interactions often occur between genes that impinge on a common essential biological function. Thus, genome-scale screens for negative genetic interactions are an effective means to chart the genetic network that underlies the functional wiring diagram for a cell or organism (Costanzo et al. 2010).

The global double mutant genetic interaction network is a valuable resource for understanding gene function and revealing key features of the global genetic interaction network (Costanzo et al. 2010; Baryshnikova et al. 2013). Genome-wide genetic interaction screens that interrogated 1,712 query genes for a total of ~5.4 million gene pairs identified ~170,000 interactions of which ~2/3 represented negative genetic interactions (Costanzo et al. 2010). This indicates that the frequency of synthetic sick/lethal interactions in the cell is approximately 2% involving diverse combinations of genes that work together to buffer mutational effects on the fitness of the organism. However, many phenotypes in simple eukaryotes and metazoans are likely influenced by more complex genetic interactions involving larger sets of genes and genetic variants (Dowell et al. 2010; Zuk et al. 2012). Indeed, conditional lethality in closely related and interbreeding yeast strains was shown to be due to multiple, often three or four, modifier loci that generate a synthetic lethal genetic interaction with the gene of interest (Dowell et al. 2010). Thus, deciphering the genotype to phenotype relationship for an individual requires that we expand our focus beyond pair-wise genetic interactions to include complex, higher-order genetic interactions.
involving more than two genes.

In this chapter, I describe a comprehensive and quantitative analysis of genetic interactions for over 150 double mutant queries chosen to enable exploration of ‘triple mutant space’. I used the modified SGA technique described in Chapter 2 to generate a compendium of triple mutant genetic interactions and then compared the features of double and triple mutant interaction networks. I show that sharing neighbours on the digenic interaction network and gene essentiality determine whether a double mutant query had many triple mutant interactions. I also extrapolated our findings to the entire yeast genome and showed that triple mutant interaction space is extensive, suggesting that analysis of complex genetic interactions will be required for understanding genetic network wiring and robustness.

3.2 Materials and Methods
3.2.1 Constructing a singleton double mutant collection with control query strains

A collection of 148 viable double mutants was constructed. The gene pairs were grouped into three general categories based on a range of features: 1) high (n = 52), intermediate (n = 56) and low (n = 40) average digenic interaction degree; 2) high (n = 45), intermediate (n = 59) and low (n = 44) functional similarity, as measured by their digenic interaction profile similarity and co-annotation to the same Gene Ontology term(s); and 3) gene pairs directly connected by zero – very weak (n = 53), weak (n = 43) or moderate (n = 52) negative interactions (refer to Table 3.1).

I predicted that if two query genes are highly functionally similar, genetically interact with each other (and thus exhibit a double mutant fitness defect) and display many genetic interactions specific to their single mutants, then the double mutant should show many trigenic interactions. In contrast, query gene pairs that do not interact directly, are functionally dissimilar and exhibit few genetic interactions specific to individual mutants, would be predicted to show few trigenic interactions. The data that are reported in Chapter 2 form the basis of these predictions. I observed that highly functionally similar paralogs tend to exhibit many triple mutant interactions, since it is not possible to gage functional similarity of a pair of singleton genes using defined evolutionary features intended for duplicated genes, the similarity of their double mutant
interaction profile should serve an analogous role. In addition, informative profile similarity necessitates at least a moderate number of interactions, thus there should be a relation with double mutant interaction degree, which itself probably reflects the activity of a gene on the assessed condition. Thus, findings from the previous chapter combined with knowledge of the double mutant interaction network allow for formulation of formal expectations related to the behaviour of double mutants on the triple mutant interaction network.

To test the frequency of trigenic interactions associated with every combination of selected features, I selected an average of 6 gene pairs from each of 27 \((3^3)\) possible feature combinations (Table 3.1). In total, I selected 148 gene pairs comprising 46 essential genes and 250 non-essential genes, spanning all major biological processes in the cell. The combinations include 109 pairs of non-essential genes, 6 pairs of essential genes and 33 pairs with mixed essentiality. I also chose 4 gene pairs that display a strong interaction \((<-3)\) in \(S.pombe\) (Ryan et al. 2012a), and no interaction in \(S.cerevisiae\) (Costanzo et al. 2010), one-to-one orthologs (Koch et al. 2012), functionally co-annotated in \(S.cerevisiae\) and exhibit a double mutant fitness of 0.7-0.95. In this manner I was able to explore the genetic rewiring that has occurred between two distantly related yeast species that are evolutionary as far apart from each other as worm and human.
The main query construction method is described below and depicted in Figure 3.1. A subset of an existing haploid lab strain collection (SN) deleted for gene 1 (MATa gene1Δ::natMX4 can1Δ::STE2pr-Sp_his5 lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 LYS2) was transformed with pGT423, which was used for downstream haploid selection steps described below. High-throughput plasmid transformations were carried out as follows: 1x master mix was prepared by mixing 50-100 ng plasmid, 20 µg ssDNA (Sigma D9156), 100 µl transformation buffer (8:1:1:1 of 50% PEG4000 : 1 M LiAC : 10 x TE : DMSO, respectively). 100 µl of master mix were aliquoted in each well of a 96-well PCR plate and a yellow tip full of freshly grown cells (passaged on the previous day) was added to each well, mixed by swirling and pipetting up and down. The cell suspension was incubated at 30°C for 1 hr if non-ts, otherwise at RT, then heatshocked at 42°C for 20 min and cooled on ice for 5 min, centrifuged 1,000 x g for 1 min, resuspended in 200 µl YEPD, transferred to a shallow 96-well plate and incubated at 30°C for 1.5-2 hrs for non-ts, otherwise at RT. SD-Leu agar plates were dried in a biohood for 15 min and 10 µl of mixed cell suspension were then spotted on the plates, which were incubated for 2-3 d at 30°C for non-ts, otherwise at RT.

In addition, a subset of strains from the EUROSCARF haploid deletion collection (DMA) carrying a deletion of gene 2 (MATa gene2Δ::kanMX4 his3Δ1 leu2Δ0 ura3Δ0 met15Δ0) was switched from kanMX4 to KIURA3 using PCR-mediated gene deletion as described earlier (section 2.1) except the KIURA3 PCR product was flanked by 55 bp of sequence homology to the 5’ end of the TEF promoter and the 3’ end of the TEF terminator of the kanMX4 cassette.
Primers for the amplification of \textit{KlURA3} with sequence of homology to \textit{MX4} cassette (5'-3'):

Forward PTEF55-\textit{klura3}_f:

\texttt{ACATGGAGGCCAGAATACCCCTTCCTTGACAGTGACGTGCGCAGCTCAGGGGC}cg
gagacaatcatatgggag

Reverse TTEF55-\textit{klura3}_r:

\texttt{CAGTATAGCGACCAGCATTCAATACGATGACGTGACGTGCGCATGATATTACTTTCTGC}GCA tctg
gaggaagtttgaggg

The resulting transformed SN and DMA strains were arrayed in 96-format on SD-Leu and SD-Ura agar OmniTrays, respectively. Then, the two arrays were mated by pinning on top of each other on YEPD and incubated for 1 d at RT. The diploids were selected by pinning on SD\textsubscript{msg}-Ura/Leu+NAT, 2-3 d, RT. Once grown up they were manually patched on sporulation plates, 7-14 d, 22°C. Then, \textit{MATα} progeny was selectively germinated by patching the swirled sporulation mix on YEPD+Hygromycin (because the hygromycin resistance gene is under the control of \textit{MATα} specific promoter, \textit{STE3}), 2d, RT. To facilitate plasmid loss the germinated haploid mix was swirled and patched onto YEPD, 1d, RT. Then, the final query strain was obtained by streaking out the germinated patch of cells for single colonies on S\textsubscript{Gal}\textsubscript{msg}-Arg,Lys,Ura+CAN,LYP,NAT, 3 d, RT. Galactose was substituted for glucose as carbon source in order to induce the toxic overexpression of \textit{KAR1} that counterselects for strains that still carry the plasmid. Colony PCR confirmation and strain preservation were performed as described in Section 2.1. Single mutant control query strains were constructed by either mating Y14391 (\textit{MATα hoΔ::natMX4 can1Δ::STE2pr-Sp_his5 lyp1Δ his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 LYS2}), which was transformed with pGT423, with transformed DMA strains or mating transformed SN strains with transformed DMA809 (\textit{MATα hoΔ::kanMX4 his3Δ1 leu2Δ0 ura3Δ0 met15Δ0}).
To facilitate the construction of double mutant query strains, haploids of opposite mating types carrying individual gene mutations were crossed (Methods). \(\text{MAT}^a\) haploid strains were obtained from the yeast deletion collection (Giaever et al. 2002), in which a mutation in one of the members of the gene pair was marked with the geneticin resistance gene (\(\text{kan}^R\)), and switched to an auxotrophic \(\text{URA}^3\) marker. \(\text{MAT}^a\) haploid strains were obtained from a second yeast deletion collection (Costanzo et al. 2010), in which a mutation in another member of the gene pair was marked with the nourseothricin resistance gene (\(\text{nat}^R\)), and transformed with a plasmid (\(\text{pGT423}\)) necessary for the selection of haploid double mutants. \(\text{pGT423}\) contains \(\text{STE}3\text{pr-hph}^R\) marker which enabled selective germination of \(\text{MAT}^a\) meiotic progeny following the sporulation of heterozygous diploids and \(\text{GAL1pr-KARI}\) enabling the counterselection for plasmid loss in the final strain.
3.2.2 Estimation of single and double mutant query fitness on standard growth conditions

Double and single mutant fitness were obtained as described in section 2.2.4 and the distributions are depicted in Figure 3.2 (top panels). Most query strains were characterized by fitness value of 0.75-1. There agreement between the fitness standards is similar to what is reported in other studies (Baryshnikova et al. 2010b) and is depicted in Figure 3.2 (bottom panels) with Pearson correlation coefficient = 0.67, $p = 0$; and 0.62, $p = 0$, for single and double mutant fitness measures, respectively.

Figure 3.2. Quantification of fitness of query strains using high-density arrays. Top panels: (left) histogram of double mutant fitness, n=149; (right) histogram of single mutant fitness, n=304. Bottom panels: (left) scatter plot comparing double mutant fitness derived from the Triple Mutant Project (this thesis project) and Lab SGA project, n=301; (right) scatter plot comparing single mutant fitness derived from the Triple Mutant Project (this thesis project) and Lab SGA project, n=141.
3.2.3 Triple mutant SGA

Triple mutant SGA screens were conducted and scored as described in Section 2.3.2 and 2.3.3. The Diagnostic array described in Section 2.3.5 was used for screens.

![Figure 3.3. Distribution of double and triple mutant interaction score magnitudes. A) Histogram of significant double mutant interaction score magnitudes. B) Histogram of significant triple mutant interaction score magnitudes.](image)

3.2.4 Precision-recall analysis

Precision-recall analysis was conducted as previously described (Baryshnikova et al. 2010b). The functional benchmarks are the following: merged protein-protein interaction standard (Gavin et al. 2006; Krogan et al. 2006; Tarassov et al. 2008; Yu et al. 2008; Babu et al. 2012), GO co-annotation is based on GO process (Myers et al. 2006), co-expression (Huttenhower et al. 2006), co-localization (Huh et al. 2003)
3.3 Results

3.3.1 Gene pair selection

I systematically assessed triple mutant genetic interactions for a subset of gene-pairs that I selected based on several criteria that I reasoned would capture the extent of higher-order genetic interaction space. The selection criteria were based on the properties characteristic of the digenic interaction network. Double mutant strains exhibiting a range of genetic interaction strength, genetic interaction profile similarity and average genetic interaction degree were used to construct all possible combinations of triple mutants for quantitative genetic interaction analysis (Figure 3.4 and Table 3.1). Co-annotation to broadly defined functional gene sets was included as an additional metric to gauge functional similarity of double mutants with a high genetic interaction profile similarity. Gene pairs with varying combinations of essentiality status were represented along with coverage of all major biological processes in the cell. Thus, gene pairs were selected to fill bins of varying attributes to enable sampling that would provide a representative view of the genetic interaction space. In total, I constructed 148 double mutant strains using a collection of 296 single mutant strains, including strains carrying 46 temperature-sensitive alleles of different essential genes and 250 deletion alleles of non-essential genes. On average 6 gene pairs were used to populate each of 27 ($3^3$) bins to test the frequency of trigenic interactions associated with every possible combination of selected features (Figure 3.4, Table 3.1).
Figure 3.4. Criteria for selecting query strains for sampling triple mutant interaction space of singleton genes in yeast. Gene pairs were grouped into three general categories based on a range of features as follows: 1) high (n=52), intermediate (n=56) and low (n=40) average digenic interaction degree (denoted by the number of black edges of each node); 2) high (n=45), intermediate (n=59) and low (n=44) functional similarity, as measured by their digenic interaction profile similarity and co-annotation to the same Gene Ontology term(s) (denoted in different colours); and 3) gene pairs directly connected by zero-very weak (n=53), weak (n=43) or moderate (n=52) negative interactions (represented by the increasing width of the red edges). Two query genes that are highly functionally similar, genetically interact with each other (and thus exhibit a double mutant fitness defect) and display many genetic interactions specific to their single mutants are predicted to show many trigenic interactions as a double mutant. In contrast, query gene pairs that do not interact directly, are functionally dissimilar and exhibit few genetic interactions specific to individual mutants, would be predicted to show few trigenic interactions. The top right corner shows the possible extent of triple mutant space compared to sampled space (depiction is not to scale).
3.3.2 Functional relation of double and triple mutant interactions

To generate triple mutant profiles, I screened the 148 double mutant queries and the corresponding 296 single mutant queries against the diagnostic array of 1,232 mutants (see Chapter 2). In total I tested 193,200 triple and 386,400 double mutants for fitness defects to discover 2,820 triple and 7,954 double mutant interactions (Table 3.2). To assess the functional information obtained from double versus triple mutant networks, I compared the distributions of double and triple mutant interactions across different cellular processes (Figure 3.5a, b). I identified combinations of biological processes that were enriched or depleted for genetic interactions relative to the expected frequency of the random background of the double (Figure 3.5a, left panel) or triple (Figure 3.5a, right panel) mutant interaction gene set. As expected, genes belonging to similar biological processes were enriched for double mutant interactions; although a similar pattern of enrichment was seen for triple mutant interactions, the fold increase over background was lower than for double mutant interactions suggesting that a three-parallel pathway model may be more rare than a two-parallel pathway model that likely underlies many double mutant interactions [see Chapter 1, (Costanzo et al. 2011)]. We also observed differences in double and triple mutation interactions that bridge biological processes. Specifically, consistent with previous observations (Costanzo et al. 2010), genes involved in chromatin, transcription, membrane trafficking and secretion interacted with genes involved in diverse biological processes, representing pleiotropic hubs on the double mutant interaction network. However, the frequency of triple mutant interactions involving genes with roles in chromatin and transcription was reduced (3.42 vs. 1.75 avg. fold increase over background), suggesting that we may be close to saturating information that can be obtained about chromatin and transcription by surveying genetic interactions in standard growth conditions. By contrast, genes with roles in membrane trafficking and secretion were more overrepresented in the triple mutant interaction dataset (3.16 vs. 5 avg. fold increase over background), suggesting that intracellular trafficking is a highly buffered process (Fig. 3.5b). Thus, double and triple mutant interactions capture similar functional relationships to double mutant interactions but exhibit distinct pleiotropic functions.

I also evaluated the functional utility of resultant double and triple mutant interactions against common functional benchmarks: annotation to the same Gene Ontology (GO) biological process, subcellular-localization pattern, protein-protein interactions and gene co-expression (Fig. 3.5c). Although genetic interactions are not always expected to connect functionally related genes, they
tend to be enriched among genes in the same biological process, so these standards serve as objective metrics for assessing the functional utility of this dataset. Consistent with our expectation double mutant interactions overlapped with co-annotated, protein-protein interacting, co-expressed and co-localized gene pairs. Triple mutant interactions were similarly enriched across these standards although the enrichment was ~40% less than that seen for double mutant interactions. The slightly reduced overlap with co-annotated pairs is consistent with the observation that within process overrepresentation is significant for triple mutant interactions but is reduced in comparison to double mutant interactions (Fig. 3.5a, b). The reduction with respect to overlap with functional benchmarks could be due to a lower magnitude of triple mutant interactions relative to double mutant interactions (Figure 3.3). Thus, like double mutant interactions, triple mutant interactions are functionally informative although there are some distinctions.

Since double and triple mutant interactions are functionally informative and display some distinctions, combining them offers a more in depth and nuanced characterization of the cellular machinery. For example, the MTC1-MDY2 double mutant is the largest hub on the triple mutant interaction network, exhibiting 155 interactions. MTC1 encodes a protein of unknown function, which localizes to early Golgi. It shares a highly similar genetic interaction profile with USO1, which is involved in vesicle-mediated Endoplasmic Reticulum (ER) to Golgi transport (Noda et al. 2007), and with RUD3, which encodes a Golgi matrix protein important for the structural organization of the cis-Golgi (Gillingham et al. 2004), thus lending support for a putative role of MTC1 in the Golgi. Mdy2p has been implicated through its protein-protein interaction with Get4p in the GET pathway, which is important for inserting tail-anchored proteins into ER membranes (Jonikas et al. 2009). The individual double mutant interactions for both MTC1 and MDY2 tended to be related to ER to Golgi trafficking due interactions with members of HMC – YDJ1, which is involved in translocation of polypeptides into the ER (Caplan et al. 1992), TRAPP – TRS20, TRS23, TRS85 and BET3, which are related multimeric guanine nucleotide-exchange factors for the GTPase Ypt1, regulating ER-Golgi traffic (TRAPPI), intra-Golgi traffic (TRAPPII), endosome-Golgi traffic (TRAPPIII) (Sacher et al. 1998), COG – COG2, COG3, COG5, COG6, COG7 and COG8, which are components of the conserved oligomeric Golgi complex that is involved in protein trafficking by mediating fusion of transport vesicles to the Golgi (Loh and Hong 2004) (Fig. 3.5d). Interactions were also seen with GET – GET1 and
GET3, which are involved in insertion of proteins into the ER membrane (Schuldiner et al. 2008), and Ric1/Rgp1/Ypt6 protein complexes – RIC1 and YPT6 which are small Ras-related GTPases involved in retrograde transport to the cis-Golgi network (Suda et al. 2013) (Fig. 3.5d). Triple mutant interactions of MTC1-MDY2 included members of the EMC complex, which is required for efficient folding of proteins in the ER (Jonikas et al. 2009), and genes encoding Sec62/Sec63, a protein complex involved in targeting and import into the ER (Young et al. 2001). These interactions suggest an inability to tolerate multiple perturbations to ER function combined with a loss-of-function mutation in a Golgi-related gene (Fig. 3.5d). A local triple hit to the system, involving the ER, Golgi and peroxisome also resulted in reduced fitness suggesting of the importance of the local architecture of the cell (Fig. 3.5d). Finally, triple mutant interactions with members of the GSE, Lst4-Lst7 and Pan1/Sla1/End3 complexes, which are involved in vesicular trafficking to the plasma membrane or from the membrane to the endosome, suggested that reduced flow through the cellular trafficking system may be a causative factor in the collapse of the cellular network (Fig. 3.5d). Triple mutant interactions with members of the elongator complex and urmylation pathway (Figure 3.6) also supported a previous observation that thiolation of the wobble base of cytoplasmic tRNA is required for translation of secretion gene mRNAs possibly due to codon bias observed in genes involved in this process (Costanzo et al. 2010; Bauer et al. 2012). Thus, integrating information derived from different types of genetic networks enhances our understanding of the biology of the cell.
Figure 3.5. Functional characterization of triple mutant interactions. A) Frequency of negative genetic interactions within and across biological processes. The fraction of screened query-array pairs exhibiting negative interactions was measured for 8 broadly defined functional gene sets. A colour was assigned to each process-process element reflecting the fold-increase over the background fraction of interactions (digenic = 0.023, trigenic = 0.016). Process-process received a count if either one of the query genes and the array were annotated to the corresponding processes. B) Comparison of heatmaps in A, whereby the trigenic interaction heatmap was divided by the digenic interaction heatmap. C) Enrichment of negative digenic (light blue) and trigenic (blue) interactions across four functional standards. PPI denotes protein-protein interactions. Red line marks no enrichment. D) Summary of digenic and trigenic interactions involving MTC1-MDY2. Protein complex annotation was obtained from Baryshnikova 2010. Only protein complexes in which > 50% of members display genetic interactions are shown. Negative genetic interactions, ε<0.08, pval<0.05, are depicted.
Figure 3.6. Trigenic interactions reveal distinct functional information for *MTCI-MDY2*. *MTCI-MDY2* trigenic interactions are enriched for tRNA processing (GO:0008033, 8.260x10^{-6}). For clarity only interactions resulting in GO enrichment are depicted. Negative genetic interactions at intermediate cut-off are depicted.
3.3.3 Triple mutant interaction space is extensive

Since we selected the gene pairs based on the properties of the double mutant interaction network, we next assessed how they determined the connectivity on the triple mutant interaction network. Consistent with our predictions, all three double mutant interaction properties which we used in our query gene selection process significantly correlated with triple mutant interaction degree, with the strongest predictor being the profile similarity of the genetic interactions of individual mutants \( r = 0.37, p = 5 \times 10^{-6} \) (Fig. 3.7a, black bars). Thus, sharing neighbours on the double mutant interaction network increased the likelihood of the gene pair exhibiting many triple mutant interactions. Controlling for the influence of the remaining two variables did not eliminate the significant correlation (Fig. 3.7a, grey bars). Thus, global genetic interaction maps based upon double mutant interactions are highly predictive of triple mutant interactions.

The number of triple mutant interactions for the double mutant also correlated with the average number of interactions observed for the individual mutants \( r = 0.33, p = 9 \times 10^{-5} \), suggesting that hubs on two different networks are similar and contribute to the higher-order robustness of the genome (Fig. 3.7a). This is consistent with the observation that the essentiality of a gene reflects its connectivity on different types of genetic interaction networks. Temperature sensitive alleles of essential genes on average show a higher number of interactions than deletion alleles of non-essential genes for single mutant queries \( p = 0.0364 \) (Fig. 3.7b). Similarly, a double mutant that carries at least one temperature-sensitive allele of an essential gene \( p = 0.0001 \) exhibits a greater genetic interaction degree than a mutant deleted for a pair of non-essential genes. Thus, essential genes are important buffers of phenotypic variation on both double and triple mutant interaction networks.

Examining the pairwise interplay of double mutant interaction network properties revealed that query gene pairs displaying the highest average triple mutant interaction degree showed a moderate average double mutant interaction degree and a significant but weak negative interaction (Fig. 3.7c). Controlling for the third variable did not affect the distribution of the average triple mutant degree across bins (Figure 3.8). Highly correlated genetic interaction profiles tended to connect functionally similar genes and since the gene pair is connected by a weak interaction, they are likely to represent collaborative elements rather than members of the same pathway.
Since double mutant interaction properties are predictive of the triple mutant interaction degree, we attempted to extrapolate our findings to the rest of the genome. Our screens showed that the frequency of triple mutant interactions was \(~1/3\) lower than that of double mutant interactions (Table 3.2). However, given that there are \(~2,000\) fold more possible triple mutant combinations globally, the triple mutant interaction space is extensive, which has broad implications for our understanding of genetic network robustness and underlying mechanisms of genetic disease states. We also found that the occurrence of interactions that are due to an exacerbated effect of double mutant interactions was \(~1.8\) lower than interactions that had not been previously seen in the double mutants, which could be due to the evolutionary pressures that negatively select against double mutant interactions.

**Table 3.2. The extent of genetic interaction space**

<table>
<thead>
<tr>
<th></th>
<th>Double mutant interaction space</th>
<th>Triple mutant interaction space</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. queries strains</td>
<td>296</td>
<td>148</td>
</tr>
<tr>
<td>No. array strains</td>
<td>1200</td>
<td>1200</td>
</tr>
<tr>
<td>No. tested combinations</td>
<td>386,400</td>
<td>193,200</td>
</tr>
<tr>
<td>No. negative genetic interactions</td>
<td>7,954</td>
<td>2,820</td>
</tr>
<tr>
<td>Freq. negative genetic interactions</td>
<td>2.1%</td>
<td>1.46%</td>
</tr>
<tr>
<td>Total possible combinations globally</td>
<td>18 million</td>
<td>36 billion</td>
</tr>
</tbody>
</table>
Figure 3.7. Relation of double and triple mutant interaction networks.

A) Correlation of trigenic interaction degree with individual attributes of the digenic interaction network. Trigenic degree denotes the number of negative genetic interactions specific to a double mutant. Negative genetic interaction strength reflects the magnitude of the fitness defect of the query double mutant. Avg. no. digenic interactions represents the average number of negative genetic interactions of single mutants of the query gene pair. Digenic interaction profile correlation represents the similarity of interacting genes directly connected to each single mutant of the query gene pair. Spearman correlation coefficient is shown in black and with other variable partialled out in grey.


C) Pairwise interplay of digenic interaction network attributes in relation to the average trigenic interaction degree, which is colour coded according to the colour map on the bottom. Mean no. queries per bin = 16.4. Negative genetic interactions, score<-0.08, pval<0.05, are depicted.
Figure 3.8. Pairwise interplay of digenic interaction network attributes in relation to the average trigenic interaction degree (residual) with partialling out the third variable. Heatmaps are colour coded according to the colour map on the bottom. Mean no. queries per composite bin = 16.4.
3.3.4 Genetic interaction network rewiring between distantly related species

We also used our triple mutant interaction dataset to ask whether the relatively low level of conservation of genetic interactions between evolutionary distant organisms might be explained, at least in part, by genetic rewiring involving higher-order genetic interactions. A literature-curated *S. pombe* network revealed that 23% of SS/SL digenic interactions are conserved in the existing *S. cerevisiae* network (Dixon et al. 2008). Since literature-curated datasets tend to be biased, the same study developed a method for *S. pombe* SGA to experimentally interrogate ~220 query genes involved in DNA replication and repair, chromatin remodeling, intracellular trafficking and other core cellular processes to map double mutant interactions. The study revealed that 29% of the tested genetic interactions were conserved between the two distant yeast species. Similarly, another study mapped genetic interactions in the fission yeast involving double mutant combinations among 550 genes related to chromosome biology and showed that up to 33% of negative genetic interactions were conserved when compared to the budding yeast (Roguev et al. 2008). Thus, despite ~75% homology between *S. pombe* and *S. cerevisiae* gene content, the level of conservation of direct digenic interactions is somewhat lower. To help resolve this difference and understand the genetic wiring of diverse eukaryotic species we performed triple mutant interactions screens involving *S. cerevisiae* homologs of *S. pombe* SL gene pairs. We performed triple mutant SGA analysis using query strains carrying deletion mutations in *SKI3-SOL2, RAD57-RAD17, EAF3-NAP1* and *ASE1-MAD2*. These pairs were selected because their double mutants are synthetically lethal in *S. pombe* but show no interaction in *S. cerevisiae*; they are co-annotated and represent one-to-one orthologs. *SKI3-SOL2* and *RAD57-RAD17* showed two and three triple mutant interactions, respectively, whereas *EAF3-NAP1* and *ASE1-MAD2* interacted with an average of 23 genes (Figure 3.9a). Overall interacting genes tended to be functionally related to the query gene pairs with some interactions belonging to diverse processes (Figure 3.9a). For example, a substantial portion of triple mutant interactions involving *ASE1-MAD2* had a role in mitosis and chromosome segregation. The distribution of the scores that were associated with the triple mutant interactions showed that many were of relatively strong magnitude falling below a stringent cut-off, $\tau < -0.12$, likely representing real reproducible effects (Figure 3.9b). Furthermore, triple mutant interactions of *ASE1-MAD2* provided clear evidence of genetic rewiring that has occurred between the two yeast species. Ase1p localizes to the mitotic spindle midzone and is required for spindle
elongation and stabilization, whereas Mad2p is a member of the spindle-assembly checkpoint complex (Yamagishi et al. 2014). Their triple mutant interactions included kinetochore proteins such as Ybp2p, Scc4p and Iml3p. Interestingly, despite homology of Iml3p in *S. cerevisiae* to Fta1p in *S. pombe*, the localization of the Ndc80 complex to kinetochores, which in turn connects to microtubules improving chromosome transmission fidelity (Hinshaw and Harrison 2013), depends on Iml3p but not on Fta1p. It has been suggested that Ndc80 complex assembles via a different pathway in fission yeast (Tanaka et al. 2009) supporting the finding that *fta1*+ was unable to rescue synthetic lethality due to the double perturbation of *ase1*+ and *mad2*+. In all pairs tested we found instances of genetic interaction conservation once we take into account higher-order interactions, suggesting that higher-order interactions offer insight into the genetic rewiring that has occurred between two distantly related yeast species and the level of conservation of their genetic networks is most likely greater than previously appreciated.

**Figure 3.9.** Triple mutant interactions of *S. cerevisiae* homologs of *S. pombe* SL pairs as query genes. A) Double mutant query strains that were used for triple mutant interaction screens are denoted by two genes separated by a hyphen; the remaining nodes represent interacting genes. In house terms (Constanzo et al. 2010) were used for gene annotation to major biological processes in the cell. B) Distribution of scores corresponding to all the triple mutant interactions is illustrated in A.
3.4 Discussion

In this chapter, I described a large-scale triple mutant interaction study involving over 150 double mutants of random singleton genes in yeast. Comparative analyses showed that triple mutant interactions are comparably functionally informative to double mutant interactions. However, within-process interactions were less frequent for triple mutants suggesting that three-parallel pathway model is more rare in the cell than the occurrence of two parallel pathways. Differences observed in double and triple mutation interactions that bridge biological processes, whereby chromatin and transcription is overrepresented on the double mutant interaction network and membrane trafficking and secretion on the triple mutant interaction network indicates genetic wiring architecture underlying the different types of genetic interaction networks and suggests the latter process is very important for buffering phenotypic variation. I also showed that a combination of double and triple mutant interactions enhances our understanding of the biology of the cell. Specifically, using the largest hub on the triple mutant interaction network, MTC1-MDY2, as an example, I showed that cell viability may depend on an optimal flow through the membrane trafficking system since the triple mutant interactions involved genes downstream of ER and Golgi with a role in vesicular trafficking to the plasma membrane or endocytosis.

By probing the properties of the double mutant interaction network I discovered that sharing neighbours is the strongest predictor of triple mutant interaction hubiness. Double mutant interaction degree and gene essentiality were also determining factors for higher-order connectivity, suggesting that essential genes represent important hubs on both types of genetic networks. Our systematic analysis revealed that despite a lower frequency of triple mutant interactions compared to double mutant interactions, they are 2,000 times more numerous, suggesting that triple mutant interaction space is extensive. This finding suggests that mapping digenic interactions alone will not saturate our understanding of genetic network robustness. Finally, I found evidence of genetic network rewiring between S.cerevisiae and S.pombe when I identified instances of conservation of synthetic lethal interactions in the fission yeast when I considered a third redundant gene in the budding yeast. This result suggests that genetic interaction network conservation may be greater than originally thought (Dixon et al. 2008; Roguev et al. 2008). It may become more feasible to map genetic interactions between distantly
related species (such as worm and human, which are evolutionary as far apart as budding and fission yeast) when considering higher order genetic interactions.

Systematic assessment of complex genetic interactions is an important step towards moving beyond pairwise genetic interactions in the direction that will bring us closer to explaining the missing heritability in humans. Given that human disease states are likely caused by interplay of a combination of multiple genetic variants (Zuk et al. 2012), mapping higher-order genetic interactions in simple model organisms is an important stepping stone. Continued triple mutant interaction experiments will fill in our gap in knowledge about this vast interaction space.
Chapter 4
4 Thesis Summary and Future Directions

4.1 Thesis Summary

In this thesis, I studied complex genetic interactions involving three genes using the Synthetic Genetic Array (SGA) analysis technique in the budding yeast, *Saccharomyces cerevisiae*. My aim was to enhance our understanding of the genetic architecture of the cell and the effects of combinatorial perturbations on cell fitness.

In Chapter 2, I described triple mutant interactions of duplicated genes to gain insight into their retention during evolution. I optimized SGA, a method originally developed to study pairwise genetic interactions, for triple mutant interaction screens and was involved in the development of a scoring strategy for quantifying triple mutant fitness effects. I also constructed a library of 275 double gene deletion mutants consisting of 242 WGD and 33 SSD paralogs, providing a valuable resource for the scientific community. By quantifying fitness of this collection of mutant strains on standard nutrient-rich and environmental stress conditions, I found that WGD and SSD paralogs contribute differently to robustness against genomic and environmental perturbations. These observations provide supporting evidence for greater functional divergence of SSD compared to WGD gene pairs and show a correlation of the evolution of two types of robustness. Most importantly, I studied the functional relationship and buffering capacity of these paralogs by interrogating them using triple mutant interaction screens. To do this I developed a quantitative measure of the level of paralog functional divergence and redundancy by combining their paralog specific and triple mutant interactions into a metric termed triple mutant interaction fraction. Using this approach, I uncovered a spectrum of retained functional redundancy among dispensable paralogs and shed light on the function of poorly characterized pairs. In collaboration with the Myers Lab, I also developed a framework to explain how paralog evolution relates to the retention of functional overlap and asymmetric divergence and proposed that structural and functional entanglement constrain paralog evolution necessitating maintenance of redundancy to achieve divergence and specialization. In this manner, I expanded our knowledge of paralog evolution through a nuanced functional assessment involving genetic interactions.

In Chapter 3, I exploited complex genetic interactions to survey the trigenic landscape of a cell on a more global level. I utilized a novel high-throughput strategy to construct 150 double
mutant query strains along with matching single mutant control strains involving unambiguous singletons spanning major biological process of the cell, including loss-of-function mutants of essential and non-essential genes. This is a valuable mutant construction strategy because it enables large-scale assessment of complex interactions by providing resources to generate multi-gene mutant query collections. The plasmid-based haploid selection relies on STE3pr-hphR marker, which selectively germinates the desired MATa mating type and GAL1pr-KAR1, which is used for counter-selection of plasmid loss in the final strain. Having constructed the library I subjected it to triple mutant genetic interaction analysis and showed that triple mutant interactions are similarly functionally informative as double mutant interactions. I provided evidence to suggest that a three-parallel pathway model is more rare in the cell than a two-parallel pathway model by observing that triple mutant interactions occur less frequently involving genes annotated to the same process than double mutant interactions. I also showed that genes belonging to chromatin and transcription on the double mutant interaction network and vesicular trafficking on the triple mutant interaction network are important buffers of phenotypic variation. By examining the properties of the double mutant interaction network I discovered that sharing neighbours is the strongest predictor of triple mutant interaction hubbiness, although double mutant interaction degree and gene essentiality are also important determining factors. Perhaps most importantly, I showed that the triple mutant interaction network is extensive, with a frequency of 1.46% in the surveyed set compared to a 2.1% frequency of double mutant interactions. Thus, even though higher-order genetic interactions are slightly more rare in the cell, the number of possible combinations is numerous and exceeds double mutant combinations by 2,000-fold suggesting that the extent of triple mutant interaction space is vast and needs to be explored to fully understand the genetic wiring of the cell. Finally, I uncovered triple mutant interactions in S.cerevisiae involving homologs of S.pombe genes involving synthetic lethal pairs. Such evidence of genetic network rewiring between distantly related yeast species may be key for understanding genetic interaction conservation between other organisms that are as evolutionary far apart from each other as worm and human.

The findings reported in this thesis are an important advancement in the field of genetics and functional genomics because they describe the first systematic and quantitative assessment of higher-order genetic interactions. Employing this approach has offered an opportunity to enhance our understanding of factors that shape the evolutionary trajectories of duplicated genes and
explore the extent of triple mutant interaction space, which is an important milestone for addressing the genotype to phenotype challenge and may be key for explaining missing heritability of human traits and disease states.

4.2 Future Directions

Below, I outline future experiments that can enhance or make use of the reagents or data that are described in this thesis.

4.2.1 Enhancing the understanding of functional relationship of duplicated genes through conditional genetic interaction screens

Systematic genetic interaction studies have enhanced our understanding of genetic network architecture and provided insight into gene function (Tong et al. 2004; Costanzo et al. 2010). For instance, genes with similar biological roles tend to share many genetic interactions. Therefore, clustering genes according to the similarity of their genetic interaction profiles is a powerful tool for predicting the functions of unknown genes (Tong et al. 2004; Costanzo et al. 2010).

However, double mutant interaction studies showed that approximately 25% of genetic interaction profiles are sparse producing signal, which cannot be differentiated from background noise, thereby hampering efforts to position them in meaningful functional clusters to illuminate their biological roles. The fraction of genetic interaction network that cannot be probed under standard growth conditions has been termed the ‘dark net’ (Chad Myers, personal communication).

Duplicated genes are significantly overrepresented among the ‘dark net’ (p < 0.05). In fact, for 122 of the 275 dispensable paralog pairs examined in Chapter 2, one or both sister paralogs belonged to this set. Screening them for triple mutant interactions uncovered approximately 15% of paralog pairs characterized by a triple mutant interaction fraction equal to or greater than 0.5 suggesting that they are highly functionally redundant and interrogation of their double mutants would be more functionally informative than single mutants. However, the remaining ‘dark net’ paralogs may require activating conditions to unravel their biological roles through genetic
interactions. By measuring the fitness of their single and double mutants on sixteen alternative growth conditions, I was able to uncover significant growth defects for 25 double mutants and 91 single mutants on at least one non-standard condition. Given previous findings (Costanzo et al. 2010, Chapter 2) that genetic interaction degree correlates with single and double mutant growth defects, it will likely be functionally informative to profile mutants involving these genes on conditions on which they display a growth defect. In fact, preliminary data for ESL1-ESL2 is encouraging in this regard, and systematically assessing the remaining duplicates in this manner is likely to uncover their functional relationships and provide deeper insight into their cellular roles.

4.2.2 Empirically showing asymmetric evolution of paralogs due to structural entanglement

In Chapter 2, I noted that one of my goals was to determine why there are paralogs with a range of triple mutant interaction fraction. I hypothesized that some paralogs may be ‘stuck’ in their evolution and not able to partition functions between the two sister paralogs (sub-functionalize) any further because there are some overlapping functions on which each sister relies for its specific functions (i.e. the paralogs are structurally and/or functionally “entangled”). These overlapping functions could be: 1) highly constrained regions of sequence (e.g. a domain); 2) secondary, tertiary or higher order folding structure; 3) regulatory mechanisms and/or 4) localization mechanism. Each of these possibilities represents a potential source of adaptive conflict, and therefore a potential escape through duplication.

Since it is not possible to measure functions directly and uniformly across all genes, measuring entanglement at the sequence level offers an optimal compromise. To this end it would be interesting to measure entanglement using co-evolving residues for all genes in *S.cerevisiae* using sequences from many species in the yeast clade. Residue co-evolution determines residues at which change is selected for in order to compensate for a random mutation at another residue and has been used to study protein conformations, ligand-mediated residue couplings and interdomain interactions (Marks et al. 2011; Morcos et al. 2011; Marks et al. 2012; Hopf et al. 2014). Presumably, it means that co-evolving residues work together (often, but not necessarily in close physical proximity) to carry out one or more functions. It is possible that the farther
apart these residues are the more likely they are to span more than one domain perhaps suggesting functionally overlapping domains. It is possible that the higher the frequency of such co-evolving residues in a pre-WGD homolog, the more likely that the post-WGD paralogs retained functional overlap and thus exhibit a high fraction of trigenic interactions and vice versa. It would also be possible to probe another more general hypothesis in relation to duplicate survivability, where I would predict that ancestral genes with a higher structural entanglement would less often be duplicated than those with lower entanglement.

4.2.3 Dissecting mechanisms of triple mutant interactions

Having generated a wealth of triple mutant genetic interaction data for duplicated genes and random singletons, the next step is to mechanistically characterize the resulting three way interactions. In Chapter 1, I discussed possible biological explanations for negative triple mutant interactions such as higher-order redundancy involving three genes carrying out a specific function; or protein complexes composed of up to two non-essential subunits; and compensatory repurposing which occurs when an alternative pathway could be rewired to substitute for a perturbation to other two parallel pathways. Other mechanisms, which were broadly described in Chapter 3, could include localized network perturbations that were observed for simultaneous perturbations to ER, Golgi and peroxisome, which could be extended to other cellular compartments; intracellular trafficking flow or signaling reduction; or pleiotropic effects involving phenotypic capacitators that buffer phenotypic variation due to higher-order genetic interactions. In addition, my triple mutant SGA scoring strategy could be used to accurately quantify positive triple mutant interactions. Some biological interpretations noted in Chapter 1 include examples of classical epistasis in a multi-component pathway or suppressive effects of a toxic effect, which could emerge in the absence of two other regulators. Constructing a multi-dimensional plot of the directionality of double mutant and triple mutant effects would shed light on such genetic interaction classes. Overlapping with GO co-annotation, protein-protein interaction, DNA-protein interaction standards would aid in guiding the follow up experiments. Thus, going forward it would be useful to continue understanding the mechanisms responsible for triple mutant genetic interactions.
4.2.4 Understanding the contribution of higher order genetic interactions in human disease and as it relates to missing heritability

Studying higher-order genetic interactions in *Saccharomyces cerevisiae* provides a simple model for dissecting complex patterns of genetic wiring. However, ultimately the goal is to understand the genetic architecture underlying human disease. Recently, it has become more widely recognized that the interplay of multiple genes may explain the missing heritability in humans with regards to physiological traits and disease states (Zuk et al. 2012). This is consistent with studies probing yeast strain-specific differences in phenotypes, which is attributed to genetic background. For example, closely related genomes Σ1278b and S288c, show phenotypic differences in growth programs resulting from multiple modifier loci frequently involving three, four or more genes (Dowell et al. 2010; Ryan et al. 2012b). With the advance of genome editing tools such as CRISPR-Cas9 system (Esvelt et al. 2013; Mali et al. 2013) and unique human cell lines such as KBM7 cells which represent a reversible gene trap collection that enables haploid mutagenesis analysis in human cells (Burckstummer et al. 2013), it should be possible to engineer and test the effects of combinatorial perturbations or interaction of disease variants directly in the human cell model. In fact, higher-order combinatorial genetics have been recently described where authors created high-coverage libraries of ~ 1,500 pair-wise and ~ 50,000 three-wise barcoded combinations of 39 human microRNA precursors in an effort to identify miRNA combinations that expose vulnerabilities of drug-resistant cancer cells to chemotherapy (Wong et al. 2015). Thus, higher-order genetic interactions are becoming widely recognized as important biological phenomena for addressing the genotype to phenotype challenge and may be key in tracking down the factors responsible for the ‘missing heritability’ of human disease.
References


