THE EXTENSIVE AND CONDITION-DEPENDENT NATURE OF EPISTASIS AMONG WHOLE-GENOME DUPLICATES IN YEAST

by

Gabriel Musso

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

Abstract

The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast
Gabriel A. Musso
PhD
2010
Molecular Genetics
University of Toronto

Immediately following a gene duplication event, if both gene copies are to be fixed into a species' genome there is a period of enhanced selection acting on either one or both duplicates (paralogs) that results in some extent of functional divergence. However, as redundancy among extant duplicates is thought to confer genomic robustness, a consequent question is: how much functional overlap exists between duplicates that are retained over long spans of evolutionary time? To examine this issue I determined the extent of shared protein interactions and protein complex membership for paralogous gene pairs resulting from an ancient Whole Genome Duplication (WGD) event in yeast, finding retained functional overlap to be substantial among this group. Surprisingly however, I found paralogs existing within the same complex tended to maintain greater disparities in expression, suggesting the existence of previously proposed "*transcriptional back-up*" mechanisms. To test both for existence of such mechanisms and for any phenotypic manifestation of their shared functional overlap I surveyed for the presence of aggravating genetic interactions between 399 WGD-resultant paralog pairs. While these paralogs exhibited a high frequency (~30%) of epistasis, observed genetic interactions were not predictable based on protein interaction overlap. Further, exposure to a limited number of stressors confirmed that additional

instances of epistasis were only observable under alternate conditions. As only a small number of stress conditions were tested, the high frequency of genetic interactions reported appears to be a minimum estimate of the true extent of epistasis among WGD paralogs, potentially explaining the lack of overlap with protein interaction data. As it is impossible to survey an infinite condition space, Synthetic Genetic Array (SGA) screening of yeast strains carrying double-deletions of paralog pairs was used to assess functional redundancy among a group of the remaining non-epistatic paralog pairs. The resulting interactions demonstrated functional relationships in non-epistatic paralogs only obvious upon ablation of both duplicates, suggesting that these interactions had initially been masked through redundant function. These findings ultimately suggest an advantage to retained functional overlap among whole genome duplicates that is capable of being stably maintained through millions of years of evolutionary time.

Acknowledgements

First and foremost I would like to acknowledge the support and insight of my two supervisors Dr. Andrew Emili and Dr. Zhaolei Zhang, without whom none of this work would have been made possible. Dr's Emili and Zhang made time to hear out my every idea, and then did whatever they could to ensure that I was best equipped to pursue it. I would also like to acknowledge the members of my advisory committee Dr. Elizabeth Tillier and Dr. Charlie Boone for years of support and guidance.

There are also several researchers at the CCBR that graciously opened up their labs so I could pursue my own research, and in that vein I would like to sincerely thank Dr. Corey Nislow, Dr. Guri Giaever, and again Dr. Charlie Boone. I would also like to thank the many students, post-docs, and technicians in the Emili, Zhang, Nislow/Gaiever, and Boone labs who took time out from their own work to show me a little bit about yeast genetics, specifically Dr. Michael Costanzo.

Finally, I would like to acknowledge my wife Natalie, who had to listen to stories about duplicated yeast genes for 4 years, but never seemed uninterested.

ABSTRACT ..... **II**

ACKNOWLEDGEMENTS ..... **IV**

LIST OF TABLES ..... **VII**

LIST OF FIGURES ..... **VII**

LIST OF ABBREVIATIONS ..... **IX**

CHAPTER 1: INTRODUCTION **1**

1.1. THE IMPORTANCE OF UNDERSTANDING GENE DUPLICATION 2
1.2. DUPLICATION EVENT TYPES AND RATES 3
1.3. EVIDENCE OF WHOLE GENOME DUPLICATION IN VARIOUS LINEAGES 8
1.4. S. CEREVISIAE AS A MODEL TO STUDY FUNCTIONAL DIVERGENCE FOLLOWING WGD 12
1.5. ADVANTAGES AND POTENTIAL BIASES OF STUDYING WGD-RESULTANT PARALOGS IN YEAST 14
1.6. MODELS OF FUNCTIONAL DIVERGENCE FOLLOWING DUPLICATION 17
1.7. THE CONTRIBUTION OF DUPLICATED GENES TO GENOMIC ROBUSTNESS 22
1.8. DIRECT FUNCTIONAL COMPARISONS OF YEAST DUPLICATES CONDUCTED TO DATE 26
1.9. PARALOGS AND EPISTASIS 28
1.10. PROJECT RATIONALE 34

CHAPTER 2 RETENTION OF PROTEIN COMPLEX MEMBERSHIP BY WHOLE-GENOME DUPLICATES IN YEAST **37**

2.1. INTRODUCTION 38
2.1.1. The importance of protein interactions in the eukaryotic cell 38
2.1.2. The retention of paralogs in protein complexes 39
2.1.3. Specific rationale and hypothesis 41
2.2. METHODS 44
2.2.1. Datasets used 44
2.2.3. Metrics used for analysis of network properties 47
2.3. RESULTS 49
2.3.1. Paralogs have different interaction properties than non-paralogs 49
2.3.2. Validation of findings 51
2.3.3. Paralogs more frequently co-complexed 55
2.3.4. Co-clustered paralogs are under tighter evolutionary constraint 58
2.4. DISCUSSION 61

CHAPTER 3 EXTENSIVE, CONDITION-DEPENDENT EPISTASIS AMONG WGD PARALOGS 63

3.1. INTRODUCTION 64
3.1.1. Specific Rationale / Hypothesis 64
3.2. METHODS 67
3.2.1. Assessment of synthetic lethality 67
3.2.2. Screening for phenotypic rescue 69
3.2.3. Information gain analysis to indicate features predictive of epistasis 71
3.2.4. Environmental screening and barcode analysis 71
3.2.6. Analysis of physical interactions 74
3.2.7. Comparisons of conservation and expression 75
3.2.8. Determination of instances of multiple paralogy 75
3.2.9. Statistical Analysis 76
3.3. RESULTS 77

3.3.1. Frequent phenotypic buffering between WGD-resultant duplicates ... *77*
3.3.2. Further buffering relationships only evident under stress conditions ... *80*
3.3.3. Experimental condition impacts composition of paralogs deemed epistatic ... *84*
3.3.4. Paralogs buffering under standard conditions are highly conserved ... *85*
3.3.5. Physical interactions are not indicative of phenotypic buffering ... *92*
3.3.6. Epistasis present among paralog pairs containing only one essential gene ... *94*
3.4. DISCUSSION ... 96

CHAPTER 4 DISCOVERING FUNCTIONAL REDUNDANCIES THROUGH TRIPLE-DELETION SGA ... 99

4.1. INTRODUCTION ... 100
4.1.1. Specific rationale and hypothesis ... *101*
4.2. METHODS ... 102
4.2.1. Construction of SGA double-deletion query strains ... *102*
4.2.2. SGA pinning protocol ... *105*
4.2.3. Scoring of interactions and batch correction ... *106*
4.2.4. Analysis of function ... *107*
4.2.5. Microscopy ... *107*
4.3. RESULTS ... 108
4.3.1. Selection of strains for SGA screening ... *108*
4.3.2. Modified SGA protocol shows reproducible results ... *110*
4.3.3. Triple-deletion SGA results require increased stringency in identification ... *114*
4.3.4. Trigenic interactions reveal insight towards overlapping functions ... *117*
4.4. DISCUSSION ... 122

CHAPTER 5 THESIS SUMMARY AND FUTURE DIRECTIONS ... 124

5.1. THESIS SUMMARY ... 125
5.2. FUTURE DIRECTIONS ... 130
5.2.1. Using further large-scale screening to identify functional overlap ... *130*
5.2.1.1. Increased breadth of triple-deletion SGA screening ... *130*
5.2.1.2. Determining functional overlap using other SGA-based techniques ... *131*
5.2.2. Assaying for the presence of transcriptional back-up mechanisms ... *133*
5.2.3. Determining the properties of paralogs in other species ... *134*
5.2.3.1. Examining the ancestral copies of epistatic paralogs ... *134*
5.2.3.2. Conservation of epistatic relationships among post-WGD yeast species ... *136*
5.2.3.3. The impact of WGD on robustness of the human genome ... *137*

APPENDIX ... 139

GENERATION OF EXPERIMENTAL DATA ... 140
Yeast-2-Hybrid ... *140*
Affinity purification ... *144*
Tandem affinity purification ... *146*
TEXT MINING ... 148
CLUSTERING OF INTERACTION DATA ... 150
Clustering algorithms ... *151*
INTRODUCTION TO GRAPH THEORY ... 153

REFERENCES ... 163

List of Tables

Table 1-1 Presence of duplicates in various eukaryotic species 4

Table 1-2 Duplications by functional category in yeast 15

Table 2-1 Size of interaction datasets used 42

Table 2-2 Elevated interaction overlap for true paralog pairs 52

Table 3-1 Results of RSA screening 78

Table 3-2 Pairs indicated as sensitive to stress 83

Table 4-1 Genes surveyed through triple-deletion SGA 111

Table 4-2 Degree of the surveyed paralog pairs 118

Appendix Table 1 Epistacy detected using RSA and GCA 155

Appendix Table 2 Detected trigenic interactions 159

List of Figures

Figure 1-1 Basic mechanisms of gene and genome duplication 5

Figure 1-2 Establishing paralogy through synteny 10

Figure 1-3 Suggested location of the ancestral WGD in the yeast lineage 11

Figure 1-4 Models of functional divergence following duplication 19

Figure 1-5 Existing epistatic growth models 31

Figure 1-6 Synthetic Genetic Array screening 33

Figure 2-1 Network interaction diagrams 43

Figure 2-2 Experimental workflow 45

Figure 2-3 Overlap in interaction coverage for the three datasets 50

Figure 2-4 Mass spectrometry detection of duplicates 54

Figure 2-5 Frequent co-clustering of paralog pairs 56

Figure 2-6 Differential properties of CC and NCC paralogs 60

Figure 3-1 Experimental outline 66

Figure 3-2 Random Spore Analysis 68

Figure 3-3 Growth Curve Analysis 70

Figure 3-4 Assaying genetic interaction in multiple conditions 72

Figure 3-5 Stress responsive paralogs 81

Figure 3-6 Functional composition of epistatic paralogs 86

Figure 3-7 Impact of additional duplicates on epistasis 88

Figure 3-8 Properties of epistatic paralogs 89

Figure 3-9 Epistasis by functional category 91

Figure 3-10 Phenotypic rescue of essential paralogs 95

Figure 4-1 Detection of redundant functions through triple-deletion SGA 103

Figure 4-2 Creation of query strains 104

Figure 4-3 Correlation in SGA profile by functional category 109

Figure 4-4 Modifications in SGA protocol 112

Figure 4-5 Similarity in interactions over replicates 113

Figure 4-6 Single-deletion hits found in double-mutant screen 115

Figure 4-7 Accuracy of triple-interaction detection 116

Figure 4-8 KIN1/KIN2 double deletion does not effect polarity 121

List of Abbreviations

Δ	deletion (gene)
^{o}C	degrees Celsius
AD	activation domain
Arabidopsis	*Arabidopsis thaliana*
BD	binding domain
bioGRID	general repository for interactions database
bp	base pair(s)
CAI	codon adaptation index
CC	co-clustered
C. elegans	*Caenorhabditis elegans*
C. glabrata	*Candida glabrata*
DDC	duplication degeneration coplementation
DNA	deoxyribonucleic acid
EAC	escape from adaptive conflict
FBA	flux-balance analysis
GCA	growth curve analysis
GFP	green fluorescent protein
GO	gene ontology
HCS	high-content screening
hyg	hygromycin B
IG	intersect group
IO	interaction overlap score
kan	kanamycin
kb	kilobase pair(s)
Ka	non-synonymous substitution rate
K. waltii	*Kluyveromyces waltii* (also known as *Lachancea waltii*)
K. polysporus	*Kluyveromyces polysporus*
MIPS	Munich information center for protein sequences database
MoBY	molecular barcoded yeast

nat	nourseothricin
NCC	non-co-clustered
NSI	non-shared interaction score
ORF	open reading frame
PCR	polymerase chain reaction
PPI	protein-protein interaction
RaR	retinoic acid receptor
RFP	red fluorescent protein
RNA	ribonucleic acid
RNAi	inhibitory ribonucleic acid
RSA	random spore analysis
S. bayanus	*Saccharomyces bayanus*
S. castellii	*Saccharomyces castellii*
S. cerevisiae	*Saccharomyces cerevisiae*
SD	minimal media
SDL	synthetic dosage lethality
SGA	synthetic genetic array
S. paradoxus	*Saccharomyces paradoxus*
S. pombe	*Schizosaccharomyces pombe*
SS	synthetic sick
SSD	small-scale duplication
SL	synthetic lethal
TAP	tandem affinity purification
ts	temperature sensitive
TM	trans-membrane
UG	union group
WGD	whole-genome duplication
wt	wild-type
Y2H	Yeast 2-hybrid
YGOB	Yeast gene order browser
YPD	rich media

Chapter 1:

Introduction

1.1. The importance of understanding gene duplication

Due to relaxation of selective constraints, duplication affords the opportunity for genes to develop novel functionality and as such has been seen as a substantial contributor to adaptation and speciation for over 70 years[1]. Generally recognized as a central figure in duplicated gene (paralog) analysis, Susumu Ohno wrote: "*natural selection merely modified, while redundancy created ... had evolution been entirely dependent upon natural selection, from a bacterium only numerous forms of bacteria would have emerged*"[2]. Ohno postulated that having an additional copy of a gene allowed it to bypass certain "*forbidden*" mutations, meaning those that would directly disrupt its function. In this sense, duplication (or perhaps more appropriately, the absence of selective pressure following duplication) provides the substrate for natural selection to mould as selective pressures dictate. While controversy exists regarding the selective forces acting before and after duplication as well as the consequent advantages of retained redundancy (discussed in detail below), the importance of gene duplication to the evolution of the eukaryotic cell is consistently regarded as being vital[3].

Retaining duplicate gene copies has notable advantages, including an increased propensity to adapt to changing environments. For example, whereas the mouse genome contains two copies of the gene encoding the photoreceptor Opsin, humans have retained multiple copies, allowing vision over a broader spectrum of light and facilitating (or perhaps facilitated by) dependence on visual sensory perception[4]. Conversely, both the human and mouse genomes contain over 1000 genes encoding olfactory receptors which are reasoned to have arisen through multiple duplications, but while many of the human genes have become pseudogenized (i.e. functionally inactivated), those in mouse have

been maintained for olfactory perception[5]. Such examples are pervasive in eukaryotic biology with duplicates abundantly spread throughout virtually all biological processes (see **Table 1-1**). It is arguably the presence of duplicates within these processes that both mediates insensitivity to perturbation and facilitates adaptation. Discussion of previous research aimed at analyzing the functional role of duplicates in extant genomes is the central theme of this chapter.

1.2. Duplication event types and rates

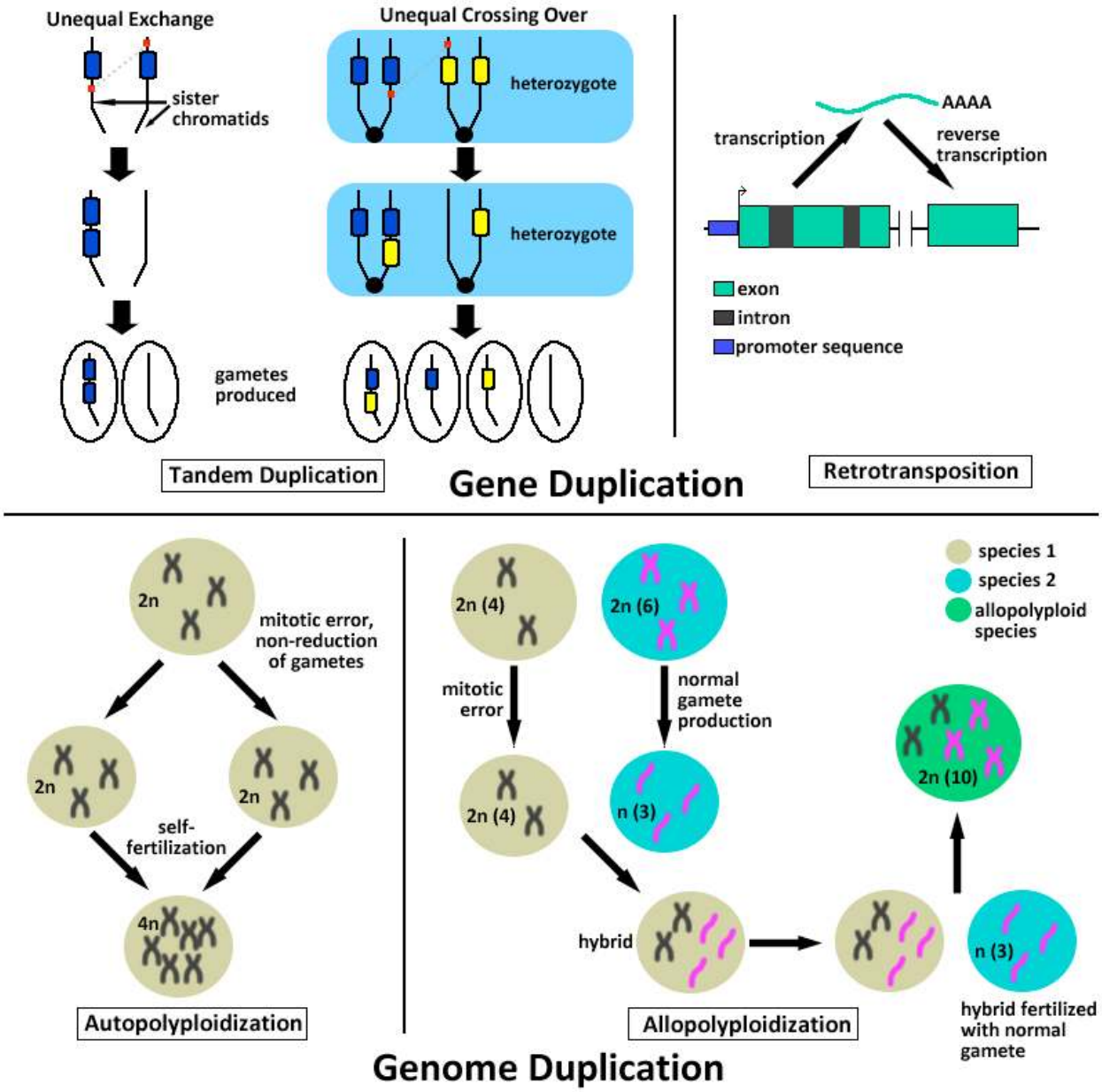
Duplication events occur through several mechanisms as part of the mitotic and meiotic processes and can involve the replication of anything from single genes to entire genomes. One of the most commonly observed consequences of duplication is the tandem duplicated segment, which by definition is multiple similar genes within close chromosomal proximity. Tandem duplications occur mainly through unequal exchange or crossing over occurring during mitosis and meiosis, respectively. In these events either paired chromatids or chromosomes exchange nucleotide sequences unevenly (see **Figure 1-1**), resulting in daughter cells of varying zygosity for genes in the exchanged region. In addition to crossing-over, tandem duplications can also occur through errors in homologous recombination or DNA damage repair[6]. If occurring in gametic cells, such duplications could eventually become fixed[2]. However it should be mentioned here that most genes arising through this or any other mechanism are subsequently lost through random mutation, suggesting that fixed genes bestow an alteration or generation of function that increases fitness (possible scenarios for this are discussed below).

Table 1-1 Presence of duplicates in various eukaryotic species

Species	Common Name	Number of Genes	Number of Duplicates	Evidence for WGD in ancestor
Caenorhabditis elegans	Nematode	18424	8971 (49%)	No
Drosophila melanogaster	Fruit fly	13601	5536 (41%)	No
Arabidopsis thaliana	Plant (thale cress)	25498	16574 (65%)	Yes
Homo sapiens	Human	20000-25000	15343 (~60-70%)	Yes
Saccharomyces cerevisiae	Budding yeast	6241	1858 (30%)	Yes

Number of duplicated genes predicted for several sequenced eukaryotic species as reported by Zhang *et al*[7]. Evidence for duplication was based on BLAST alignment[8] and the number of assigned duplicates can vary depending on stringency in definition.

Figure 1-1 Basic mechanisms of gene and genome duplication



Depicted are several common mechanisms for both gene and genome duplication. Beginning at top left and going clockwise, two well-described mechanisms for tandem duplication are unequal exchange or crossing over occurring due to mis-alignment (indicated by red squares and grey dotted lines) during mitosis and meiosis respectively. Retrotransposition involves the reverse-transcription of transcribed mRNA sequences into the genome as cDNA. Allopolyploidy events involve the combination of the genomes of two species to increase the genetic complement (one described case depicted). In contrast, autopolyploidies typically result from errors in the reduction of gametes among a single species. Portions regarding auto and alloploidization adapted from Campbell and Reece[9], and regarding tandem duplication adapted from Ohno[2].

Over time, further unequal crossing over or translocation events may disturb the tandem orientation of duplicates. Consequently, either inter- or intra-chromosomally-located duplicated genome segments over 1kb in length and with greater than 90% overall sequence similarity are referred to collectively using the blanket term 'segmental duplications'[10] (although due to this definition, segmental duplications also can include remnants of larger-scale duplication events such as aneuploidies or genome duplications; more below). Regions of segmental duplication cover 5% of the human euchromatic genome[11], and are thought to be selectively maintained[12].

Another less common mode of creating functional duplicates is retrotransposition, the process by which mRNA is reverse-transcribed and integrated into the genome. Since this involves a transcribed gene, the parental intronic structure and core promoter sequence are typically lost. For this reason, genes duplicated through retrotransposition are generally chimeric, containing a promoter sequence from another gene[13] while those genes that do not acquire a promoter sequence will eventually become pseudogenes. Perhaps indicative of the fact that obtaining a promoter is an unlikely occurrence, gene duplication via retrotransposition is appreciably rare, with a retrogene being fixed into a population's genome at approximately 1 gene per million years in the human evolutionary lineage[14] (by contrast, eukaryotes typically fix variants generated through segmental duplications at a rate of between 0.002 to 0.2 per gene per million years[3]).

Although not occurring with the frequency of tandem duplications or retrotranspositions, but creating duplicates on a tremendous scale, Whole Genome Duplication (WGD) represents a near or complete doubling in genomic content. There are two basic categories of genome duplications: autopolyploidies, in which an organism

doubles its own DNA content, and allopolyploidies, meaning that each genome copy came from a distinct species[15]. Both allopolyploidies[16] and autopolyploidies[17] have been suspected to occur in the lineages of eukaryotic species; however allopolyploidies are generally more common in plants[18] (however see Spring[19]). For clarity, all paralogs not resulting from a WGD event will herein be referred to as Small Scale Duplication (SSD)-resultant paralogs.

WGD events are presumed to occur due to a lack of disjunction among chromosomes after DNA replication[7] arising through either failures during mitotic segregation or following meiosis (termed genomic doubling and gametic nonreduction, respectively), or through polyspermy[18]. WGD events are frequent within plant species (between 30-80% of plants are polyploid[20]) although the reason for their being more tolerated is unclear. One notable, widely conserved ancestral gene family thought to have diversified through WGD is the *hox* genes, which are involved in morphological development. While the homeobox superfamily is reasoned to have duplicated largely through tandem (or segmental) duplication[21], presence of only one *hox* cluster in amphioxus, which diverged from the vertebrate lineage approximately 500 million years ago, versus 4 clusters in ray finned fish suggests that *hox* genes specifically may have expanded through 2 ancestral duplication events[22].

The increase in ploidy resulting from a WGD event is unstable[23], presumably due to an inability to undergo proper chromosome segregation[24]. This leads to increased pressure for rapid gene loss and subsequent genome degradation[17]. In the case of yeast, gene loss occurred to the extent where post-WGD species were shown to contain only approximately 10% more extant genes than a related species diverging immediately prior

to WGD[17]. However even at eventual an increase of 10% in genomic content, WGD events are profoundly influential and their determination in ancestors of extant species has helped to explain advances in species complexity.

1.3. Evidence of whole genome duplication in various lineages

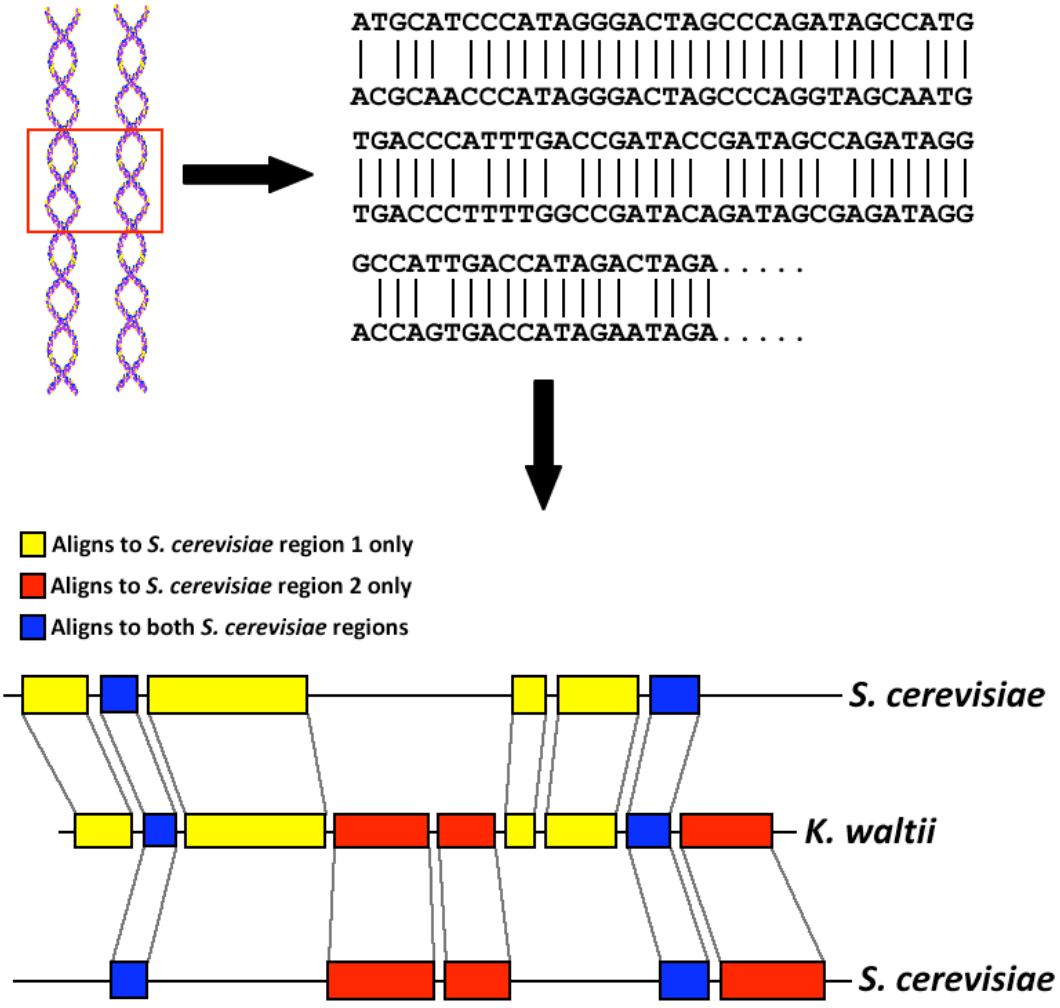
WGD events represent a unique potential source of increased organismal complexity and functional diversity and have been described as a major determinant in the evolutionary history of yeasts[17,25], plants[26], and vertebrates[27-29]. Because of the large amount of genomic substrate provided by their occurrence, WGD events are often associated with speciation, far more so than other modes of duplication. For example, occurrence of a WGD event in the vertebrate lineage is thought to have produced the teleost fishes[30], and as evidence of this WGD event (and likely one other) still have detectable impacts on the human genome today. Thus, understanding the nature of gene retention following WGD events could further our understanding of human genome architecture.

Before the availability of genome sequence data, evidence for genome duplication events was largely inferred using either map-based (meaning identification of duplicated chromosomes or chromosome sections) or tree-based (consistency among ratios of branch lengths for duplicated groups of genes) techniques[31], and was more circumstantial than concrete. In 1997 the then-recently sequenced budding yeast *S. cerevisiae* produced the first suggestion of ancestral genome duplication using a sequenced genome[25]; however despite the obvious presence of large, non-overlapping blocks in the genome sequence, controversy existed as to whether these had been created by a single event, or

multiple, smaller duplications[32,33]. Later Skrabanek and Wolfe identified as criteria for proving the existence of a WGD event that there be evidence of conserved, non-overlapping gene order in paired chromosomal regions of the genome, and that phylogenic support exist for a 2:1 orthological relationship with an outgroup[34]. These orthological relationships were firmly established in 2004 with the publication of genome sequences for additional yeasts[17,35]. Most notably, using the sequence of a species diverging close to the purported event (*Lachancea waltii*, formerly *Kluyveromyces waltii*), Kellis and colleagues aligned the *waltii* sequence to that of *cerevisiae*, showing syntenic (see **Figure 1-2**) regions apparent occurred twice as often in *cerevisiae*, and identified 457 gene pairs present in the *S. cerevisiae* genome as a result of this ancient genome duplication.

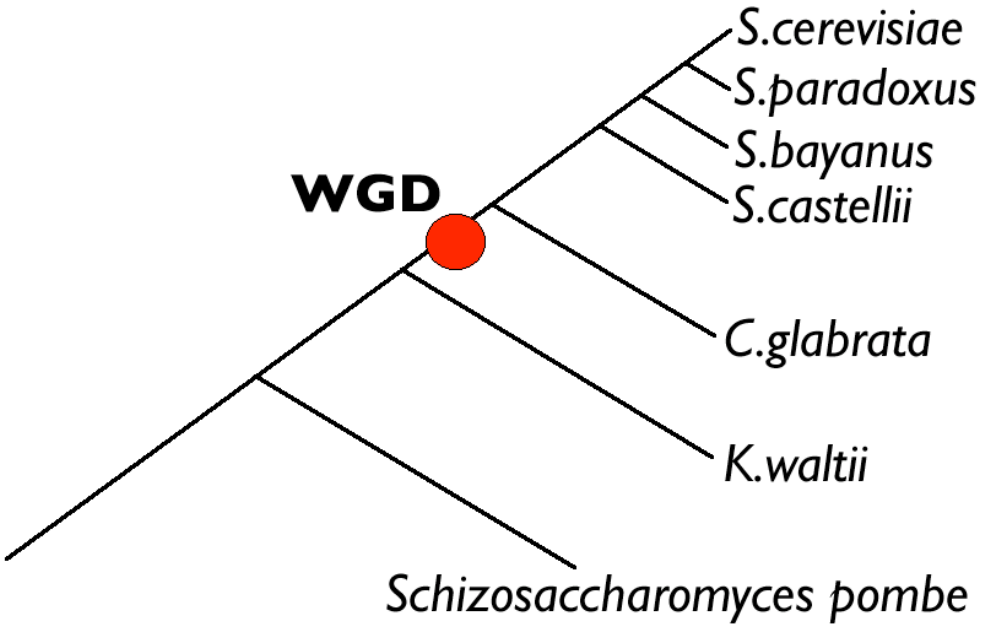
Adding additional outgroup species would allow Byrne and Wolfe to expand on the syntenic approach through the use of their Yeast Gene Order Browser (YGOB), and ultimately demonstrate that the number of *S. cerevisiae* WGD parlog pairs was as high as 551, with 599 WGD gene pairs also retained in *Saccharomyces castelli* and 404 in *Candida glabrata*[36] (see **Figure 1-3**). Interestingly, results from the YGOB study show that post-WGD species exhibit varying patterns of gene loss at 20% of nearly 3000 investigated loci[37]. Further, in approximately 5% of cases single-copy genes alternated their patterns of gene loss in post-WGD species, and are subsequently not true orthologs (i.e. asymmetrical gene loss). These differences in gene retention are thought to contribute to reproductive isolation following WGD[37] and may reflect subsequent adaptation to varied environmental pressures.

Figure 1-2 Establishing paralogy through synteny



Depiction of synteny modeled after that presented by Kellis and colleagues[17]. The process of assigning synteny begins by first finding regions of high similarity between two genomes through nucleotide sequence alignment of known or predicted open reading frames (as in upper right). Using genes conserved between the two species, regions containing orthologous genes in conserved order (syntenic regions) can be identified. Finding syntenic regions doubly-conserved in *S. cerevisiae* and once in *K. waltii* bolstered the case for an ancient WGD event occurring in the yeast lineage. The 457 genes in the double-conserved *S. cerevisiae* blocks were taken to be WGD-derived paralogs.

Figure 1-3 Suggested location of the ancestral WGD in the yeast lineage



Indicated is the suggested temporal location of the Whole Genome Duplication (WGD) event reasoned to occur between 100-150 million years ago in the yeast lineage. Phylogenic relationships are adapted from Wang *et al*[38].

A syntenic procedure was also used in the analysis of a teleost fish genome to demonstrate genome duplication in an ancestral vertebrate[39], and (albeit somewhat weaker due to lack of convincing outgroup comparison) evidence has been used to infer genome duplications in *Arabidopsis thaliana*. Since the *Arabidopsis* genome contains large duplicated segments in no more than 3 copies each, they are thought to have been created simultaneously[40]. Comparatively, the recent sequencing and analysis of the amphioxus genome[41] provides confirmation that two rounds of genome duplication occurred in the vertebrate lineage. The authors of the amphioxus study examined synteny on two gross orders: macro-synteny, meaning the conservation of whole chromosomes, and micro-synteny, conservation of local gene order. By doing so the authors demonstrated clear conservation of macro, but not necessarily conserved micro-synteny[41], ultimately suggesting that 2 ancestral duplication events had been followed by extensive gene loss. This phenomenon is at least anecdotally similar to what had been described in yeast[17], suggesting that a greater understanding of the nature of selective forces acting post-WGD in yeast may help explain selective pressures that influenced gene retention in vertebrates.

1.4. S. cerevisiae as a model to study functional divergence following WGD

Containing orthologs of many human genes[42], the budding yeast *S. cerevisiae* represents a valuable model eukaryote from which to gain insight into the functioning of our own cells. Due to its compact genome, ease of culture, and ability to exist in both the haploid and diploid state, budding yeast is one of the most commonly studied and functionally annotated model organisms, and is often the testing ground for advanced

high-throughput experimental techniques (many of which could be insightful in investigation of the function of paralogs). The model yeast *S. cerevisiae* was the first sequenced eukaryote[43], with its gene expression patterns investigated extensively through DNA microarrays[44,45] and protein abundance and sub-cellular localization studied via large-scale fluorescence labeling[46,47] analysis, and the first to be used for large-scale protein interaction screening including by our group[48,49]. Large-scale experimental results plus a wealth of data resulting from over a half-century of experimental assessment (much of which is assembled, manually curated and publically available at the Saccharomyces Genome Database[50]) have lead to an un-paralleled functional categorization of yeast gene products.

In addition to in-depth descriptions of gene-product function, the genome of *S. cerevisiae* can be easily manipulated[51], facilitating the creation of a strain collection containing deletions of nearly every known open reading frame (ORF) not encoding an essential gene[52]. This collection was designed with unique nucleotide sequences inserted at each deletion site in a process known as molecular barcoding[52]. Barcode sequences can be identified through microarray hybridization analysis, facilitating parallel assay of pooled yeast strain cultures and thus the systematic identification of gene-environment interactions. This is of specific relevance to this project as it could be used to assess whether functional overlap among duplicates is retained specifically to deal with adaptation to alternate environments (more on this in *Chapter 3*).

One of the most surprising findings to stem from the ordered deletion of genes in yeast is that very few genes (less than 20%) are required for cell viability under standard laboratory growth conditions[52]. One potential explanation for this finding is a high

degree of robustness afforded biochemical pathways by the presence of a large number of duplicated genes[53] (in this context I define robustness as the ability to withstand mutation or gene-product inhibition). In addition to having over 1000 (or more, depending on stringency in identification) genes reasoned to have arisen by SSD events[54], nearly 15% of ORFs in the extant *S. cerevisiae* genome originated from a single, ancestral duplication event[17]. The presence of such a large body of paralogs, coupled with the vast amount of phenotypic and functional data collected to date makes yeast ideal not only for overall comparisons of retained function, but to analyze these properties as they relate to other elements such as level of conservation, expression, or functional categorization.

1.5. Advantages and potential biases of studying WGD-resultant paralogs in yeast

While duplication events arise through seemingly random events occurring as a by-product of DNA replication, the paralogs that tend to be subsequently retained are anything but arbitrarily selected (see **Table 1-2**). In yeast (as in plants[55] and animals[56]), duplicated genes show a striking bias to contain high levels of transcription factors, kinases and transporters[57], although this trend can vary somewhat species to species[58]. The selective pressures underlying this functional enrichment can be subjects of speculation, but are currently uncertain. Further, although initial gene loss is rampant following WGD, the genes fixed in the genome over long expanses of evolutionary time represent a functionally unique subset of genes that make an interesting body to study. Specifically, genes in *S. cerevisiae* originating from the ancient WGD event are highly

Table 1-2 Duplications by functional category in yeast

Functional category	Number of genes with classification	WGD Paralogs in set (%)	SSD paralogs in set (%)
molecular function (unclassfified)	2782	28.22	17.28
transferase activity	627	11.62	11.52
structural molecule activity	331	11.43	2.65
hydrolase activity	752	11.25	25.35
protein binding	530	7.44	5.99
DNA binding	326	6.81	4.95
transcription regulator activity	319	6.53	2.07
transporter activity	368	6.26	14.29
oxidoreductase activity	265	5.35	7.49
enzyme regulator activity	208	5.26	1.73
protein kinase activity	129	4.90	2.76
RNA binding	288	3.72	0.58
other	286	3.45	6.22
ligase activity	166	2.27	2.76
lipid binding	70	1.72	0.69
isomerase activity	56	1.45	1.15
phosphoprotein phosphatase activity	47	1.36	1.15
signal transducer activity	44	1.27	1.27
peptidase activity	108	1.09	4.26
translation regulator activity	52	1.09	0.46
lyase activity	83	0.73	2.53
nucleotidyltransferase activity	71	0.73	0.92
motor activity	16	0.54	0.35
helicase activity	83	0.36	5.76
triplet codon-amino acid adaptor activity	299	0	0

Shown are the number of duplicates originating from putative WGD and SSD events in each functional category for *S. cerevisiae* (1102 and 811 paralogs representing WGD and SSD respectively). WGD Paralogs are those identified by Byrne *et al*[36], and SSD paralogs were stringently identified in a manner similar to that used by Gu *et al*[53]. WGD paralogs are deficient for hydrolase and transporter activity, but are enriched for genes with structural molecule activity[59], and have a greater proportion of genes with unclassified function.

enriched for ribosomal and catalytic proteins when compared to SSD paralogs (again, as in plants[60]). Further, SSD paralogs show an under-enrichment for transcription regulators that is not seen in WGD paralogs[61].

In addition to having unique functional biases, yeast whole-genome duplicates are far less likely to contain genes essential for cell viability (less than 10% essential versus nearly 20% genome average), and also tend to retain shared physical interactions at a rate greater than SSD-resultant paralogs of comparable age[59] (more on this below). One potential explanation for this retained overlap in function is thought to be due to maintained balance. For example, it has been suggested that duplication is preferential among genes with limited connectivity within the interaction network[62], perhaps attributable to the fact that duplicating a gene with many interactions will disrupt too many processes as to be compensated for[62,63]. However, since WGD events entail the duplication of entire networks or pathways, associations can be maintained post-duplication and balance is not affected, potentially facilitating retained functional redundancy among WGD-paralogs[64]. Assertions regarding the retention of physical interactions for WGD-resultant duplicates remain speculatory however, as protein interaction data has traditionally been too sparse to facilitate comprehensive examinations (discussed in more detail in *Chapter 2*).

Although distinct from SSD paralogs from an evolutionary and functional standpoint, WGD-resultant duplicates represent a group of genes with inherent diversity in functional composition and conservation. As mentioned briefly above, since all genes in this set have an identical time since duplication, large-scale comparisons of functional overlap with properties such as sequence conservation and co-expression can be

conducted without needing to scale or correct for duplicate age. Additionally, since WGD events are established through the detection of synteny and not merely on the basis of sequence similarity, the case for shared ancestry is strengthened. Among the questions remaining to be answered by this set are: (i) to what extent duplicates of substantial age (in this case 100-150 million years) retain functional associations, (ii) if overlapping function is predictable based on other observable properties such as sequence conservation and co-expression, and ultimately (iii) what model best describes the nature of function retention for paralog pairs. Since they vary in functional composition, any findings resulting from analysis of WGD-resultant paralogs may not be directly applicable to other duplicates of alternate origin, but the general properties of selection following WGD events may provide insight into understanding the human genetic landscape as nearly 2000 extant human genes are thought to have directly resulted from two ancient WGD events in a vertebrate ancestor[29,41].

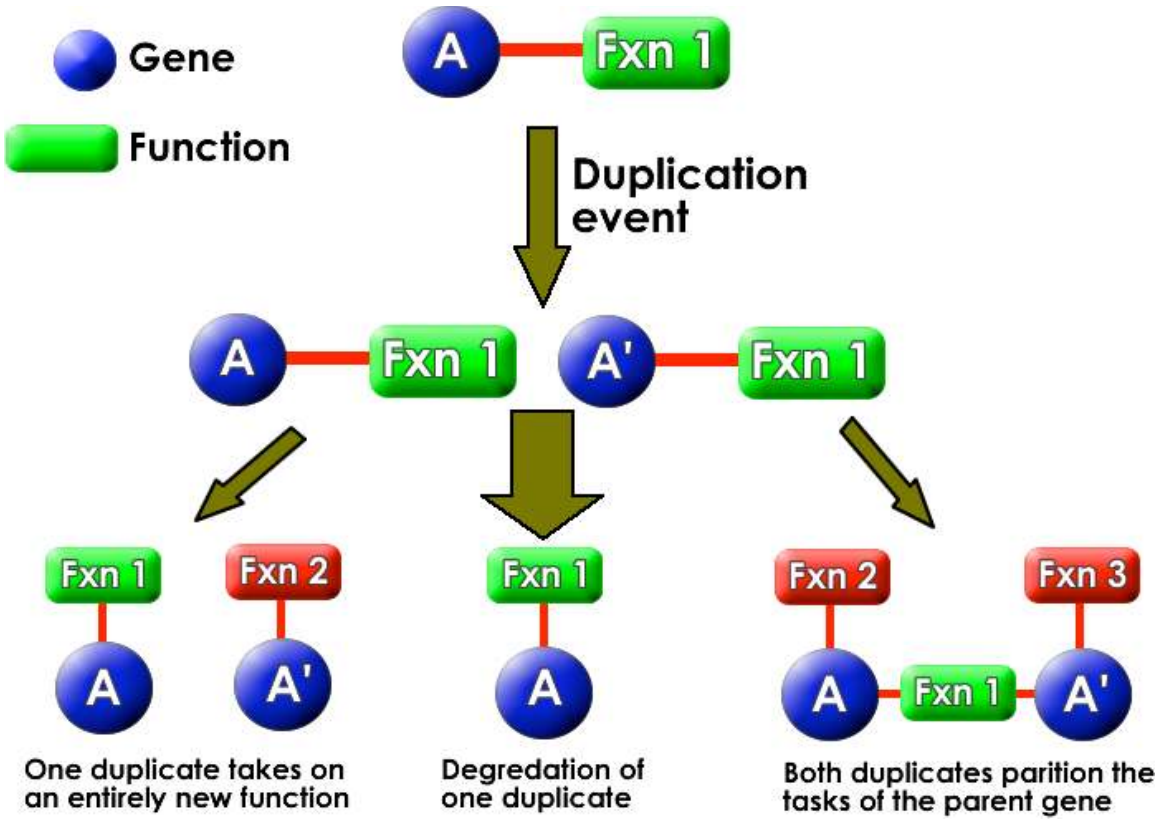
1.6. Models of functional divergence following duplication

Immediately following a gene duplication event, the resulting gene copies are presumably identical in sequence and expression, a condition not thought to be maintainable from an evolutionary standpoint[2,65]. Conceptually, if two genes are both performing the same function, even if this function is essential for cell viability, one gene can be subject to random mutation with the function still maintained, allowing some degree of functional dispersal. As mentioned briefly above, the vast majority of early research conducted on paralogs agreed that by far the most likely occurrence following duplication was degeneration and subsequent pseudogenization of one duplicate due to

random mutational innactivation[1,66,67]. In the many known instances of duplicate retention however, there has long been controversy as to what forces influence the preservation of multiple gene copies. Two main contradicting theories are commonly used to describe the potential long-term retention of paralogs (see **Figure 1-4**), each of which is discussed in detail below.

The classical theory used to describe functional dispersal following duplication is known as neo-functionalization, and purports that duplication allows one gene to perform (and retain) the ancestral function, while the second undergoes natural selection and ultimately derives a new function, else it becomes non-functionalized. In this model, retention of a given duplicate is based on the acquisition of a mutation in one paralog that endows a novel, selectable function[2] (a rare occurrence[68], thus explaining the great preponderance of duplicates to non-functionalize and become eliminated). As Kimura and Ohta wrote: "*Gene duplication must always precede the emergence of a gene having a new function*"[69]. The neo-functionalization theory is often originally attributed to Ohno, who noted that duplicated genes typically evolve at different rates, implying that the slowly-evolving gene is preserving the ancestral function. The observation of asymmetrical evolution has also more recently been noted for a great number of yeast WGD-resultant paralogs[17], initially suggesting that they had largely neo-functionalized. However, there is nothing conceptually regarding neo-functionalization that purports that the newly-obtained function be less rigorously protected through purifying selection. Further, Ohno himself acknowledged the potential for an alternative explanation when discussing isozymes with varying expression "*Because of differential use, the duplicated genes are exposed to different pressures of natural selection*"[2].

Figure 1-4 Models of functional divergence following duplication



In the event of gene duplication, by far the most likely occurrence is degradation of one gene copy from the genome. In the case where both genes are fixed however, two competing theories describe the nature of paralog retention: one ascribing acquisition of functional novelty to be vital to maintenance of both duplicates, the other a potential division of the ancestral function. As depicted here, partitioning of ancestral function has allowed for the generation of functional novelty (as in some escape from adaptive conflict models), however this not strictly a tenet of the Duplication Degeneration Complementation (DDC) model.

While the neo-functionalization model did offer an explanation as to how duplicates could be retained, the frequency of duplicates fixed in known genomes, specifically after supposed genome duplications[67,70-72], was too high as to be described by acquisition of randomly acquired, advantageous mutations[73]. Further, the neo-functionalization model did not seem to suit the case of multifunctional genes, which had been demonstrated in some instances to divide functions immediately following duplication (i.e. without requiring a period of non-functionalization for one duplicate[74]). In 1999, Force and colleagues proposed an alternative theory, that by which paralogs acquire complementary, degenerative mutations and by so-doing, partition the original functions of the ancestral gene among them (model as presented also known as Duplication-Degeneration-Complementation, or DDC)[73]. The authors described these sub-functions as any that: "*might involve the expression of a gene in a specific tissue, cell lineage, or developmental stage, or individual functional domains within the polypeptide coding portion of the gene*", and the process of dividing such functions among derived genes was termed sub-functionalization. This model suggested that degenerative mutations occurring in non-coding regions (or potentially even within the ORF[75]) could result in establishment of a complementary relationship among duplicates whereby varying functions could be distributed.

Since it relied on degenerative rather than advantageous mutations, the sub-functionalization model seemed more apt to describe the large number of duplicates retained following WGD events[76]. Evidence of this in yeast was elegantly demonstrated by Ambro van Hoof, who showed for several gene pairs that the single ancestral gene copy from *S. kluyveri* was capable of performing the individual functions of the paralogs

in *S. cerevisiae*[77]. However, systematic evidence demonstrating either the sub- or neo-functionalization models is lacking, with models based on population genetics typically supported only by sporadic tangible examples. Consequently, neither model currently is viewed as the dominant mode of functional dispersal post-duplication.

Since the publication of the neo- and sub-functionalization models, numerous variations and hybrid models have been presented to accommodate observed genetic relationships that do not seem to fit either mould[78-83]. One particularly notable variation is the proposed escape from adaptive conflict (EAC). What is sometimes seen as a limitation of sub-functionalization is that (as originally presented), the DDC model did not allow the potential to acquire adaptive mutations[82]. The term adaptive conflict specifically refers to the case in which a multifunctional gene cannot optimize one function without limiting its capability to perform others. Therefore in the EAC model, following sub-functionalization one duplicate can undergo either regulatory or coding-sequence changes that adapt its function towards new environments, as was demonstrated in yeast for two genes functioning in the galactose signaling pathway, *GAL1* and *GAL3*[82]. While the ancestral gene was likely to encode both an enzyme and a transcription factor, two genes in *S. cerevisiae* uniquely perform each of these functions.

Additionally, Nowak and colleagues[84] present three cases for the long-term preservation of redundancy: genes *A* and *B* perform a given function with equal efficacy, genes *A* and *B* perform the function at varying efficiencies, but are compensated for by reciprocal mutation rates, and lastly, *A* and *B* perform different functions, but *B* retains the capacity to perform the function of *A* at lower efficiency. However, just as above, practical examples of these mechanisms are even less common than for sub- and neo-

functionalization, presumably due to the unlikely occurrence of exact balance between either between selective force on the duplicates, or selective pressure and expression. Due to their low practical occurrence, existence of such mechanisms for long-term preservation of exact functional redundancy can be considered minimal.

The fact that no single model dominates in terms of being commonly accepted to describe evolution immediately post-duplication is perhaps owed to the fact that meaningful functional information has traditionally been lacking. Currently the amount of functional overlap retained by duplicated genes as well as the frequency and extent of functional retention ultimately remains unclear with compelling evidence presented[17,53,85-87] to demonstrate both limited and significant functional overlap among extant paralogs in the yeast *S. cerevisiae.* At stake is the central question of not only what influenced the initial retention of duplicates, but also to what extent they share function, and by extension, contribute to robustness of the genome towards mutation or other insult (discussed in detail below). Resolving these issues using large unbiased datasets is a central concern in this thesis.

1.7. The contribution of duplicated genes to genomic robustness

As data exists both supporting and contradicting a retained functional overlap among yeast duplicates, there has been a long-standing debate as to whether the presence of retained duplicates in a genome serves to increase genomic robustness (as defined above). Specifically, availability of gene-deletion mutants in yeast has facilitated analysis of the fitness costs associated with deleting a paralogous gene and has subsequently sparked controversy. In 2000 Andreas Wagner reasoned that if duplicates

were contributing to robustness, increased amino-acid sequence similarity should correlate negatively with their fitness defect upon deletion[88]. Instead, when studying 45 yeast paralogs, Wagner did not find a correlation between either sequence similarity and deletion cost, or sequence similarity and co-ordination of expression (also reasoned to be a hallmark of duplicates contributing to robustness) to be statistically significant, therefore concluding that genetic robustness was more greatly derived from non-duplicates[88]. However as this was a small subset of the large number of yeast duplicates, there was a substantial chance that this subset contained a bias that influenced the results.

Gu and colleagues later re-visited the concept of duplicates and genomic robustness using fitness data for 5,766 yeast ORFs (ultimately comparing growth rates among 1,147 and 1,275 genes that could clearly be considered duplicates and non-duplicates, respectively), finding differences between the fitness properties of these two groups[53]. The authors noted that duplicated genes had a significantly lessened deletion cost (also subsequently demonstrated to be true in nematode using growth following inhibitory RNA addition[89], and more recently in mouse using knockout phenotypes[90]). Further, deletion cost was positively correlated with similarity between duplicates, since less-similar duplicates were less likely to compensate, albeit still at a level greater than functionally-related singleton genes[53]. The same group would further demonstrate in a later publication that there was a relationship between the age of a duplicate and expression similarity, with genes becoming more divergent in expression over time[91]. Taken together these findings do suggest that functional overlap is maintainable over long expanses of evolutionary time, however that duplicates had a smaller fitness cost upon deletion has spurned two competing explanations.

The first presented explanation for the seeming expendability of duplicates was that there exist "*transcriptional back-up*" mechanisms between duplicated genes, whereby dissimilarly expressed paralogs can alter their expression patterns upon mutation or ablation of their sister gene[92]. Specifically, Kafri *et al* examined correlation in expression over 40 experimental conditions for hundreds of duplicated gene pairs and reported that duplicates with the greatest variance in co-expression, and thus the greatest potential for expression modulation, showed the least consequence upon deletion[92]. Further, the authors demonstrated that similarities in the promoter sequences of duplicates are actively maintained, suggesting that purifying selection preserves this mechanism. The same group would also later show that similar transcriptional mechanisms were present in multiple other eukaryotic organisms[93], suggesting that expression modulation was not a species-specific phenomenon.

In terms of more general evidence for transcriptional re-wiring following WGD, comparisons of the *C. albicans* and *S. cerevisiae* transcriptional networks during anaerobic growth showed alterations in the expression patterns for genes in *S. cerevisiae*, and Ihmels and colleagues attributed these changes to regulatory elements following the ancient WGD[94]. It should be mentioned when interpreting this analysis however that *C. albicans* is a pathogenic fungus, and therefore may have had an unusually accelerated evolutionary path following speciation. However, taken together these results provide at least anecdotal evidence for mechanisms of transcriptional compensation that could greatly benefit from validation through large-scale directed phenotypic assay (more below).

An alternative theory presented to explain the decreased fitness cost of duplicated gene deletions was that duplication was reserved for genes of limited importance to the organism[95]. He and Zhang showed that *S. cerevisiae* genes duplicated by both WGD and SSD in 7 other yeast species (but which were singletons in *S. cerevisiae*) were less likely to be essential, and generally showed less fitness cost upon deletion[95]. A notable caveat to this survey however is that 6 out of 7 of the species used in the analysis by He and Zhang differentiated from *S. cerevisiae* before the WGD event, meaning that this comparison does not take into account how functional relationships may have changed post-WGD. He and Zhang later demonstrated that the correlation between deletion fitness cost and expression similarity could be explained through correlation with a third variable, that of the number of protein interactions[96] (i.e. interaction degree). Specifically, He and Zhang suggested that high protein interaction degree, and not low correlation in expression, indicated essentiality for paralog pairs. Kafri and colleagues responded to this assertion by demonstrating that when controlling for degree there was still a clear relationship between fitness and concerted expression[97]. This demonstrated clearly that co-expression was inherently predictive of the functional relationship between duplicated genes. However, while analyses of expression and fitness can potentially highlight systematic mechanisms of retained function, they do not directly illustrate the extent, and nature of that functional overlap. Studies that directly compare the functional overlap among both WGD and SSD duplicates are discussed in detail below.

1.8. Direct functional comparisons of yeast duplicates conducted to date

While the above mentioned surveys of fitness and expression have instilled controversy regarding mechanisms and advantages of functional dispersal, examinations of the biological properties of extant paralogs based on the putative physical interactome[54,98-100] and reconstructed metabolic network[85,86] in *S. cerevisiaie* have generally implied extensive functional similarity among both WGD and non-WGD resultant paralogs, supporting an advantage to retaining substantive functional overlap.

The first large-scale analysis of duplicate function based on protein interaction data was published by Andreas Wagner[101] and was generated using a high-throughput Yeast-2-Hybrid protein-protein interaction dataset[102] (for a description of high-throughput interaction screening methodologies, as well as an introduction to the properties of graph theory as pertaining to interaction network analysis, see Musso *et al*, *Chemical Reviews*[103], portions of which are provided in the **Appendix**). Since paralogs were not any more likely than random to share protein interaction partners or exist in the same network sub-graph, Wagner concluded that yeast duplicates had largely lost their initial functional overlap. However as only a small number of paralogs were represented in the dataset used, this conclusion may have resulted from a lack of a truly reliable or comprehensive interaction dataset.

A later similar analysis using more extensive and likely more accurate interaction dataset (the manually curated set maintained in the Munich Information center for Protein Sequences (MIPS) database[104]) in yeast found alternate support for widespread functional overlap[100]. Specifically, Baudot and colleagues[100] used a functional classification algorithm on a network of over 4000 putative protein-protein interactions (generated

largely from small-scale surveys and collected using literature-mining techniques; see **Appendix**) to demonstrate that 41 yeast paralogs resulting from the WGD could largely be functionally grouped into the same gene ontology (GO)[105] category based on physical associations. This implies that paralogs generally maintain enough similar physical associations so as to be functionally linked, however this does not inherently support either neo- or sub-functionalization as a model for their divergence. Further, the small number of paralogs surveyed limits widespread interpretations as to the extent or advantages of retained functional overlap between paralogs in the interaction network. However, the publication of two landmark genome-scale proteomic surveys of protein complexes in budding yeast, one by our group, using affinity purification – mass spectrometry based on the Tandem Affinity Purification (TAP) procedure[48,49] several years after Wagner's initial publication facilitated re-analysis of the function of gene duplicates based on the properties of a far more extensive physical association network[54,99]. The results of these surveys were the initial focus of this project and are discussed in detail in the next Chapter.

In addition to function elicited through physical association with other proteins, function among duplicates has also been assessed using enzymatic information from the yeast metabolic network. Two recent studies investigating the gross function of paralogs showed a preponderance for duplicated genes to exist in yeast central metabolism pathways (i.e. glycolysis, pentose phosphate shunting and the citric acid cycle), which are demonstrably highly robust[106,107]. However these studies do not necessarily provide direct evidence of compensation, as they do not eliminate the possibility that genes existing in well-buffered pathways are more likely to retain duplicates. Specifically, it is

unknown whether it is the duplicates themselves that contribute to the robustness of these pathways, or whether it is the robustness of a system that permits tolerance to gene duplication.

To directly assay whether paralogs contribute to system robustness through compensation, Papp and colleagues used Flux Balance Analysis (FBA), a method for *in silico* modeling of the metabolic network and subsequent analysis of perturbation effects, to demonstrate that between 15-28% of extant duplicates functioning as metabolic enzymes in *S. cerevisiae* can be compensated for by their paralog through altered metabolic flux[85]. It should be noted as a caveat however that in FBA metabolic reactions are modeled as opposed to the underlying genes. Therefore while predictions can be made about whether compensation is on behalf of a paralog, ultimately experimental evidence is needed to verify these claims. Harrison and colleagues performed a similar FBA procedure, and although not studying paralogs specifically, demonstrated via multiple gene deletion assay that approximately 60% of their 98 predictions of compensation to be accurate. Notably Harrison *et al* also describe that the majority of their reported compensatory mechanisms were not observable under standard laboratory conditions. Therefore in order to effectively demonstrate compensation among paralogs, direct fitness experimentation, potentially conducted in multiple conditions, may be needed.

1.9. Paralogs and epistasis

The term epistasis is consistently used to refer to a genetic interaction between genes, although the exact meaning can vary depending on context. William Bateson

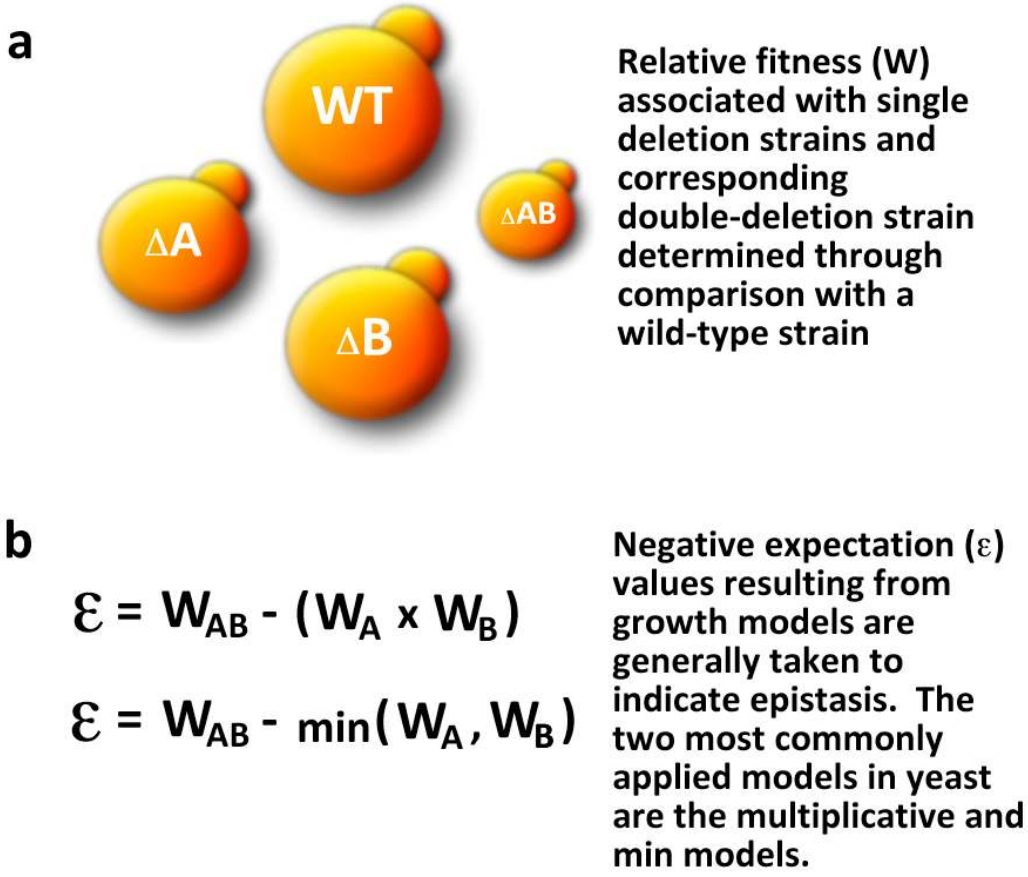
originally coined the term epistatic (literally "*standing over*") in 1909 to describe a case in which expression of an allele at a given locus masked the effects of a variant allele at a second locus[108]. Bateson used this term initially to describe heredity of coloration in rabbits, noting that alleles controlling coloration could be either epistatic or hypostatic depending on whether or not they were masking or being masked, respectively: "*We shall then speak of the determiner for grey as epistatic to that for black*". Bateson reasoned that the alleles were connected in the sense that they were acting in pathways ultimately related to the same phenotype[109]. One other notable property of this operational definition is that it reflects an asymmetric relationship: if allele *A* is epistatic to allele *B*, the reverse cannot possibly be true. Alternatively, RA Fisher would later refine this classical definition by considering epistasis in terms of quantitative traits. Fisher described epistasis as a deviation from the linear expectation of the contributions to phenotype when two alleles are expressed at different loci[110]. Notably, Fisher did not believe that epistasis substantially affected the evolutionary process, asserting that interactions among alleles in different loci did not contribute to overall fitness, and doubted whether they were transmissible. Fisher would have a long-standing debate on this issue with Sewall Wright who believed they were integral[111]. However while Fisher and Wright debated the influence of gene combinations on the overall fitness of a population, in the context of this thesis I will consider only how genes combine to affect the fitness of an individual.

Throughout this thesis I will refer to epistasis using a definition that is akin to that employed by the population geneticist[112]: when the effect of perturbing the function of two genes within the same species results in a phenotype not predictable based on the

individual deletions alone, those genes are said to be epistatic. This type of relationship is also referred to as a "*synthetic genetic interaction*", if a systematic experiment is performed to examine gene relationships (rather than measuring natural variance in a population). A commonly reported type of synthetic interaction is synthetic lethality, in which individual gene mutations or deletions result in a viable cell, but a double-mutant causes inviability[113]. Similarly, synthetic sickness describes a substantial worsening of phenotype upon allelic combination. Those epistatic interactions causing an un-expected worsening of phenotype (models used to derive phenotypic expectation are discussed in more detail below) are also generally termed "*aggravating*", and those that result in a suppression of a sick phenotype are usually referred to as "*alleviating*" (see **Figure 1-5**). Classically, aggravating genetic interactions are thought to occur among genes in alternate pathways that converge on a related essential process (as the cell can compensate for compromised function of either pathway individually, but not to defects in concert), and alleviating interactions among genes within the same pathway (since the pathway has already been compromised by single loss-of-function alleles, further intra-pathway perturbations do not have any consequence)[114].

Various models have been proposed to determine what is an appropriate expectation for epistatic genes when combining defects in growth (see **Figure 1-5**), however a recent comparison of the various growth models shows multiplication of fitnesses to be the most accurate at predicting known instances of epistasis[115]. Basically, this model supposes that if the double-mutant has a quantifiable fitness defect worse than expected based on multiplying the fitness defects of the two constitutive single-deletions,

Figure 1-5 Existing epistatic growth models



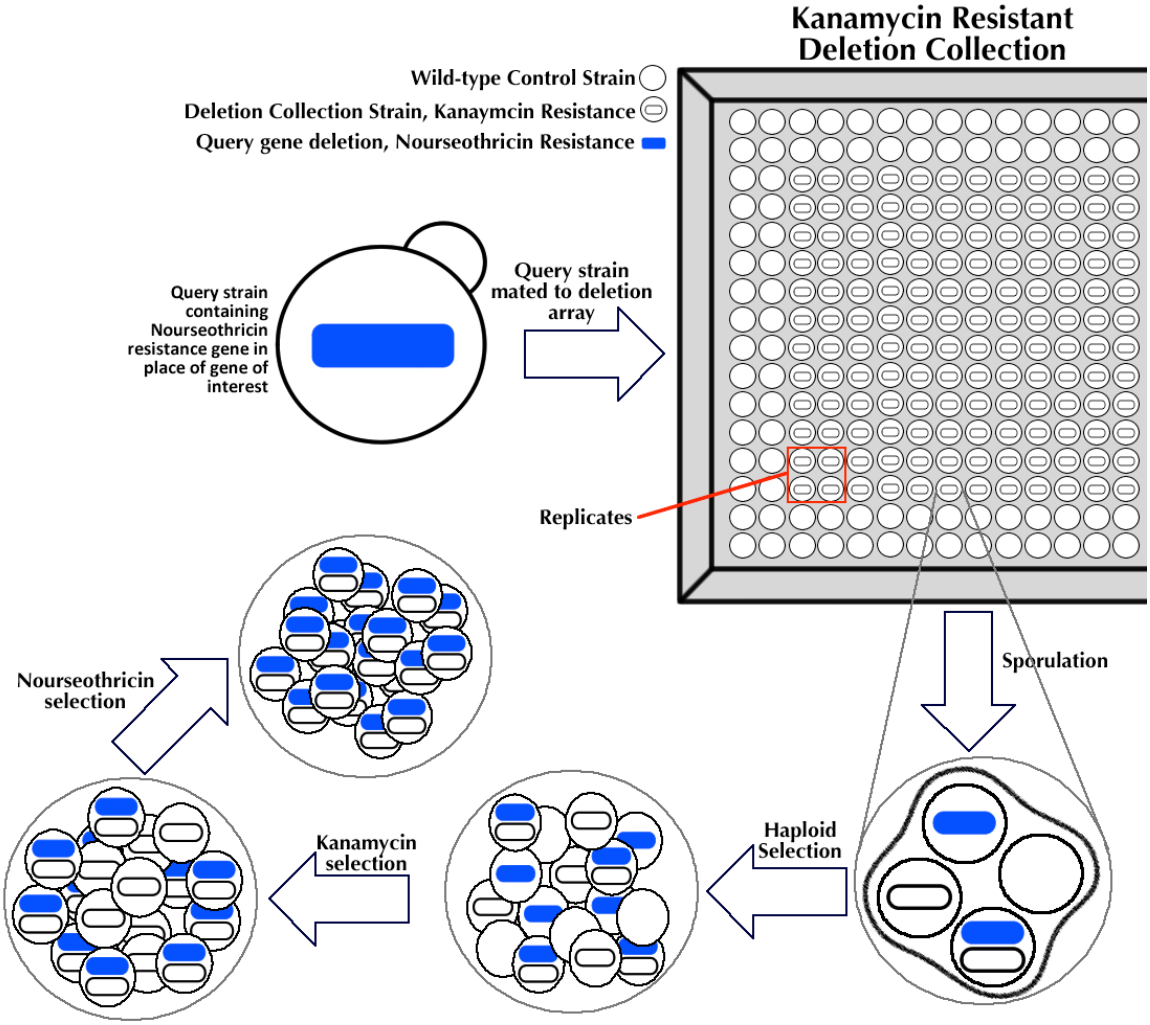
Observation of epistasis generally involves first determination of the quantitative growth defects associated with particular mutations (**a**). This is typically achieved through comparison with an appropriate wild-type control. Once growth defects are measured for single- and double-deletion mutants, the presence of epistasis is determined through comparison with an expectation model (**b**). While the min model (lower equation) is commonly applied in yeast genetic interaction screening, recent evidence suggests that the multiplicative model is reportedly more accurate at predicting bona fide instances of epistasis[116].

an aggravating genetic interaction is assumed. Due to the nature of these experiments, their assay is typically restricted to genes of non-essential function, although use of conditional allelic variants (such as temperature sensitive, or *ts* alleles) does allow for survey of genetic interactions involving an essential gene. Also, in addition to identifying the effects of reducing gene dosage (i.e. loss of function phenotypes), identifying gain-of-function phenotypes through over-expression can also be useful in identifying gene function[117], however gain of function experiments seem less likely to identify potential mechanisms of phenotypic buffering through functional compensation, and therefore are not addressed experimentally in this work.

The concept and functional informativeness of genetic interactions has been established in yeast for decades, facilitating systematic surveys of the functional architecture of biological processes such as the secretion system and cytoskeleton[113,118]. However the capacity to perform genetic interaction screening on a large-scale was first demonstrated in 2001 with the development of the Synthetic Genetic Array (SGA) screening technique by our Departmental colleagues Charlie Boone and Brenda Andrews and associates[119]. Their approach involves the mating of a single gene deletion strain (known as the query strain) against an ordered array of single gene deletion strains representing the majority of non-essential ORFs in the yeast genome and each carrying an alternate selectable marker (see **Figure 1-6**).

Large-scale screening conducted using the SGA technique[119,120] and subsequently-derived similar methods[121,122] has facilitated the creation of high confidence genetic interaction networks that can be used to infer gene function. These maps generally show

Figure 1-6 Synthetic Genetic Array screening



Depicted is the SGA procedure as published by Tong *et al*[120]. Briefly, a strain carrying *Nourseothricin* resistance in place of a gene of interest (represented by blue bar) is mated against a collection of over 4000 individual deletion mutants with *Kanamycin* resistance (represented by white bar) through pinning in 384-spot format. Each individual deletion strain is present in quadriplicate in the 385-spot plate, outer colonies contain control strains (*Kanamycin* resistance inserted in place of a null allele). Following mating, strains are sporulated, and haploid cells showing appropriate dual resistance to both *Nourseothricin* and *Kanamycin* selected through successive pinning steps. The plates are imaged, and colony sizes analyzed to determine growth defects.

an overall binary gene interaction rate of 0.5% (comparable density to physical interactions) indicating that there is substantial buffering of deletion effects in yeast (as has long been suggested[114]). However, given that the interaction frequency is 8-10 times more likely for genes with similar functional annotation, interactions for genes queried can be used to infer function (so-called "*guilt-by-association*") and have been useful in defining the molecular functions of unannotated yeast genes[120]. Given that they serve as an indicator of molecular function, and that they can be used to assay for phenotypic buffering (i.e. the cell should be able to cope with loss of either buffered paralog individually but not both in the same strain, manifesting as an aggravating genetic interaction), genetic interactions seem an ideal avenue to test for retained redundancies and compensatory mechanisms existing between yeast duplicates. Over the past 2 years, our group[123] and others[124] have used SGA to examine epistasis between paralogs in detail, the results of these and subsequent analyses are discussed in detail in *Chapter 3*.

1.10. Project Rationale

Despite the importance of gene duplication events, patterns of functional divergence of paralogs following a duplication event are poorly understood. Availability of an increasing number of genome sequences has facilitated postulation of various theories regarding the evolutionary mechanisms of duplicate retention, and emergence of various high and low-throughput interaction screening techniques has allowed detailed analysis of duplicate gene function; however several major questions remain un-answered. Notably, many genes retain functional overlap for long spans of evolutionary time and the extent, advantages, and nature of this retained redundancy remain unclear.

The purpose of the studies presented in this thesis is to examine in depth the overlap in terms of paralog function based on a large body of extant duplicates created by the WGD event, how extensive this overlap is, what potential role it has played in the natural evolutionary process, and to determine what biological properties define the subset of paralogs that have maintained detectable functional redundancy. This work serves to test one major and consistent hypothesis: that retained functional overlap is extensive among WGD duplicates in yeast, and that previous indications suggesting widespread functional dispersal were in fact based on incomplete or biased representations of functional relationships. I address this hypothesis by analyzing the physical and genetic interactions between these extant pairs.

Protein-protein interactions mediate the proper operation of most cellular processes, and genetic interaction screens can reveal the existence of maintained compensatory mechanisms. Therefore, in *Chapter 2* I use two newly-generated high-throughput interaction datasets to re-visit analyses of functional overlap among extant duplicates through shared physical associations. I demonstrate that there is substantial retention of interaction partners for WGD-resultant paralogs, and notably that those most similar in interactions are also most dis-similar in patterns of expression. As this provides support of previously proposed mechanisms of transcriptional back-up, I next turn to genetic interaction analysis to survey function among paralogs.

In *Chapter 3* I describe a series of experiments directly assaying for aggravating genetic interactions occurring between 399 surveyed WGD-resultant yeast paralogs. Finding that epistasis was very high among these paralogs, I also note that experimental condition has a notable effect on detection of epistasis. As survey of the near-infinite

condition space is not feasible, I turn to novel applications of the SGA technique for functional overlap among the remaining non-epistatic paralogs.

In *Chapter 4* I describe modifications made to the standard SGA procedure in order to survey for interactions using query strains containing double-deletions of non-epistatic paralogs. By comparison of these double-deletion profiles with the profiles of the constitutive single-deletion strains I attempt to identify notable differences, and subsequently, evidence of functional redundancy. I present as a result not only a protocol that can be used specifically to assay function of any gene pair, but also notable evidence of redundancy among non-epistatic paralogs.

Through these experiments I demonstrate the prevalence of functional association among WGD-resultant paralog pairs, ultimately suggesting that there is an advantage to long-term retention of functional overlap.

Chapter 2

Retention of protein complex membership by whole-genome duplicates in yeast

Portions of this chapter have been reprinted or adapted from *Musso *et al*[99]

I performed all experiments in the corresponding manuscript; Andrew Emili and Zhaolei Zhang supervised and advised the experimentation.

2.1. Introduction

2.1.1. The importance of protein interactions in the eukaryotic cell

Physical interactions among proteins mediate virtually every molecular process and as such their survey allows elucidation and ultimately classification of protein function. Traditionally physical associations among proteins have been assayed on the order of single complexes through the use of techniques like immunoprecipitation and co-sedimentation, however the recent emergence of high-throughput interaction detection techniques such as Yeast-2-Hybrid[125] (Y2H) and Tandem Affinity Purification followed by Mass Spectrometry[126] (TAP-MS), has facilitated large-scale determination of the network of protein interactions underlying biological processes. Delineation of associative protein units from these networks is useful in elucidating the mechanistic basis of complex molecular systems and in functionally characterizing interacting clusters of proteins (for an in-depth review of the formation of protein clusters from large-scale datasets, see **Appendix**).

Protein complexes, consisting of stable protein-protein interactions (PPIs), are ubiquitous and essential to the proper conduct of all eukaryotic functional pathways, serving to coordinate virtually every aspect of cellular biology[127]. The term 'protein complex' has traditionally been used to describe heteromeric groups of tightly associated proteins that interact to form a unified cellular component such as the ribosome or proteasome (approximately 30% of the gene products in yeast are involved in such complexes[104]). Yet, as large-scale interaction data has become increasingly available, and global interaction networks discovered, the idea of the protein complex has evolved somewhat to the notion of interconnected 'modules' consisting of groups of physically-

associated proteins functioning in a unified manner, although not necessarily with exclusive membership[128]. This has introduced a dichotomy in the interpretation of experimental datasets, as some would define the protein complex as a stable macromolecule, while others see it as a more dynamic, non-exclusive set of interacting proteins. Indeed, recent experimental evidence derived from genome-scale studies using yeast as a model system has begun to blur the heuristic boundaries that have historically been applied to define protein complexes as discrete biological articles. Consequently, researchers have begun to note heterogeneity, both in terms of the apparent limited correlation of attributes such as gene co-expressions and functional incongruence of putative members of certain protein complexes[129-131] and in profiles of genetic interactions[122].

One other intriguing aspect of resolved interaction networks has been the prevalence of cross-connections between various protein modules, as projected by certain high-throughput screens in yeast[132,133], which suggests a preponderance of cross-talk among biological systems. Therefore while shared membership within a complex or module is indicative of shared function, substantial functional overlap may still exist among non-co-complexed proteins. The role that duplicated genes might play in both the evolution of the protein complex, as well as the unique functional associations that link these modules remains unresolved.

2.1.2. The retention of paralogs in protein complexes

The functional bias of WGD-resultant paralogs as compared to those resulting from SSD events (discussed in detail above) are thought to be due at least in part to

maintenance of dosage among proteins within a complex or pathway. Sensitivity to altered dosage could potentially explain why genes resulting from the WGD event have unique functional properties when compared to other duplicates[61], specifically, an enrichment for ribosomal genes which are particularly sensitive to imbalance[134] and therefore less amenable to individual duplications. The concept of haploimbalance was originally used to describe the formation of inactive complexes due to an imbalance through either increased or decreased dosage of one member when examining haploinsufficiency among transcription factors[135]. The subsequent dosage balance hypothesis purported that duplicating a sub-component of a complex alters its inherent stoichiometry and is potentially harmful, demonstrating that genes with dosage sensitivity were more than twice as likely to be involved in protein complexes[136] and that many subunit pairs with associated fitness defects are co-expressed[136].

WGD events represent a potential mechanism for the duplication of entire complexes or pathways, and as such relieve constraints of imbalance. For this reason, genome duplications are thought to not only contribute a set of genes that vary in function from other duplication events, but also that add uniquely to genome complexity. Retention of entire functional modules is thought to facilitate morphological gain, with suggestions that this has played a role in plant body plan evolution[137]. Comparable evidence in yeast had been more controversial. Pereira-Leal and Teichmann suggested that since retained duplicates do not frequently exist in known complexes, the contribution of WGD to this mechanism is minimal, rather novel modules emerge in a step-wise fashion[87]. However, this contradicts work by Conant and Wolfe who demonstrated based on expression that there is distinct network partitioning following

genome duplication[64]. Consequently, the role of complex retention on the functional overlap of paralogs (who would inherently be in alternate complexes) is difficult to predict.

2.1.3. Specific rationale and hypothesis

Although previous investigations had found minimal overlap in shared physical interaction partners for gene duplicates, these observations were made on a limited interaction network, and thus may have provided an inaccurate interpretation (see **Table 2-1** and **Figure 2-1**). Recently, two landmark genome-scale proteomic surveys of protein complexes in budding yeast were published[48,49] (herein referred to as Gavin and Krogan, after their respective first authors) describing rigorous mass-spectrometry-based analyses of global collections of purified protein complexes systematically isolated from engineered yeast strains using Tandem-Affinity-Purification (TAP). The TAP method is recognized as being among the most accurate and comprehensive experimental methods for determining PPI and the sub-unit composition of protein complexes[138]. Hence, to more conclusively re-examine the degree of functional divergence between duplicated gene products, I compared the extent of overlap in the physical interactions mediated by paralogous proteins reported in these two datasets.

My focus was on the 450 extant paralog pairs resulting from the ancient WGD event in *S. cerevisiae*[17,25] as these would have had a substantial (~100-200 million years) and equal divergence time, thus providing a basis for horizontal comparisons (excluding partial duplicates and pseudogenes). As a specificity control, I compared these PPI

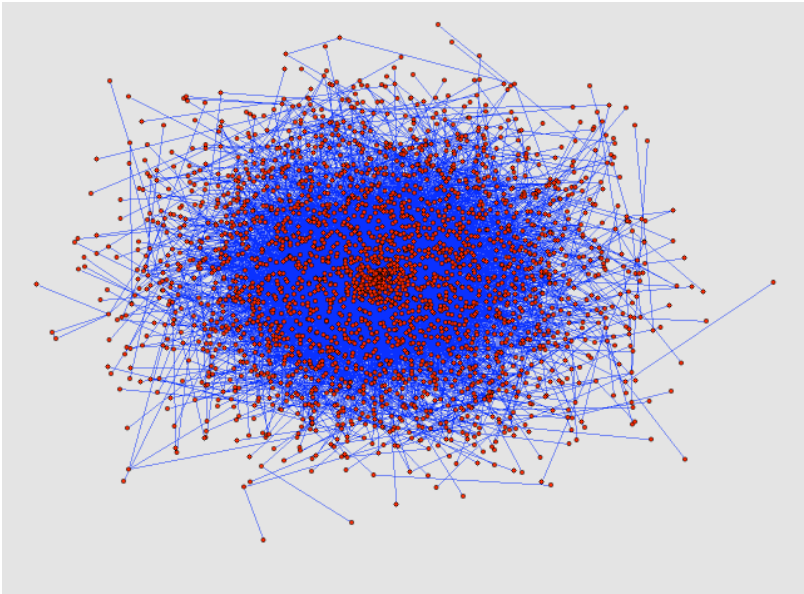
Table 2-1 Size of interaction datasets used

Source	Proteins	Interactions	Average Degree
Krogan *et al*	2708	7123	2.63
Gavin *et al*	1462	6942	4.75
BioGRID	3162	11348	3.59

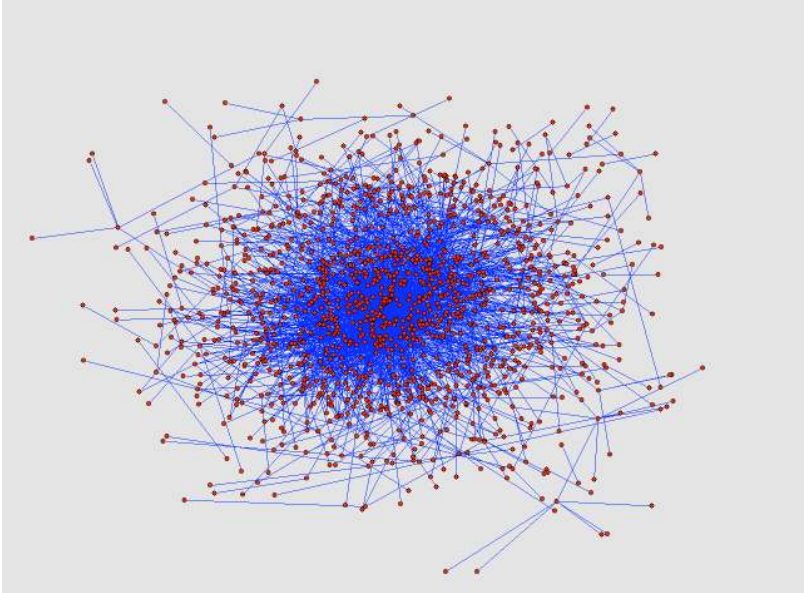
Indicated is the number of physical interactions contained in each dataset used for this analysis. Degree represents the average number of interactions per protein in the dataset. BioGRID data was filtered to remove data resulting from large-scale survey (specifically TAP and Y2H) before use.

Figure 2-1 Network interaction diagrams

A.



B.



Visualizations of the protein interaction data using proteins as nodes and interactions as edges (**A** represents data published by Krogan *et al*, **B** data published by Gavin *et al*).

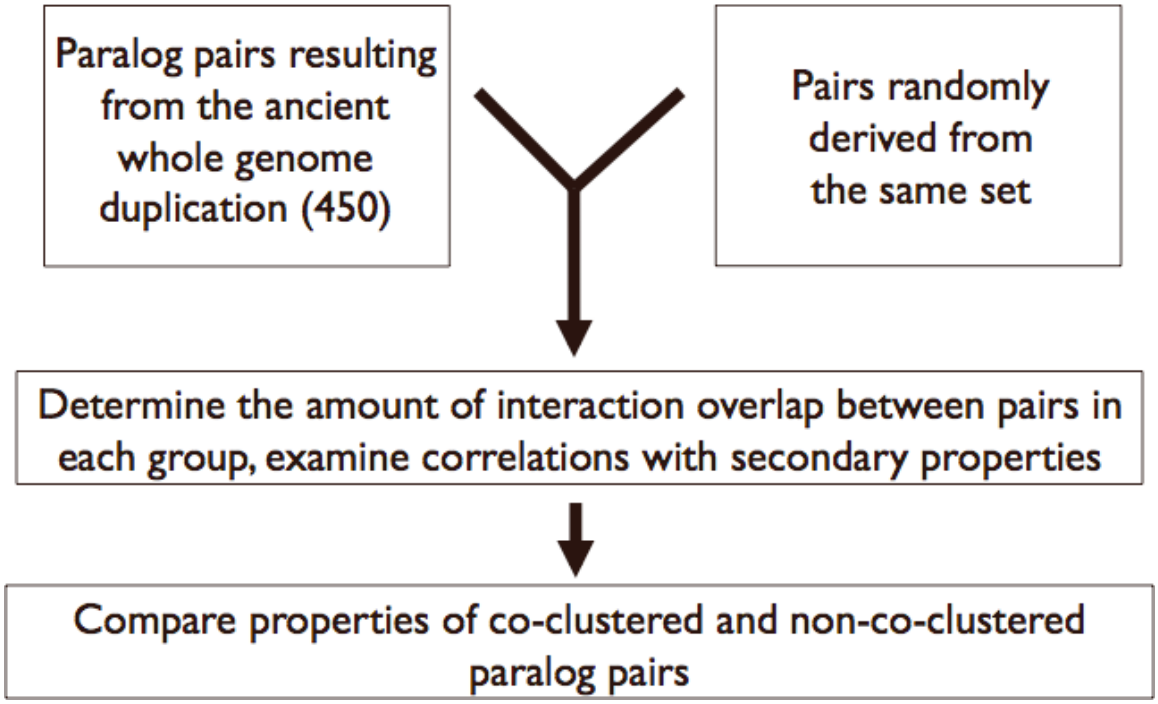
patterns and complex memberships with those derived for randomized pairs selected from the same sub-networks (averaged over 10,000 iterative permutations; see **Figure 2-2**). Given previous observations regarding the advantages of their retained functional overlap (discussed above), and the nature of gene balance hypothesis I felt this analysis would reveal a widespread tendency to share physical interactions and complex membership between WGD-resultant paralogs.

2.2. Methods

2.2.1. Datasets used

All paralog sequences were obtained from the supplementary section of the 2004 publication by Kellis *et al*[17]. In this paper the authors described 457 paralog pairs as resulting from the ancient whole genome duplication event in *S. cerevisiae*. Of these 457 duplicated genes, 450 resulted in the creation of exactly 2 genes; these are the pairs analyzed here. Random sets were drawn from the same set of proteins present in the valid paralog set. The Krogan *et al* TAP interaction data[49] as well as cluster data were generously provided by the Emili and Greenblatt labs at the University of Toronto (http://tap.med.utoronto.ca). This dataset represented the high confidence interactions obtained through the machine-learning algorithm implemented by the authors (7123 unique non-self protein interactions)[49]. Protein clusters containing less than 3 members were discarded from this dataset. The interaction cluster set published by Gavin *et al*[48] was obtained from the supplementary section of their related publication. All protein interactions with a Socio-Affinity Index above 5 were used in this analysis (this was the

Figure 2-2 Experimental workflow



Experimental design for the comparison of true paralog pairs with a random subset to derive an empirical p-value. After randomly shuffling the 450 WGD paralog pairs, true paralogs were compared to random in terms of various properties of interaction, co-expression, and conservation (see **Methods**). Shuffling procedure was repeated 10,000 times.

value indicated by the authors to have high reproducibility), which was 6942 interactions[48]. Clusters used from the Gavin *et al* publication were derived from the core cluster set. The full BioGRID[139] dataset (version 2.0.24) was downloaded from the publicly available website and all genetic interactions, protein interactions resulting from high-throughput experimental protocols (specifically TAP and Y2H), and interactions not found experimentally in *S. cerevisiae* were manually removed. There were 11,348 remaining interactions that were then used in the described analyses (**Table 2-1**). MIPS[104] complex data was also web-downloaded and consisted of 55 protein complexes. Due to the immense size of two MIPS complexes (266 members in the first and 195 in the second compared to the remaining average complex size of 15.04), they were removed from the dataset.

2.2.2. Randomization protocol

All random sets were comprised of 10,000 randomly paired full sets derived from the true paralog list. For example, there were 124 paralog pairs having protein interactions in the Krogan dataset. Each of the comparable 10,000 random sets was then comprised of pairs drawn at random from the resulting 248 paralogs (true pairs were excluded). The reasons for performing the randomization this way were two-fold. First, by shuffling only the respective pairs being examined and not the interactions themselves the network topology is exactly preserved. Second, comparing true paralogs against randomly-derived proteins from the same set maintains any interaction properties pertaining to that specific group. Comparing against outside proteins may have provided

a greater pool from which to draw random protein pairs, but may have also made for less accurate comparisons.

2.2.3. Metrics used for analysis of network properties

Interaction scores

The Non-Shared Interaction (NSI) score was calculated using the method described by de Lichtenberg *et al*[140]. This score is a measure of the number of non-shared interaction partners for each paralog set. Specifically, the formula was:

$$-\log ((N_1+1)(N_2+1))$$

where N_1 represents the non-shared interaction partners of protein 1 and N_2 represents the same for the second protein. The interaction overlap score was calculated as the number of overlapping interactions of two paralogs divided by the total number of unique interactors of the two proteins. It was calculated as:

$$(A \cap B) / (A \cup B)$$

where A and B represent the set of interactors of given proteins.

Gene and protein expression ratio

Gene expression data was based on the datasets compiled by Greenbaum *et al*[141]. These expression levels represent absolute gene expression for each paralog. For each

pair, the expression ratio was described as the greater expression value divided by the lesser. A similar method was used to calculate the protein expression ratios using the expression data collected by Ghaemmaghami *et al*[46].

Evolutionary rate and variation

Non-synonymous substitution rate (Ka) was calculated using Codeml[142] (as part of the EMBOSS suite of programs[143]). For every paralog pair, this value was calculated pairwise (between paralogs) and against the described *K. waltii* ancestor (as described by Kellis *et al*[17]). All sequence identity comparisons were derived from global sequence alignments of amino acid sequences generated using Needle[143] from the EMBOSS package. Paralog pairs that have been evolving at an equal rate should have similar Ka values when compared to their common ancestor. Therefore, the Ka values were plotted for each paralog pair and the evolutionary variation between paralogs was calculated as the Euclidean distance between the corresponding graph point for that pair and the diagonal line. The formula for this was:

$$\text{Distance} = ((K_1 - K_2)^2/2)^{1/2}$$

Where K_1 is the Ka value for the gene 1 and K_2 is the Ka value for the corresponding paralog.

Edge distance and Interaction Graphs

The minimum edge distance was calculated for all paralog pairs using both interaction sets independently. Minimum edge distance represents the minimum number of edges (interactions) required to link two nodes (proteins) on a network graph (for a more detailed overview of graph theory, see **Appendix**). For each interaction dataset all minimum distances were calculated using Pajek[144].

2.2.4. Statistical analysis

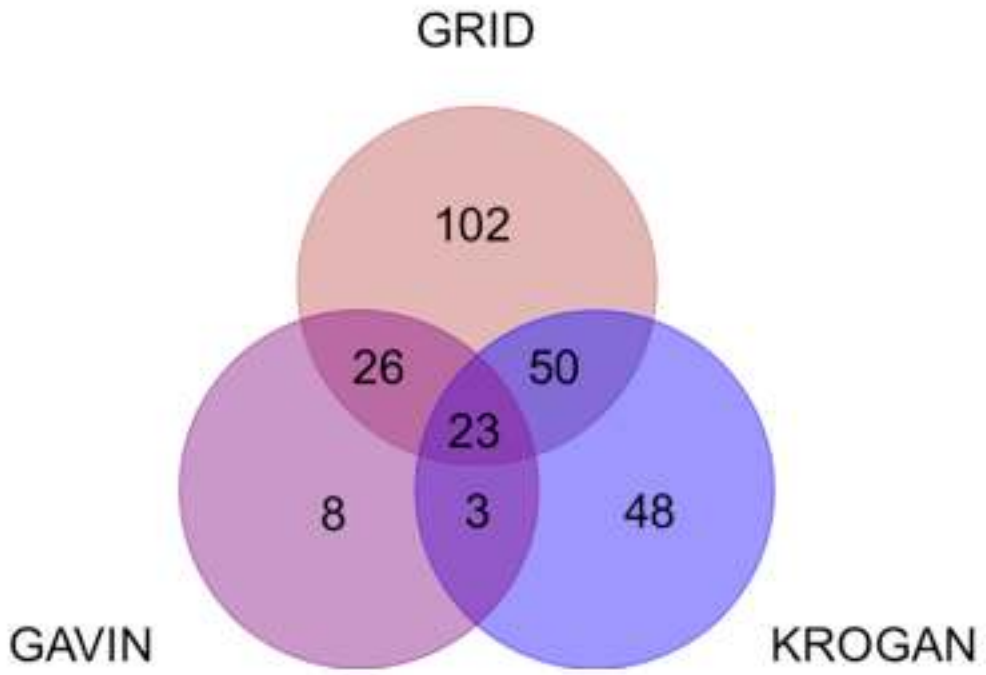
When comparing against random groups, all reported p-values were empirically derived. All reported statistical comparisons between co-clustered and non-co-clustered paralog groups were conducted using 2-tailed student t-tests assuming unequal variance.

2.3. Results

2.3.1. Paralogs have different interaction properties than non-paralogs

High confidence PPI data was available for over one-third (158) of the WGD paralog pairs (124 pairs in Krogan, 60 pairs in Gavin; see **Figure 2-3**). While both Krogan and Gavin employed similar TAP experimental protocols, the subsequent computational assignment of confidence to putative PPI was distinct, resulting in two high-confidence but only moderately overlapping datasets[145]. For this reason, all analyses described herein were performed independently on both the Krogan and Gavin datasets with the logic being that concordant trends are further validated. I applied two complementary metrics to measure the degree of functional overlap: First, a non-shared

Figure 2-3 Overlap in interaction coverage for the three datasets



Venn diagram illustrating the overlap in paralog pairs covered by the three applied datasets (GRID represents all physical interaction data contained in the BioGRID database at the time of analysis). Minimal overlap between the Gavin *et al* and Krogan *et al* datasets was due largely to variations in analysis and clustering techniques.

interaction (NSI) score, which represents the product of the number of non-shared PPI detected between two paralogs (in logarithmic scale, as reported in de Lichtenberg et al[140]); and second, an interaction-overlap (IO) score, which represents the number of shared interactions divided by total number of PPI reported for the paralogs (i.e. intersection divided by union). As shown in **Table 2-2**, I found that the true paralog pairs were far more likely to (i) share interaction partners, (ii) interact with each other, and (iii) to lack non-shared interactions, as compared to randomized datasets ($p<0.001$ in all cases for both Krogan and Gavin, as compared to random). Likewise, comparative analysis of the complete PPI networks (see **Methods**), with proteins represented as nodes and the interactions as edges, indicated a much closer average functional similarity among the extant paralog pairs (**Table 2-2**). True paralogs had a shorter average minimum edge distance, representing the number of PPI separating any two nodes, as compared to the shuffled pairs ($p<0.001$ Krogan, $p=0.002$ Gavin). This demonstrates that duplicated gene products generated by the ancient WGD are closely linked in the physical networks, and hence more likely to be functionally associated.

2.3.2. Validation of findings

To eliminate any potential ambiguity (i.e. mis-assignments) in the original proteomic datasets caused by highly similar protein sequences among the paralogs (see **Methods**), I re-analyzed the interaction data after excluding the 80 of 457 paralog pairs that had sequence identities greater than 80%. All mass spectrometry experiments suffer from the failing of not being able to distinguish between closely related sequences.

Table 2-2 Elevated interaction overlap for true paralog pairs

A.

Maximum Similarity (%)	Group	Number of pairs in set	Number of pairs with shared interaction(s)	Number of pairs that interact with each other	Interaction overlap score (IO)	Non-shared interaction score (NSI)	Minimum edge distance	Frequency of co-clustering (%)
80	Random Pairs	367	<1	<1	0	-1.09	5.63	<1
80	Paralogs	367	28	24	0.1*	-1*	3.14*	41
100	Random Pairs	447	<1	<1	0	-1.19	5.9	<1
100	Paralogs	447	42	48	0.11*	-1.09*	2.66*	56

B.

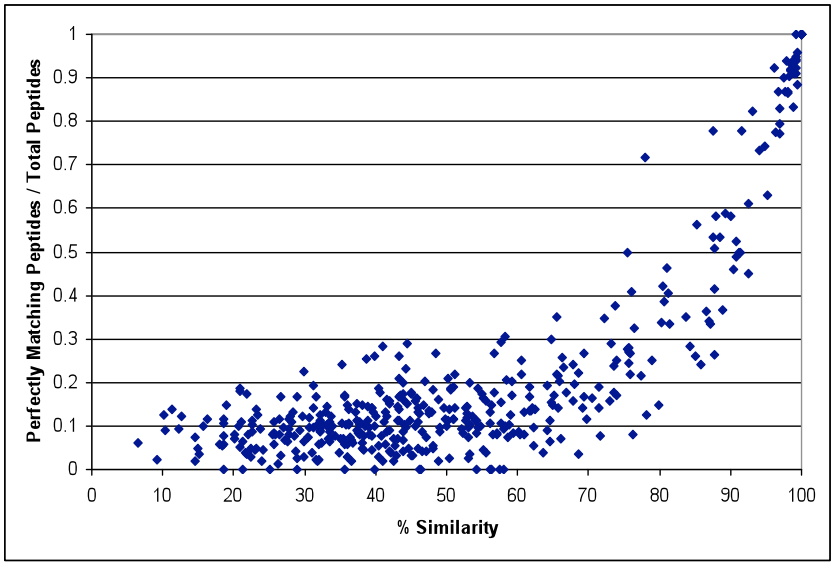
Maximum Similarity (%)	Group	Number of pairs in set	Number of pairs with shared interaction(s)	Number of pairs that interact with each other	Interaction overlap score (IO)	Non-shared interaction score (NSI)	Minimum edge distance	Frequency of co-clustering (%)
80	Random Pairs	367	<1	<1	0	-1.42	5.51	<1
80	Paralogs	367	13	10	0.23*	-1.06*	1.52*	50
100	Random Pairs	447	<1	<1	0	-1.44	5.67	<1
100	Paralogs	447	26	14	0.13*	-1.25*	3.19*	41

Interaction overlap among true paralog pairs (all pairs, or only those pairs with less than 80% amino-acid similarity) as compared to randomized pairs drawn from the same datasets (averaged over 1000 iterations). While Kellis and colleagues identified 457 putative paralog pairs as resulting from the WGD, 7 of these pairs had paralogs divided across multiple open reading frames and 3 contained proteins with no confirmed amino-acid sequence. Therefore these 10 pairs were removed from our analysis. Panel **A** represents comparisons made using the interaction data from Krogan dataset[49], while Panel **B** represents data drawn from Gavin[48]. All scores (IO, NSI and minimum edge distance) were compared statistically (see text); significance is indicated with an asterisk. Analyses of both datasets show a significant difference in the interaction properties of valid paralogs, indicating an elevated level of functional similarity.

During mass spectrometry, the protein of interest is digested via trypsin, and the sequences of the resulting peptides are mapped to proteins using a mapping algorithm such as SEQUEST[146]. As these algorithms tend to take the first FASTA sequence to provide a perfect match with the sequence of interest, it is conceivable that SEQUEST mistakenly returns the paralog of the true protein instead of the true protein itself. To verify this, I ran a mock trypsin digestion of all 900 members of the paralog set and determined the fraction of peptides that could also match to the corresponding paralog. When plotting this fraction against percent amino acid sequence identity, I found an exponential curve, drastically increasing above 80% similarity (see **Figure 2-4**). Therefore, analyses were re-performed only using proteins below 80% sequence similarity. Again, extensive PPI overlap was observed among the remaining paralog pairs ($p<0.01$ for all comparisons). The fact that even after removing the most conserved duplicated gene pairs still revealed substantive evidence of functional overlap among the more divergent paralogs demonstrates the robust prevalence of this propensity, and indicates that functional relatedness is not a trait exclusive to paralogs under the most rigid evolutionary constraint or due to a possible confounding effect of near identical duplicates or close isoforms.

Next, as both the Krogan and Gavin interaction datasets were generated using TAP, I sought to determine whether the apparent PPI conservation was influenced by artifacts specific to this particular experimental procedure. Hence, I obtained the latest literature-curated interactions from the BioGRID[139] web resource (version 2.0.24), removing all PPI resulting from high-throughput experimentation, reasoning that the

Figure 2-4 Mass spectrometry detection of duplicates



The peptides that would result from Trypsin digestion were matched against the corresponding paralog. The ratio of peptides that match exactly for each paralog pair versus sequence identity is indicated and shows a clear increase above 80% sequence similarity.

remaining interactions, while not necessarily more accurate, would not be subject to the same bias as for TAP. With over 200 paralog pairs having curated PPI in the filtered BioGRID dataset, the trend was not only mirrored, but further pronounced across all statistical tests performed ($p<0.001$).

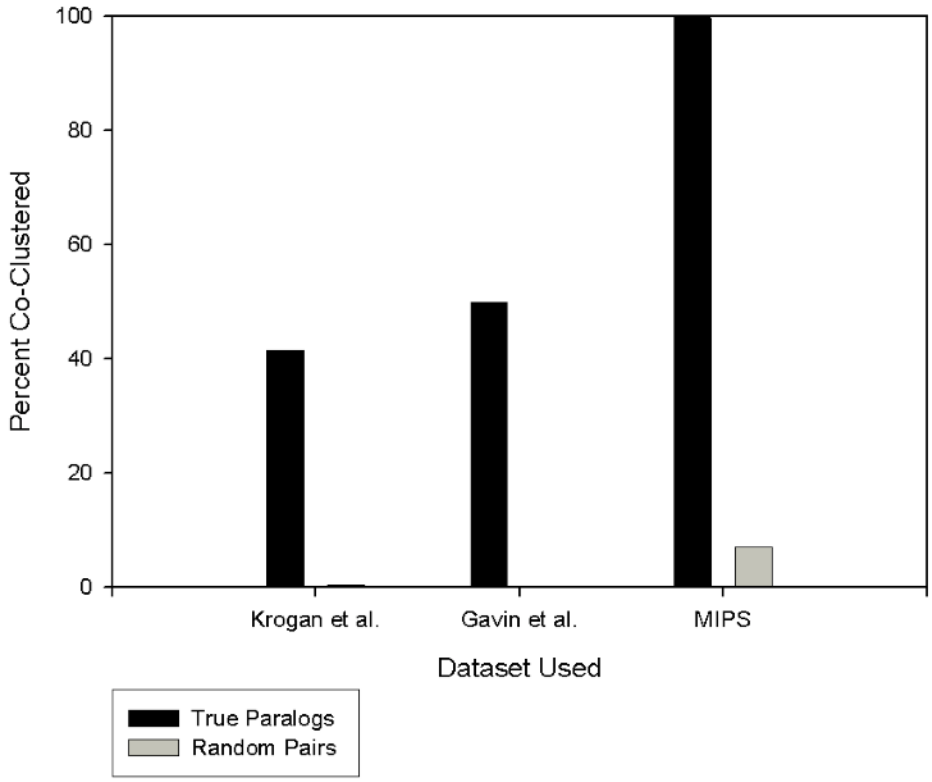
I also re-performed the analyses after randomly removing one half of the PPI from each of the three datasets (Krogan, Gavin and BioGRID). Again, the true paralog pairs showed significantly more PPI overlap than the comparable randomized pairs across all performed tests, indicating resilience to incomplete, inaccurate or missing PPI.

2.3.3. Paralogs more frequently co-complexed

Using the protein complexes reported by Krogan and Gavin[48,49] as an alternative, more rigorous gauge of functional relatedness, I observed that the WGD paralogs were far more likely to be associated within the same complex as compared to random pairs. Among all cases where both sister paralogs had been confidently assigned to a given complex, nearly half (40-56% depending on the dataset and amino acid similarity cut-off) were assigned to the same protein complex (**Figure 2-5**). In contrast, less than 1% of the shuffled paralogs were typically assigned to the same complex ($p<0.001$ for both Krogan and Gavin).

Co-clustered paralog pairs were shown to have slightly more protein interactions on average than no-co-clustered paralogs (5.98 and 4.20 respectively), however these proteins were also shown to exist in smaller interaction clusters (average of 7.3 members versus 10.03 members in no-co-clustered), indicating that co-clustering was not an

Figure 2-5 Frequent co-clustering of paralog pairs



Co-assignment of pairs of duplicated gene products together in the same putative protein complex (interaction clusters) in both the Krogan and Gavin predicted interaction clusters as compared with randomized pairs. The figure shows the percent of true paralogs wherein both members were assigned to the same complex (co-clustered) at varying amino-acid similarity cut-offs as compared to control pairs (averaged over 10000 iterations). In all cases, the true paralog pairs exhibit a statistically significant level of overlap.

artifact of within-cluster protein interactions and thus potentially influenced by the experimental method. Also, as there has been evidence presented showing hubs to have distinct evolutionary properties from non-hub proteins[147], I sought to determine if the presence of hubs may be biasing my noted trend of increased conservation in co-clustered proteins. Assuming hubs to be in the top 10% in terms of connectivity (a liberal definition when compared to the approximately 7% initially used by Jeong et al[148]), this classifies all proteins with 13 or more interactions (286) proteins within the Krogan interaction dataset as being hubs. As there are 385 paralogs with interactions in the Krogan set, one would expect 38 (approximately 10%) to be represented as hubs by this definition, when in fact only 27 are, making WGD paralogs less likely than average to be hub proteins. To extend one step further, there are 11 hub paralogs studied in my pair-wise analysis (included because both sister paralogs had interactions in the dataset), 7 of which were deemed to have co-clustered with their partner. Upon removal of these 11 paralogs and their respective partners, co-clustered paralogs were still significantly more conserved than non-co-clustered ($p<0.001$), indicating that conservation is not a result of the inclusion of hub proteins (it is worth noting here that there were no studied paralog pairs in the Gavin *et al* dataset containing hub proteins).

As a final stringent benchmark, I re-performed my analysis using a public "*gold-standard*" reference consisting of the well-established manually curated MIPS protein complex database[104]. Even more prominently, all of the true paralog pairs were found to be co-clustered in the same complex as compared to only ~7% in the randomized group (n.b. the smaller number of MIPS complexes resulted in an inflation in the proportion of co-clustered pairs within the random set), confirming the validity of my major finding

that the products of gene duplicates produced by WGD have an elevated propensity to be functionally linked.

2.3.4. Co-clustered paralogs are under tighter evolutionary constraint

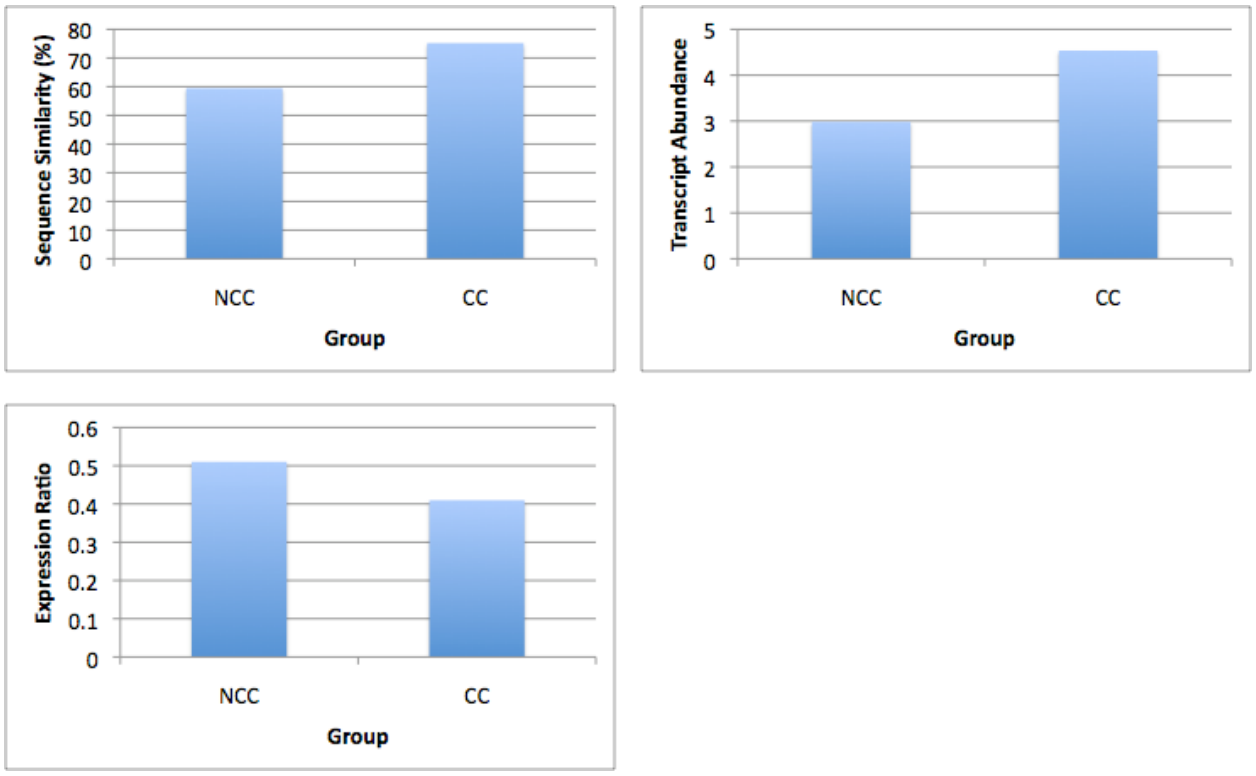
In an effort to better understand the evolutionary adaptations resulting in the physical complexing of duplicated gene products, I next asked whether the two classes of paralog pairs, i.e. those that are co-clustered in same complex (herein referred to as CC) versus the non-co-clustered (grouped into different complexes and referred to as NCC), exhibited other commonly studied functional properties in common. The properties investigated were similar in nature to those previously applied to examine functional overlap in paralog pairs[17,53,85-87]. These included: (i) the level of amino acid sequence similarity (inferred from sequence alignment), (ii) concordance in combined gene/protein expression levels (as assessed by Greenbaum *et al*[141]), (iii) the correlation of expression patterns, (iv) the codon-adaptation index[149] (CAI), (v) the non-synonymous substitution rate (Ka) which served to gauge the conservation rate as calculated against predicted orthologs in *K. waltii*[17] and, lastly, (vi) the Euclidean distance between the Ka values of the paralog pairs was calculated between to assess differences in selective pressure (see **Methods** for full description of the methods used). Due to the small number of paralog pairs in the Gavin dataset, I restricted all subsequent comparisons to the Krogan dataset (123 clustered pairs, see **Methods**).

As a group, the CC paralog pairs had a higher average sequence similarity ($p<0.001$) than the NCC paralogs. Likewise, when comparing Ka, which quantifies the evolutionary rate against an ancestral out-group, the CC paralogs were found to be far

more conserved ($p<0.001$) than their NCC counterparts (see **Figure 2-6**). Intriguingly, however, these two classes showed no difference in the overall distributions of Euclidean divergence distances among the respective paralog Ka values. As this value indicates the relative symmetries of divergence between sister paralogs (or, in other words, the degree of concerted evolution), this finding implies that while the sister paralogs associated in the same protein complex are generally more highly conserved than those not, the extent of bilateral conservation is generally similar to that observed for the NCC paralogs.

As a relationship between evolutionary conservation and gene expression has previously been established for WGD paralogs in yeast[150], I next sought to examine the co-expression properties of CC and NCC paralogs. As might be expected, I found that both the CAI and expression levels (as determined through combination of various gene and protein expression datasets[141]) of the CC paralogs were significantly higher than NCC ($p<0.01$; **Figure 2-6**). Surprisingly, however, there was no difference in the correlation of expression patterns among the CC and NCC sister paralogs; indeed, comparison of the expression ratios (see **Methods**) showed a trend ($p<0.06$) towards a greater difference in expression abundance between pairs of CC paralogs. This trend was further exaggerated ($p<0.05$) when examining only the protein abundance data reported by Ghaemmaghami *et al*[46]. These observations support a previously proposed mechanism[92] (see comment by Koonin[151]) whereby one copy of a duplicated gene maintains a minimally supportive role, as evidenced by lower relative expression under normal conditions, but which can be up-regulated as needed to support cell viability under perturbed growth conditions.

Figure 2-6 Differential properties of CC and NCC paralogs



Those paralog pairs existing in different annotated complexes (non-co-clustered, or NCC) showed properties varying from those pairs existing in the same complex (co-complexed, or CC). The NCC paralogs generally had lower sequence similarity and lower abundance, however the CC paralogs showed a lower expression ratio (see **Methods**), suggesting that they are actively maintained at varying transcript levels.

One obvious question that arises from these new results is whether the distinctive properties of CC paralogs are a consequence of their physical association or rather whether the increased conservation and co-expression predisposes them to the retention of PPI. During my investigations, I found no overall correlations between various measures of similarity, including sequence identity, conservation, CAI, gene expression correlation, and any previously mentioned interaction quantification score (such as the number of shared interactions, IO, NSI, or minimum edge distance) among the WGD paralog pairs reported in the Krogan, Gavin and BioGRID datasets. This implies that similarity in such former parameters does not necessarily imply that paralogs are likely to share the same interaction partners and that co-membership in a protein complex is not merely a consequence of elevated sequence conservation. This observation has also subsequently been noted by Guan *et al* using an interaction dataset assembled from multiple sources through machine-learning[54]. It is, however, presently impossible to determine whether the physical associations of sister paralogs (i.e. co-clustering) causes increased conservation and gene expression or vice versa.

2.4. Discussion

Although PPI information was available for only about 1/3 of all surveyed paralog pairs produced by WGD, the results presented here indicate that after at least 100 million years of evolution, a substantial number (about half) of extant paralogs retain substantial functional conservation as evidenced by co-clustering into the same interaction complexes. While different lines of evidence suggests that the WGD paralogs are under different initial evolutionary constraint than singleton duplicates[3], clouding extrapolations

towards paralogs of non WGD-origin, my analysis represents one of the largest comparisons of the functional behavior of paralogs in a eukaryotic model setting.

Increasing, albeit somewhat speculative evidence for the so-called 'Duplication-Degeneration-Complementation' (DDC) models of sub-functionalization has emerged over the past few years[73,74]. Although my data do not rule out neo-functionalization as a means of initial functional divergence (since it is at present impossible to calibrate the rate of functional divergence without knowledge of PPI for orthologs in the ancestor), the surprisingly high degree of extensive functional overlap (co-clustering) among paralog seems more compatible with the DDC hypotheses. Consequently, my results contrast with previous reports that species with large population sizes are unlikely to experience substantial sub-functionalization of gene duplicates[73,152], highlighting the need for alternate measures to accurately determine patterns of functional divergence.

While the algorithmically-derived complexes identified in the Krogan and Gavin datasets are not macromolecular associations in the traditional sense, data clearly demonstrate that duplicated genes frequently retain close association. This seems to contradict the balance hypothesis in that duplicates to not appear to remain sequestered in independent functional modules. Therefore while maintained balance[135] may have mediated their initial retention, subsequent gene loss coupled perhaps with a drive for maintained functional overlap has forced their co-retention inside a pathway or complex.

Chapter 3

Extensive, condition-dependent epistasis among WGD paralogs

Portions of this chapter have been reprinted or adapted from *Musso *et al*[123]

Corey Nislow, Andrew Emili, Zhaolei Zhang, Charles Boone, and Michael Costanzo supervised and/or advised the experimentation and I performed all relevant analysis. I also performed all of the described experimentation with the exception of microarray analysis (I was assisted in this by Andrew Smith and Larry Heisler), tetrad dissections (performed by Manqin HuangFu under my supervision) and growth curve analysis (assisted by Bryan Joseph San Louis and Jadine Paw).

3.1. Introduction

As mentioned briefly above, Harrison *et al* used FBA to predict functional compensation among yeast genes, ultimately confirming 12 of 17 (~60%) predicted cases[153]. Further, 3 of the 5 non-interacting pairs were shown to be epistatic upon deletion of a third, additional gene. This represented the first systematic test of functional compensation between genes in yeast, and importantly the first evidence that detection of genetic interactions among metabolic genes was dependent on environmental condition[153]. The capacity for duplicates to retain redundancy so as to facilitate environmental adaptability was initially suggested by Papp *et al* using FBA[85] and is akin to a finding of Gu and colleagues[53] when investigating single gene deletions (see above). Gu *et al* would importantly suggest that examination of fitness under standard media conditions may bias observations for duplicated genes as: "*it is possible that when a gene deletion showed no effect in any of these conditions it was not due to compensation by other genes but was because the gene deleted was not related to the growth conditions used*"[53]. These authors also divided their systematic results by functional category, noting that effects varied based on the nature of the genes identified[91]. Ultimately the extent of epistasis among both metabolic and non-metabolic paralogs remained largely unknown.

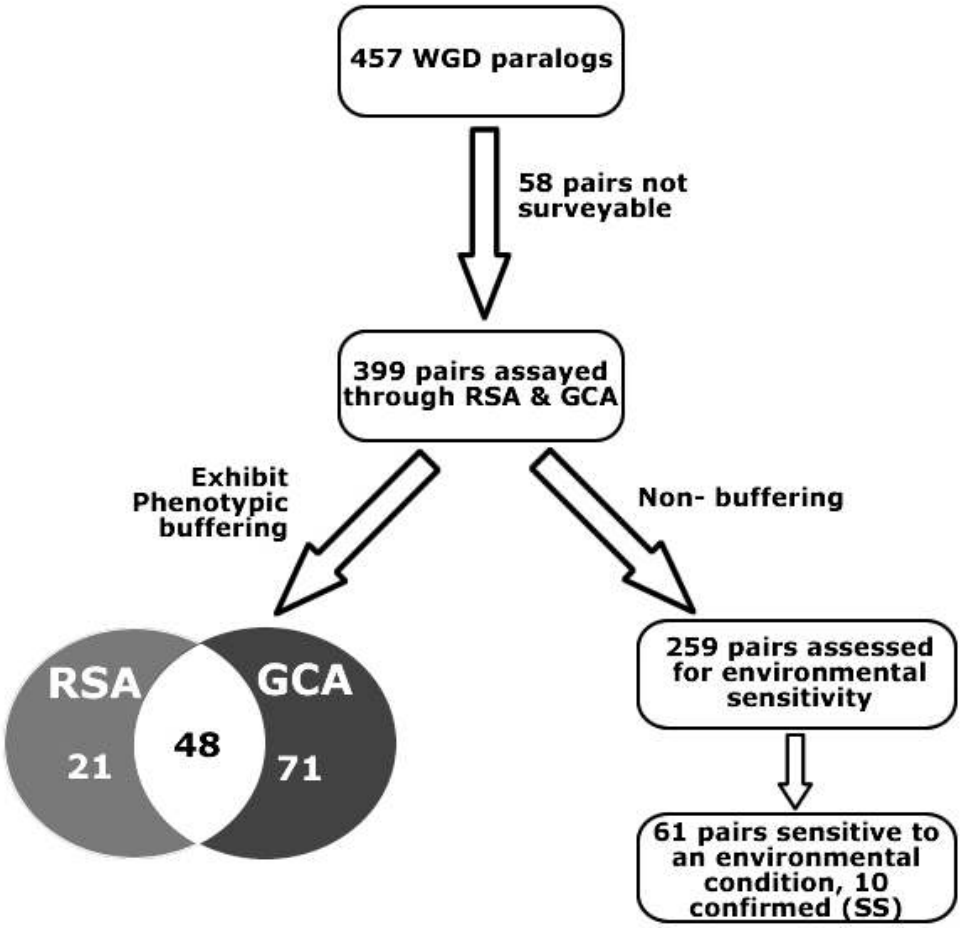
3.1.1. Specific Rationale / Hypothesis

Previous examinations of the biological properties of extant paralogs based on the physical interactome[54,98-100], metabolic networks[85,86], and single-gene deletion phenotypes[53] in the budding yeast *S. cerevisiaie* have implied extensive functional

similarity among both WGD and SSD resultant paralogs, supporting an advantage to retaining substantive functional overlap. Conversely, analysis of synthetic genetic interactions of a small subset of yeast WGD and SSD paralogs has lead an alternate hypothesis that while some paralog pairs have maintained the ability to buffer loss of a respective sister, this mechanism is limited in scope and does not function over a wide range of compromising environmental conditions[154]. This assertion contrasted with previous suggestions that duplicates may be preferentially retained to compensate for cellular stresses or perturbations[53,99]. Consequently, the extent and context of functional buffering among WGD-resultant duplicates as well as the molecular properties of buffering paralogs remain to be resolved.

To address these issues directly, I examined the relative fitness of yeast strains bearing single and double deletions of all surveyable WGD-resultant paralog pairs in yeast. Further, as I had previously asserted that any potential "*transcriptional back-up*" mechanisms might be retained specifically to deal with stress or pertubation, I (with assistance from members of the Nislow & Gaiever labs) examined interactions not only in normal growth media, but in several alternate conditions designed to mimic common cellular stressors (see **Figure 3-1**). Since evidence suggests not only a widespread functional overlap among extant paralogs but also mechanisms of transcriptional compensation, I expect to observe extensive genetic interactions between WGD paralogs in yeast.

Figure 3-1 Experimental outline



The 399 WGD paralog pairs with viable constituent single-deletion strains[52] were analyzed for genetic interaction using Random-Spore-Analysis (RSA) and Growth-Curve-Analysis (GCA). The overlap in terms of duplicates exhibiting significant phenotypic buffering is shown in inset Venn diagram. The 259 WGD paralog pairs not exhibiting epistasis (and hence deemed to not buffer phenotypically) in rich growth media were further analyzed through Environmental Screening in five alternate media conditions designed to either mimic common cellular stress states or introduce competition[52].

3.2. Methods

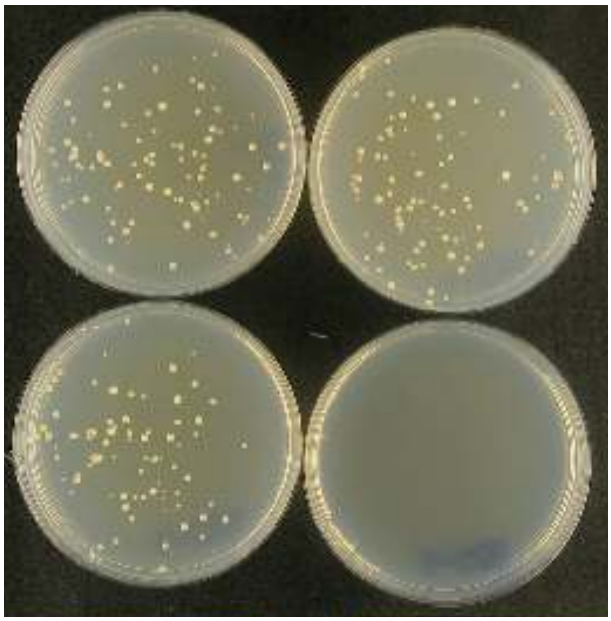
3.2.1. Assessment of synthetic lethality

Method for screening of synthetic lethality through Random Spore Analysis (RSA) was based on that previously used[119,120]. Briefly, yeast strains carrying a deletion of a given paralog were mated with strains harboring deletion of the corresponding paralog (with deletion confirmed through two resistance markers; *Kanamycin* and *Nourseothricin*) and of opposite mating type. Following mating, cells were sporulated at 22°C for five days. Individual colonies were then diluted and grown on appropriate selective media (SD – Arg – His + Canavanine + G418/NAT; see **Figure 3-2**) for 2-3 days at 30°C, photographed, and classified by two independent observers as being either synthetic lethal, synthetic sick, having no interaction, being unclear, or being abnormal (abnormal indicates that a single deletion or control strain was inviable). Any strain determined by both observers to be either synthetic lethal or synthetic sick was classified as such. Strains classified by one observer as synthetic lethal or synthetic sick but by the other as unclear were further investigated by tetrad analysis (see **Figure 3-2**). Lastly, all abnormal crosses (either one deletion strain or control strain inviable) were re-analyzed.

For Growth Curve Analysis (GCA), MAT**a** spore progeny obtained as described for RSA were grown in 96 well format in media selecting for double mutants (YPD + Canavanine + G418 + NAT). Growth rate was monitored for 5 generations using a TECAN reader and corresponding curves analyzed visually. The resulting growth-curves deemed to either: have a lower point of saturation, have an obvious growth lag, or have a decreased slope in exponential growth phase when compared to plate-specific controls

Figure 3-2 Random Spore Analysis

A.



B.

Double mutant colonies with obvious lethal (a) or sick (b) phenotypes scored as genetic interactions. In both (a) and (b) top left colony represents cells with no mutation, top right represents mutation of one paralog, bottom left deletion of the corresponding sister paralog, and bottom right the strain carrying the dual-deletion. Media used was SD – Arg – His containing (clockwise from top left): Canavanine, Canavanine + G418, Canavanine + G418 + NAT, Canavanine + NAT.

were labeled potential interactors. Next, growth rates of the constitutive single-mutant deletion strains of potential interactors were assayed in duplicate, and area under growth curve calculated after 20 hours growth (see **Figure 3-3**). Genetic interaction between gene pairs was assessed as that beyond the predictions of the multiplicative model (analogous to[115]):

Let W_x equal the fitness of mutant strain x (as compared to plate-specific control), W_y the fitness of the strain carrying the deletion for the corresponding sister, and W_{xy} the fitness of the dual-deletion strain. A genetic interaction then is characterized as:

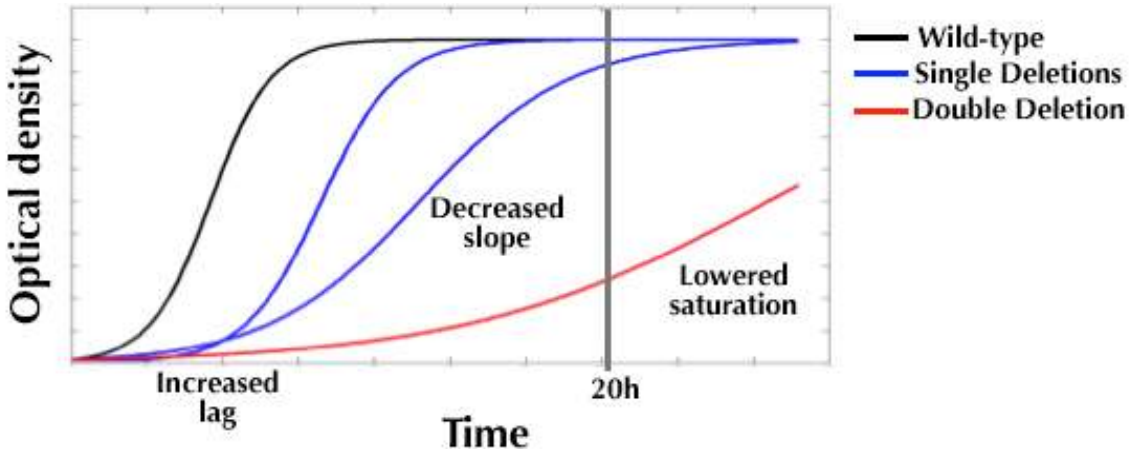
$$W_{xy} < (W_x * W_y) - (\sigma_x + \sigma_y)$$

Where σ_k represents the standard deviation of deletion strain k measured over replicates.

3.2.2. Screening for phenotypic rescue

To assay for phenotypic rescue among genes essential for cell viability, strains containing a temperature-sensitive (*ts*) allele of the gene of interest were used where available and transformations performed using both the corresponding paralog and the gene itself on both low (MoBY[155]) and high copy (2μ) plasmids using a standard Lithium Acetate transformation procedure. Following appropriate selection at room temperature growth for *ts* strains was assessed using spot dilution.

Figure 3-3 Growth Curve Analysis



After 20 hours of growth the area of the curve is calculated and compared to wild-type strains (black). Use of the area method ensures that the three common phenotypes indicative of delayed growth (increased lag from stationary phase, decreased slope during exponential growth, and lowered point of colony saturation) are captured. A genetic interaction would be observed if the growth defect associated with double-mutant (red) compared to constitutive single-deletions (blue) surpassed expectations of the multiplicative model (see **Methods**).

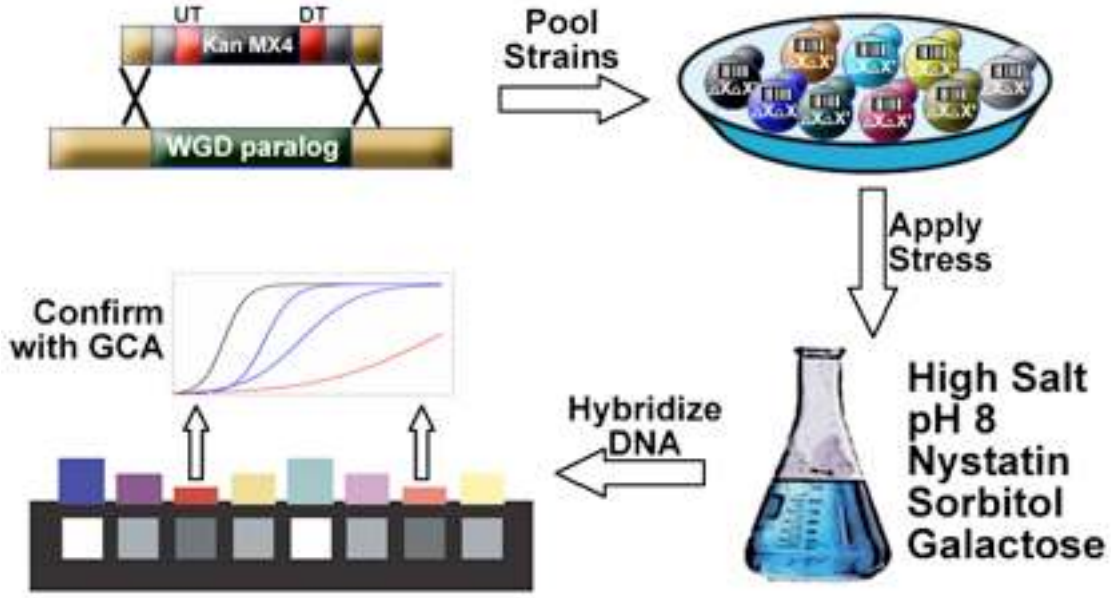
3.2.3. Information gain analysis to indicate features predictive of epistasis

In order to determine properties generally indicative of epistasis for paralog pairs a large list of pertinent features was compiled and analyzed for predictive value. These features included: the number of additional duplicates (see *Chapter 2* above), sequence similarity (Ka, local sequence alignment percent identity, local sequence alignment percent similarity), physical interaction degree and overlap (using data from BioGRID[156], Krogan *et al*[49], Gavin *et al*[48], and Batada *et al*[132]), difference in transcript and protein expression magnitude (as calculated in *Chapter 2* above) and co-expression across multiple conditions (as published by Kafri *et al*[92], and Tirosh & Barkai[157]). Information gain of each of these properties with epistasis for paralog pairs was calculated using the Weka software environment[158] both for all paralog pairs, and for paralog pairs divided by GO SLIM functional category[105].

3.2.4. Environmental screening and barcode analysis

Combined lethality of double-mutant haploid yeast strains grown as described above and combined into a pool of 499 members (399 WGD paralog pairs, 100 double-mutant control strains; slow-growers supplemented to obtain equal starting concentration of each double-mutant strain) was examined in 5 media conditions previously designed[52] to either introduce competition or mimic common stress states (1M NaCl, 1.5M Sorbitol, YPGal (where 2% galactose was substituted for dextrose), 10μM Nystatin, and YPD pH 8; see **Figure 3-4**). Equalized concentrations of cells were grown in respective media conditions for 5 generations, at which time cells were lysed using a Qiagen DNeasy kit

Figure 3-4 Assaying genetic interaction in multiple conditions



The presence of molecular barcodes was used to facilitate screening of genetic interactions in multiple conditions. In addition to *Kanamycin* resistance, each strain in the deletion collection contains two unique nucleotide sequences (denoted as UT and DT above). This allows the strains to be pooled in liquid culture and analyzed; facilitating the determination of condition-specific strain sensitivity. Once a double-deletion strain was determined to be sensitive to a give condition, growth of that strain as well as the constitutive single-deletion strains were analyzed in that condition and epistasis determined.

and DNA was isolated. Both UPTAG and DNTAG barcodes were amplified via PCR, and hybridized to high-density oligonucleotide Affymetrix (as described previously[52], except with use of Tag4 arrays[159]).

3.2.5. Determination of condition-specific synthetic lethality

For each of the five media conditions tested, intensity values resulting from hybridization (described above) were Lowess normalized using ~1000 barcoded strains which had been independently grown in YPD and additionally hybridized to each chip used (done to remove potential spatial bias). Abundance data for WGD paralogs was then normalized both by row and column. Row values (dual-deletion strain abundance) for the five experimental conditions were normalized using each strain's hybridization in control YPD. Columns (conditions) were normalized using the average abundance value of hybridizing non-WGD double-mutant deletion strains grown in each condition. Normalization was performed independently for both experimental runs and for UPTAG and DNTAG expression. Data were then combined, and those strains with below 75% normalized abundance in a given condition were treated as potential interactors.

A subset of the most consistently affected potential interactor strains was then selected to be further analyzed through GCA. To determine this subset, all strains in the initial pool of 499 (399 paralog double-deletion strains + 100 internal controls) with abundance two-fold beyond background in control YPD (background determined as the average abundance value of non-existing strains) were given an incremental rank in each condition according to their abundance magnitude. For each strain, the difference between rank within the control state and in any given media state was calculated

(changes in rank for entire cell population were found to be roughly normally distributed around 0 for each condition). Those strains indicated as above to be potential interactors and with changing rank beyond one standard deviation in both experimental runs (in both UPTAG and DNTAG expression) were selected. Corresponding single-mutant and double-mutant strains were then grown and analyzed similar to described above for GCA, however in this instance, GCA was performed entirely in the given media condition. Those paralog pairs passing the multiplicative model (see above) were confirmed as being sensitive to the given condition.

3.2.6. Analysis of physical interactions

Physical interactions of genetic interactors and non-interactors were compared as above using two published tandem affinity purification interaction datasets[48,49], and the data compiled in BioGRID[139] following removal of all data generated using high-throughput assay (performed manually). Lists of protein complexes were also obtained from Krogan *et al* and Gavin *et al* publications, as well as the MIPS database[104]. Briefly, genetic interactors were compared against non-interactors in 5 categories: number of shared interactions, interaction overlap score (number of shared interactions per unshared interactions), non-shared interaction score (previously described as the negative logarithm of the fraction of non-shared interactions[140]), propensity to interact with each other, propensity to co-cluster (the latter two analyzed by Fisher exact test, all else analyzed using Mann-Whitney rank sum test).

3.2.7. Comparisons of conservation and expression

Sequence identity among paralogs was evaluated through application of a global alignment algorithm (Needle as part of the EMBOSS suite of programs[143]) calculated on amino-acid sequences. Non-synonymous substitution rate (Ka) was calculated using Codeml[142] (again through EMBOSS[143]) against the appropriate *K. waltii* ancestor[17]. While the ratio of non-synonymous to synonymous substitution rate is the preferred metric when assessing evolutionary conservation, the synonymous substitution rates were saturated and therefore un-usable. Protein[46] and mRNA[141] expression data were obtained as published. To compare temporal expression patterns of paralog pairs, expression data compiled at various time points throughout the cell cycle[44]. This data was first normalized based on the logarithm of each gene's median intensity value, then the Pearson Correlation Co-efficient for every paralog pair over the first 8 time points was determined (internal analysis revealed that all points correlated poorly beyond the first 8 time-points, data not shown).

3.2.8. Determination of instances of multiple paralogy

Additional (i.e. non-WGD-resultant) cases of paralogy were determined for 449 of the initial set of 457 WGD paralog pairs (the 7 pairs initially described[17] as being split into multiple open reading frames as well as one pair containing a categorized pseudogene were excluded), similar to the method previously described[160]. Briefly, protein sequences corresponding to every known cDNA sequence in *S. cerevisiae* (excluding hypothetical or dubious open reading frames, 5880 total) were downloaded from the SGD database (www.yeastgenome.org) and BLAST analysis was performed

aligning each of the 898 individual WGD paralog genes against all *S. cerevisiae* protein sequences with an expectation (e-value) cutoff of 0.1. From there, resulting alignments in which the aligned region covered at least 50% of the sequence of the larger protein were retained (50% used as the cutoff for the aligned region as opposed to the more common value of 80% in order to identify a greater number of paralogs (as previously described[53]). As a final criterion, alignments were required to meet a threshold of percent sequence identity in order to be retained[53]. For alignments where the aligned region was greater than 150aa in length, a minimum sequence identity of 30% was required. Based on previous empirical evidence indicating that a more stringent sequence identity cutoff is needed for smaller proteins[161], for all alignments shorter than 150aa, sequence identity was required to surpass the value determined by a pre-determined formula[161]:

$$n + 480 * L^{-0.32 \times (1 + e\text{-}L/1000)}$$

Where $n = 6$ (as previously established[53]) and L represents the length of the aligned region.

3.2.9. Statistical Analysis

All statistical calculations and correlations were performed using SigmaStat 3.1. Heatmap was drawn using MATLAB. Representation of statistically enriched GO terms drawn using Cytoscape[162].

3.3. Results

3.3.1. Frequent phenotypic buffering between WGD-resultant duplicates

I used two complementary experimental growth assays to systematically monitor the fitness of single and double mutants to determine the extent of phenotypic buffering among putative yeast WGD paralog pairs[17] under standard culture conditions. I was unable to assess 7 pairs because one or both paralogs was split into multiple open reading frames[17], while 51 pairs were excluded from analysis due to the inviability of one or both of the single-deletion strains (these strains were investigated using temperature-sensitive alleles, see below) leaving 399 surveyable pairs out of the initial set of 457.

Random-Spore Analysis (RSA) was first applied to measure the overall viability of the progeny of genetic crosses between individual single gene deletion strains. Haploid yeast strains containing deletions corresponding to either one or both paralogs of a WGD pair were grown on solid minimal media and selected for based on specific drug sensitivities (deleted genes were replaced by drug resistance cassettes, see **Methods**). Visual inspection conducted by two independent evaluators was ultimately used to define 51 obvious cases of synthetic sickness or lethality (see **Table 3-1**). Tetrad dissection additionally confirmed 18 of the 31 pairs initially deemed non-obvious by either or both evaluators, ultimately leading to the identification of epistasis among 69 WGD paralog pairs (17% of all pairs tested, 15% of all WGD paralog pairs). This frequency of epistasis for WGD paralog pairs is well beyond what would be expected for randomly selected gene pairs (<1% based on synthetic genetic array data[120]), and furthermore, beyond the 8- to 10-fold increases in epistasis expected for gene pairs with similar or identical GO annotations, respectively[120].

Table 3-1 Results of RSA screening

		Scorer 1				
		SS	**SL**	**Unclear**	**No Interaction**	**Dead Cells**
Scorer 2	**SS**	37	5	10	0	0
	SL	1	8	8	0	0
	Unclear	0	1	12	10	0
	No Interaction	0	0	1	278	0
	Dead Cells	0	0	3	13	12

Resulting colonies from all assayed WGD paralog pairs were photographed and assessed by 2 independent researchers (indicated as Scorer #1 and Scorer #2). Pairs evaluated by both scorers as being either synthetic sick (SS) or synthetic lethal (SL) were treated as genetic interactors (indicated in red). Any pairs evaluated by either researcher as being 'Unclear' (indicated in blue) were analyzed through tetrad dissection.

Next, Growth-Curve Analysis (GCA) was applied as an alternate means to quantify growth rates to detect attenuated instances of epistasis among WGD paralogs. Unlike in RSA, growth for GCA is assayed in rich liquid media and culture growth monitored through optical density, allowing more precise calculations of fitness, and hence, identification of more subtle growth defects (see **Methods**). Through GCA, 119 double-mutant strains were identified as having growth decreased beyond that predicted by multiplying the fitness defects of the corresponding individual mutations (i.e. using a multiplicative model; see **Methods**), 71 of which had not been witnessed through RSA (see full list in **Appendix Table 1**). These data suggest that epistasis exists to some extent among 140 WGD paralog pairs (35% of those surveyed, 31% of all 457 WGD paralogs).

The RSA and GCA data were not completely overlapping (see **Figure 3-1** and **Appendix Table 1**), as nearly one-third of paralog pairs with obvious growth defects detected in RSA showed normal growth properties through GCA. These remaining 21 pairs may have a function that specifically limited double mutant growth in the minimal media of the RSA experiments (for example, included in this list are *ENO1* and *ENO2* which are known to function in concert only in low-glucose conditions[163]) or the ability to grow on solid media.

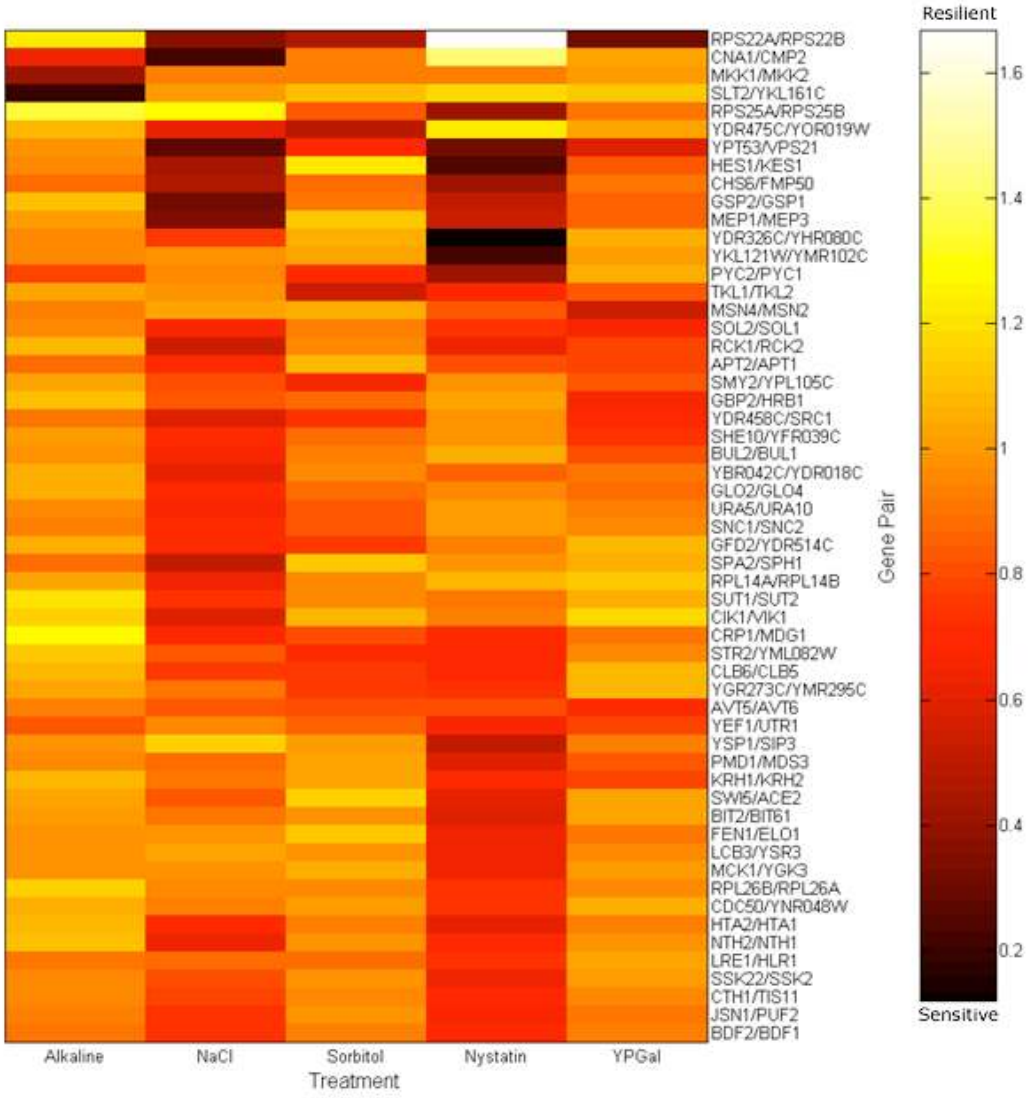
Overall, these results indicate that epistasis occurs among approximately one-third of WGD paralog pairs. However, the difference in epistasic relationships as measured across liquid and solid media suggests a prevalence of condition-specific lethality, underscoring the importance of assaying genetic buffering under multiple conditions.

3.3.2. Further buffering relationships only evident under stress conditions

To further explore epistasis in alternate conditions, and to investigate previous hypotheses that buffering relationships may specifically be maintained to cope with cellular perturbations[53,99], I expanded the assay to include a series of stresses. The relative fitness of the 259 double-deletion mutants corresponding to non-buffering WGD pairs (i.e. those not defined to be epistatic by either RSA or GCA; see **Methods**) were monitored for sensitivity in five media designed to induce cellular stress (alkaline pH, high salt, the presence of the antifungal agent Nystatin (decreases membrane permeability), high levels of Sorbitol, or with galactose substituting for dextrose as the main carbon source)[52] (**Figure 3-4**). As a competitive growth assay suited to examining multiple media types, we measured the relative fitness of bar-coded double-mutant strains simultaneously by quantitative microarray hybridization[52]. Double-deletion strains exhibiting sensitivity within any of the 5 stress conditions (**Figure 3-5**) were further analyzed using GCA conducted in the specific condition, allowing independent validation of these results.

To be considered epistatic in a stress condition, the paralog pair had to satisfy two criteria. First, the pair had to not be epistatic in standard media conditions, indicating that if epistatis was detected in a stress condition, it was unique to that condition. Second, epistatic paralogs had to pass the multiplicative model in the given condition, meaning that either individual knockout could not be responsible for the stress sensitivity. We first detected candidate condition-specific fitness defects through microarray, as indicated by a >25% decrease in strain abundance for a given condition (see **Methods**), among 61

Figure 3-5 Stress responsive paralogs



In 10 of the 261 WGD dual-deletion strains, significantly decreased abundance (decreased detection of tag in microarray hybridization indicated in red above) were observed in at least one of the 5 media conditions tested. Rankings indicated were averaged for all 10 tags present on array, and over 2 independently conducted experimental runs. Only pairs showing abundance rank decreasing beyond one standard deviation of average (see methods) independently in both experiments. All 16 pairs meeting this criteria were later further confirmed through GCA in the appropriate condition.

paralog pairs not detected as buffering in standard media by either RSA or GCA (23.5% of those tested). To ensure that fitness defects were the result of the combined mutation and not the deletion of either single mutant, we conducted GCA for each of these 61 pairs in the appropriate sensitive condition. I found that 10 of the 61 pairs met the stringent confines of a multiplicative fitness model, and thus were epistatic under the given stress conditions (see **Table 3-2**).

These 10 WGD pairs (herein referred to as Sensitive to Stress or SS) both confirm previously described observations and suggest new functional synergies. Among those relationships confirmed are the hypersensitivity to high salt of the *BUL1/BUL2* double mutant[164], and the hypersensitivity to carbon-source starvation of the *MSN2/MSN4* double mutant[165]. Interestingly, the *PYC1/PYC2* double mutant with known hypersensitivity to glucose[166] was found here to be epistatic in multiple conditions, but showed normal growth with galactose as the primary carbon source. Other characterizations of epistasis can also be explained given known protein functions. For example, *YGK3* & *MCK1* are members of the cell-wall integrity pathway[167], while *CHS6* & *BCH2* are involved in the transport of membrane proteins[168]; both pairs were found here to be epistatic in the presence of Nystatin. Further, *CNA2* and *CMP2* which were epistatic here in 1M NaCl are involved in ion homeostasis and confer tolerance to high levels of sodium[169]. My interpretation of these findings is that WGD paralog pairs which are non-epistatic in standard conditions may be epistatic in a condition that emphasizes their principal function, despite the fact that neither individual paralog is essential for survival in that function (i.e. epistasis can be evolved for specific circumstances).

Table 3-2 Pairs indicated as sensitive to stress

Systematic Name 1	Systematic Name 2	Standard Name 1	Standard Name 2	Condition
YML111W	YMR275C	*BUL2*	*BUL1*	NaCl
YLR433C	YML057W	*CNA1*	*CMP2*	NaCl
YBR218C	YGL062W	*PYC2*	*PYC1*	Sorbitol, Nystatin
YPR074C	YBR117C	*TKL1*	*TKL2*	Nystatin
YJL099W	YKR027W	*CHS6*	*BCH2*	Nystatin
YNL307C	YOL128C	*MCK1*	*YGK3*	Nystatin
YDR326C	YHR080C	*YSP2*		Nystatin
YKL062W	YMR037C	*MSN4*	*MSN2*	Galactose
YCR073W-A	YNR034W	*SOL2*	*SOL1*	NaCl, Nystatin, Galactose
YGL228W	YFR039C	*SHE10*		NaCl, Galactose

Listed are the set of 10 non-buffering WGD paralog pairs assessed to be sensitive to at least one of the five experimental conditions examined (first identified through reduced hybridization of competitively grown bar-coded double-mutants to a microarray, then subsequently confirmed using GCA).

Based on our observations of condition-specific epistasis among SS paralogs, I am able to propose function for some genes that are either uncharacterized or poorly understood. Specifically, as *YSP2* and its paralog *YHR080C* are epistatic in multiple conditions, I propose that the known pro-apoptotic protein *YSP2* may also exercise an inherent rescue function in conjunction with its paralog in response to cell wall or plasma membrane damage. Also, the epistasis of *SOL1* & *SOL2* in multiple conditions suggests that they may be involved in a general stress response. Lastly, I submit that the putative protein of unknown function *YFR039C* functions synergistically (either in conjunction or within an alternate functional pathway) with the glycosylphosphatidylinositol anchored protein *SHE10* in response to cell stress, notably high salt and alterations in carbon source.

3.3.3. Experimental condition impacts composition of paralogs deemed epistatic

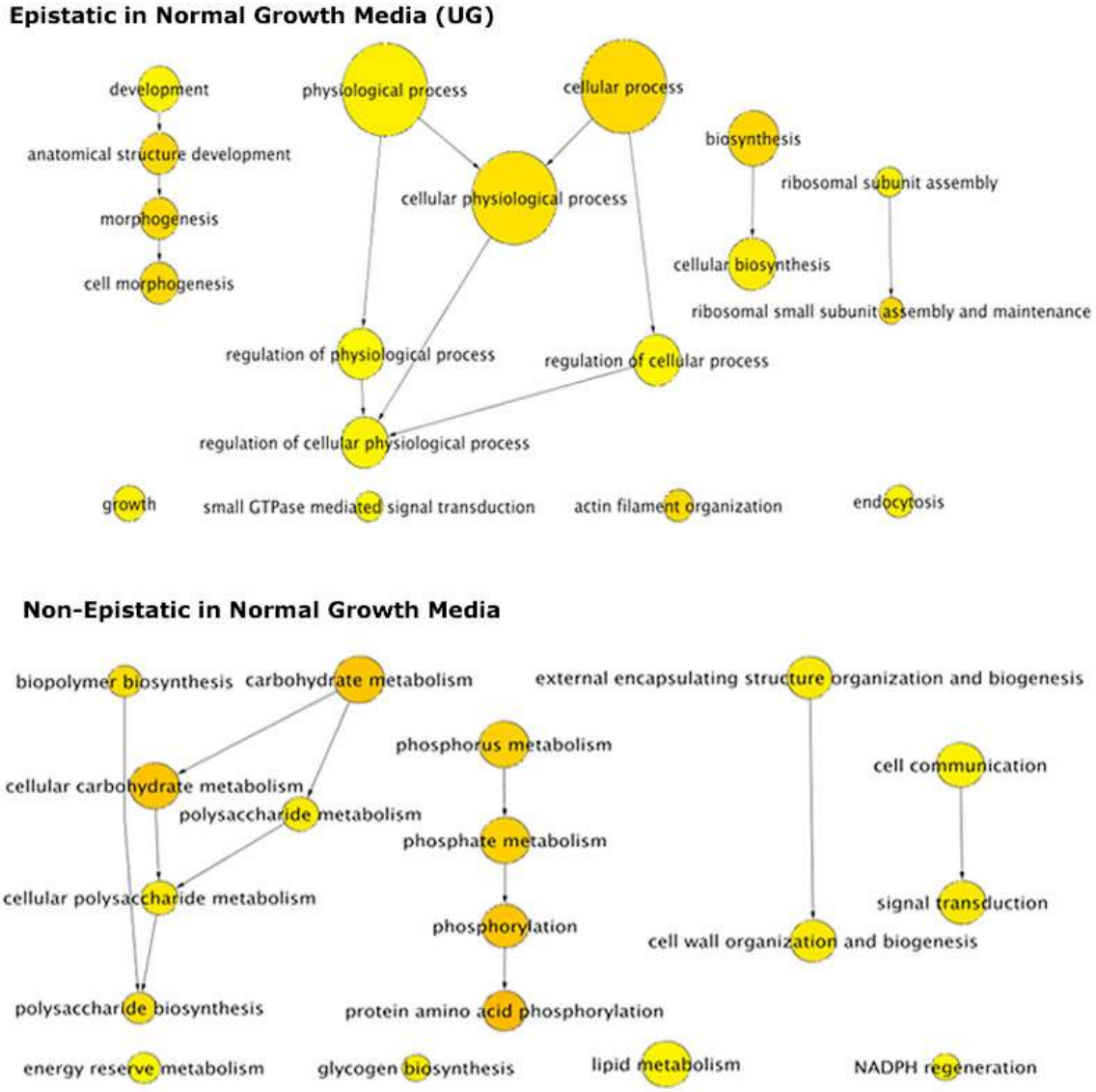
As investigation above had revealed SS paralogs to generally have functions related to their incident condition (analogous to what has been previously reported for metabolic enzymes using flux balance analysis[153]), I next tested whether those paralogs epistatic only under standard conditions (i.e. those initially detected by RSA and GCA) had any common functional properties. Enrichment analysis[162] of functional categorization as assessed by the Gene Ontology (GO) database[170] indicated that these paralogs were biologically disparate from their non-buffering counterparts. Specifically, WGD paralogs epistatic under standard laboratory conditions were more frequently involved in cell-growth, protein metabolism, and division.

Many buffering paralogs (RSA and GCA) are annotated as ribosomal proteins or metabolic enzymes (23 pairs in each). However, even upon removal of these proteins, functional GO enrichment analysis revealed that, although somewhat more functionally diverse, the remaining paralogs epistatic under normal growth conditions were still significantly enriched for growth related processes. More specifically, while paralogs detected jointly by both RSA and GCA were mainly involved in protein production, those detected solely by GCA were notably enriched for involvement in the cell-cycle (i.e. septin ring formation, cell-cycle checkpoint, regulation of cell-cycle), and those detected solely by RSA were seemingly enriched for metabolic processes and structure development. No paralog pairs detected to be epistatic jointly by both RSA and GCA contained genes of unknown function. In contrast, non-epistatic paralogs were significantly enriched for elements of signal transduction (e.g. amino-acid phosphorylation, signal transduction, cell communication; see **Figure 3-6**), and frequently (25% of pairs) one or both paralogs was of unknown molecular function. These results suggest a distinct impact of survey condition on the functional properties of detected epistatic relationships.

3.3.4. Paralogs buffering under standard conditions are highly conserved

I next investigated whether buffering paralogs had any additional unique evolutionary or genomic properties correlating with their epistatic capacity. To eliminate any potential method-specific bias, when evaluating paralogs epistatic under normal

Figure 3-6 Functional composition of epistatic paralogs



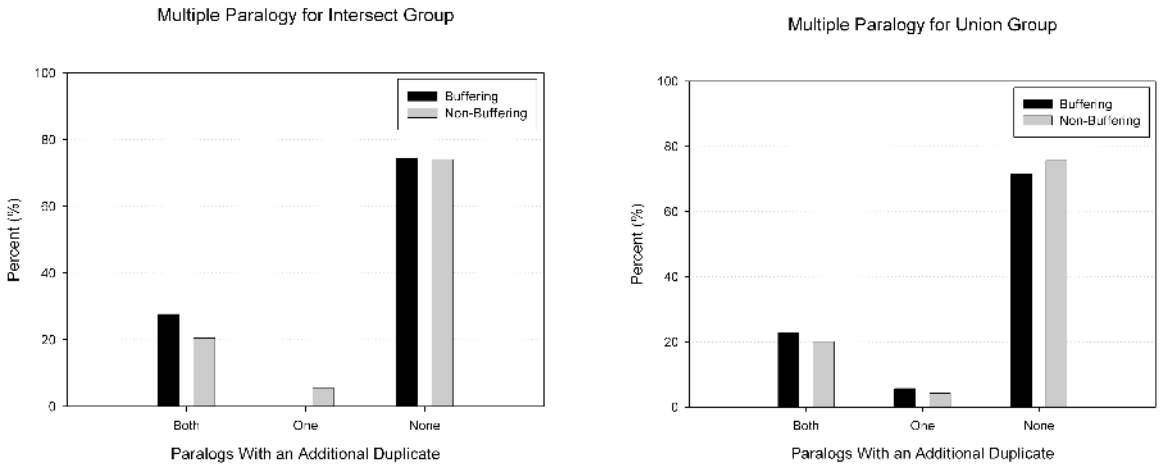
Depicted are the GO functional categories over-represented for paralogs both epistatic (above) and non-epistatic (lower half) under normal growth conditions (i.e. those detected by either RSA or GCA, the union group). The size of the node reflects the fraction of paralogs involved in that over-represented process, and darker coloring denotes higher significance. While the significance of each term was calculated independently (regardless of placement within the GO classification system), arrows indicate continuance along the path of the GO directed acyclic graph. Increased statistical stringency ($p < 0.0005$) was used for depicted terms in order to account for increased representation of more general terms. Picture created using the BinGO plug-in for the Cytoscape software environment.

laboratory conditions I confirmed results using paralogs separated on the basis of joint detection by both the RSA and GCA (intersect group, IG), and using either of these two methods alone (union group, UG).

First, I examined whether presence of additional (i.e. non-WGD-resultant) duplicates impacted the potential for phenotypic buffering. To analyze the affect of multiple paralogy on epistasis I sub-divided WGD paralogs based on the existence of additional duplicates and compared frequencies of epistasis. Based on these identifications, I sub-divided WGD paralog pairs into three groups: those pairs where both members had additional paralogs, those pairs where only one member had an additional paralog, and those pairs where neither had additional paralogs (21%, 4% and 75% of all WGD paralog pairs respectively). Upon comparison, epistatic and non-epistatic paralogs showed no appreciable differences in the representation of the three groups of WGD duplicates (see **Figure 3-7**). Therefore I conclude that presence of an additional duplicate does not influence the propensity of a WGD paralog pair to be epistatic.

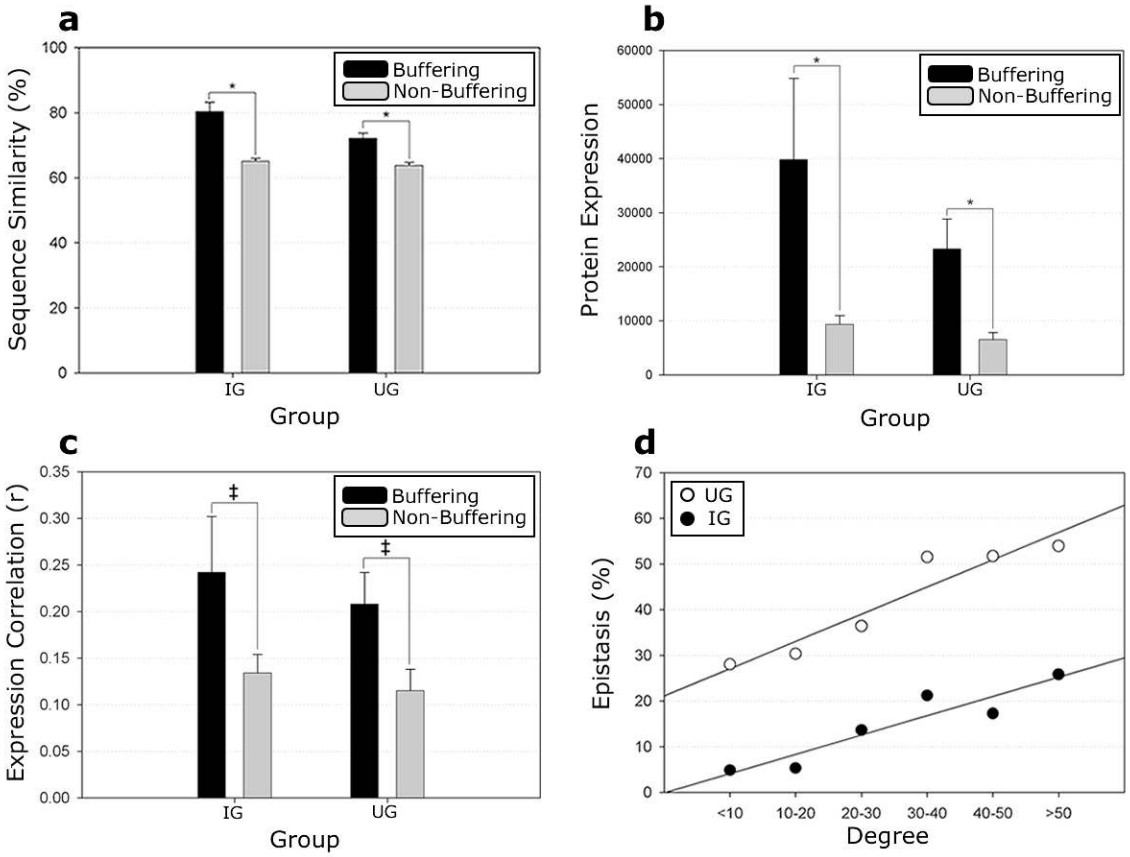
Next, upon examining sequence conservation, buffering paralogs (IG and UG) had significantly higher average sequence similarity (86% for IG, 77% for UG, versus 65% for non-buffering) than corresponding non-buffering paralogs, and a consequent decrease in non-synonymous mutation rate (tallied using *K. waltii* as the ancestral out-group; see **Methods**). These findings were statistically robust ($p < 0.05$ for all comparisons, Mann Whitney Rank Sum test; see **Figure 3-8a**) against the removal of ribosomal proteins which exhibit disproportionately high levels of sequence conservation.

Figure 3-7 Impact of additional duplicates on epistasis



WGD paralogs were divided into three groups based on the presence of additional (non-WGD) paralogs. Paralogs were classified as being: Both (both members of a pair of WGD paralogs have additional duplicates in the *S. cerevisiae* genome), One (only one pair member has additional duplicates), or None (neither WGD paralog has additional duplicates). Depiction of epistasis as assessed by both RSA and GCA (a), or by either RSA or GCA (b) indicates no overall difference in the composition of WGD paralogs with additional, non-WGD, duplicates.

Figure 3-8 Properties of epistatic paralogs



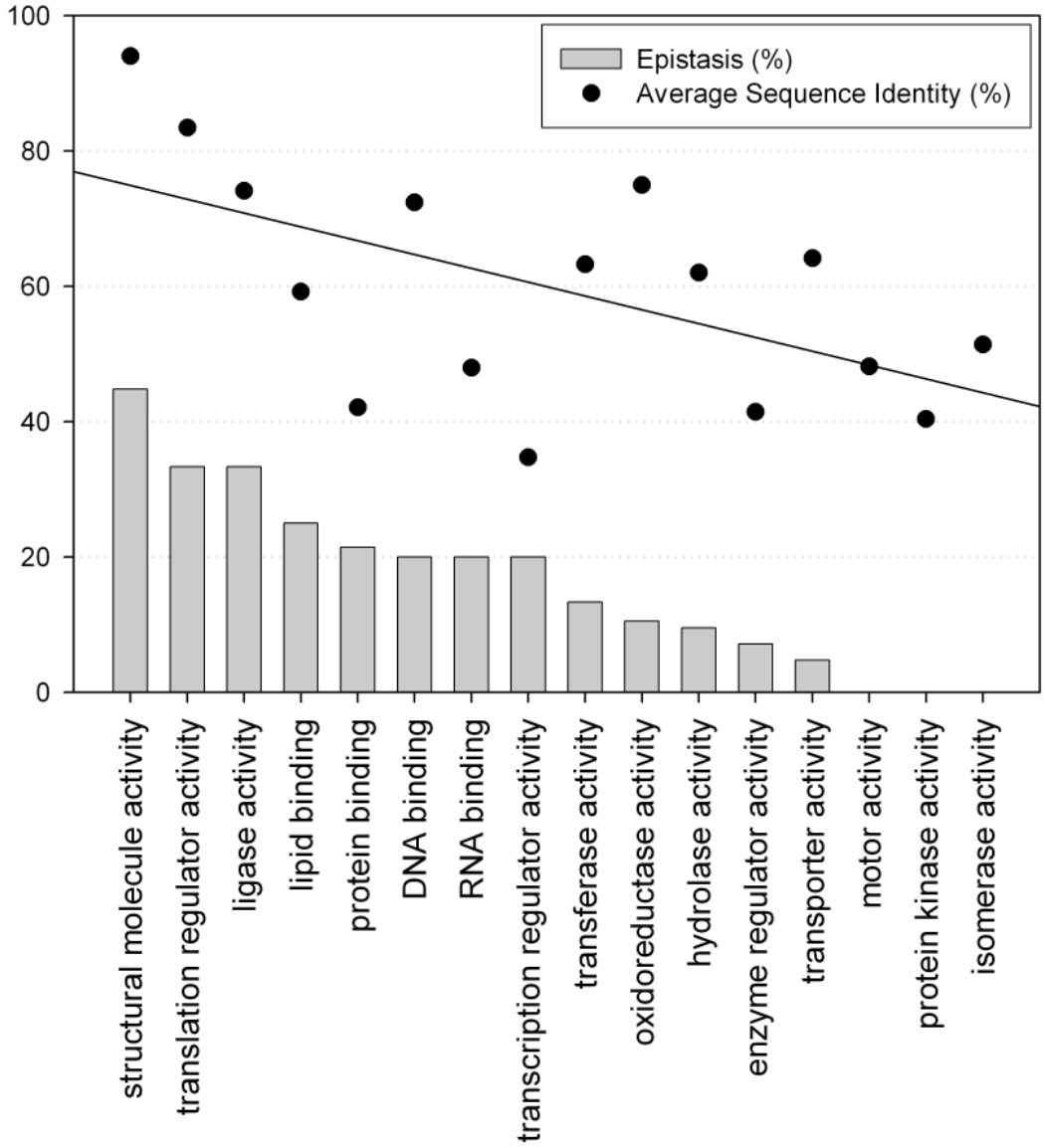
All relationships depict comparisons between epistatic paralogs (IG and UG) and respective non-epistatic following the removal of ribosomal proteins. **a** Buffering paralogs have significantly more conservation of sequence than non-buffering ($p < 0.001$ for IG and UG, as indicated by asterisk). **b** Buffering paralogs are significantly more highly expressed than non-buffering (protein abundance depicted, same results noted for transcript abundance) and additionally exhibit a trend towards ($p < 0.1$ as indicated by ‡) being more highly correlated throughout the cell cycle (displayed in **c**). **d** All WGD paralog pairs were binned based on their number of combined protein interactions (degree) within the BioGRID dataset, with the respective percentage epistatic (both IG and UG) indicated, demonstrating a clear relationship between the number of physical interactions and the likelihood for epistasis under normal growth conditions.

Despite these trends, it is interesting to note that this increased within-species conservation is not exclusive to buffering paralogs. For example, the UG contained 3 buffering paralog pairs sharing less than 40% sequence identity (*PHD1* & *SOK2*, *MGA2* & *SPT23*, and *BOI1* & *BOI2*), implying that conservation is not a pre-requisite for phenotypic buffering. Furthermore, SS paralogs did not exhibit the same stringent conservation of sequence as IG and UG paralogs (average sequence similarity among the 10 confirmed SS paralogs was 68%).

On average, buffering paralogs (IG and UG) also exhibited significantly higher basal mRNA[141] and protein[46] expression levels ($p < 0.05$, Mann-Whitney Rank Sum test, **Figure 3-8b**) when compared to non-buffering (robust against removal of ribosomal proteins, see **Figure 3-8b**). As expression magnitude is intrinsically linked to sequence conservation for yeast paralogs[150], and epistatic paralogs are more highly conserved, this finding is not unexpected. However, while IG and UG paralogs initially had more highly correlated mRNA expression both throughout the cell cycle ($p < 0.01$; see Methods), and across varying cell conditions[97], following the removal of ribosomal proteins, these relationships are no longer significant. Therefore, expression correlation does not appear to be generally predictive of epistatic capacity for WGD-resultant paralogs.

A breakdown of WGD paralogs by functional category as assigned by GO Slim (http://geneontology.org/GO.slims.shtml) demonstrates that epistasis under standard conditions is much more frequent among those paralogs involved in certain growth and division related processes (**Figure 3-9**). Also, there is a direct linear correlation between the fraction of buffering paralogs within a functional group, and the overall sequence

Figure 3-9 Epistasis by functional category



WGD paralogs were grouped into functional categories based on the broad definitions of the GO slim hierarchy and ranked in decreasing order based on the percentage of paralogs found to be buffering (IG), indicated with bars. Only those paralog pairs with single, matching annotations were included, resulting in 315 depicted pairs (categories 'other' and 'molecular function' were removed). The juxtaposed dotplot indicates the average percent sequence identity of functional groups (buffering and non-buffering paralogs combined), the overlayed line indicates linear regression. There is a significant correlation ($r=0.64$, $p<0.005$) between the epistatic capacity of a functional group and the conservation of paralogs contained therein.

similarity between paralogs in that group ($r=0.64$, $p<0.005$, linear regression $p < 0.05$; see **Figure 3-9**). Therefore, as function and conservation appear to be intertwined, the possibility that functional bias is responsible for the observed increase in conservation of paralogs epistatic under normal conditions cannot be immediately discounted.

Information gain analysis was used to further examine the nature of epistasis in various functional gene categories (see **Methods**). While high sequence similarity and co-expression are generally predictive of epistasis for all pairs detected through RSA, division by functional category suggests that high average physical interaction degree is predictive of epistasis among genes classified as RNA, DNA and protein binding. Further, correlation across various experimental conditions is predictive of epistasis for genes involved in transcriptional regulation. Together these results suggest that while certain features of epistatic duplicates are generally true, greater insight regarding epistatic relationships can be gleaned through functional dissection.

3.3.5. Physical interactions are not indicative of phenotypic buffering

To test a recently observed correlation between the number of physical interaction partners (the 'degree' of a protein) and the propensity to backup a sister paralog[97], I next compared interactions of epistatic and non epistatic paralogs. Using the identical dataset that this assertion had been based on (the full interaction dataset contained at BioGRID[139]), both IG and UG paralogs had significantly more interactions per pair than non-buffering (**Figure 3-8d**, $p < 0.05$; Mann Whitney Rank Sum Test). The significance of this relationship was robust against removal of ribosomal proteins which had a disproportionately high number of interactions. Alternately, SS paralogs were not

significantly greater or lesser in degree than the remaining WGD paralogs, suggesting again that increased degree may be linked to the functional bias, increased expression, or increased conservation of IG and UG paralogs, and not necessarily an inherent property of epistatic relationships. The difference in degree between sister paralogs had no noticeable correlation with epistasis.

One could logically assert that paralog pairs capable of buffering deletion of their sister should be more congruent in function than those that do not. I next attempted to test this assertion using annotated physical interactions as a direct proxy of function[100]. I compared the extent of shared protein-protein interaction partners of buffering versus non-buffering WGD sister paralogs based on the results of two high-throughput proteomic screens[48,49]. To eliminate potential biases created by high-throughput assay, I further confirmed all results using the aforementioned BioGRID dataset[139] following the manual removal of data resulting from high-throughput screens. Buffering and non-buffering WGD paralogs were compared in the following aspects: (i) propensity of the sisters to physically interact with each other, (ii) propensity to be co-grouped into associative protein complexes, (iii) the number of shared interacting partners, and (iv) the proportion of non-shared partners (first two analyzed by Fisher exact test, all else using Mann-Whitney rank sum test; see **Methods**).

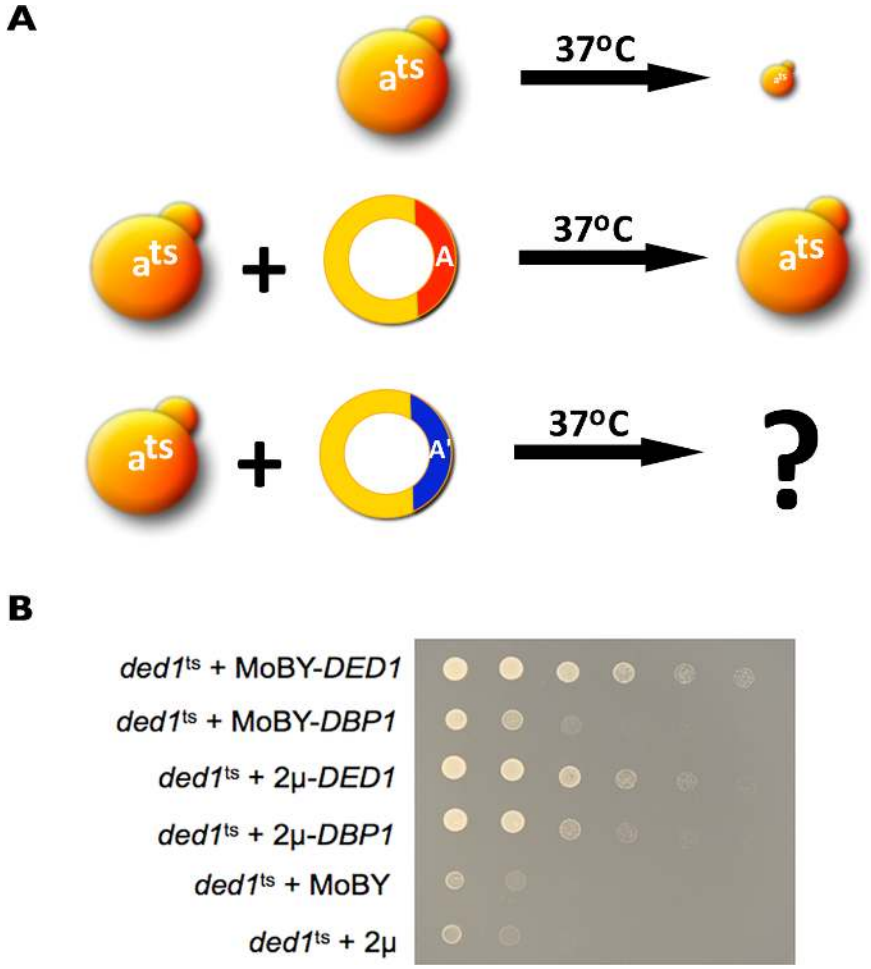
Only one comparison achieved marginal statistical significance. IG paralogs were more likely to interact with each other in the Krogan *et al* interactome dataset ($p < 0.05$), but this trend was not confirmed in either the Gavin *et al* genome-wide study or using the BioGRID dataset. Although my findings depend on the assumption that the current physical interaction data provides a fairly complete and consistent coverage of the WGD

paralogs, these results perplexingly suggest that paralog pairs deemed to be buffering under standard laboratory conditions are not necessarily more functionally redundant. As previous studies have indicated a pervasive overlap in terms of the physical interaction partners of WGD-resultant duplicates[54,99,100], one potential explanation for the lack of correlation between physical interaction data and epistatic capacity is that not all epistatic relationships have been revealed.

3.3.6. Epistasis present among paralog pairs containing only one essential gene

Finally, examination of phenotypic rescue using strains containing temperature-sensitive (ts) alleles of paralogous essential genes demonstrated that some of these pairs have maintained capacity for phenotypic rescue (see **Figure 3-10**). Of six WGD-resultant paralog pairs with available *ts* strains, two showed obvious phenotypic rescue (*DED1/DBP1* and *GSP1/GSP2*) confirming previously observed relationships[171,172]. Analysis of conservation and co-expression do not reveal any properties that generally differentiate these two pairs from the four not demonstrating phenotypic rescue. Although a small sample set, these results clearly demonstrate the possibility for functional retention among paralogs with one essential member, contradicting assumptions that these genes had lost their ancestral (essential) function[17].

Figure 3-10 Phenotypic rescue of essential paralogs



Schematic illustrating the principle of phenotypic rescue detection among WGD paralogs. A strain carrying a temperature-sensitive allele of essential gene '*A*' is inviable at 37C, however expression of '*A*' via plasmid allows growth. **(B)** Results for the paralog pair *DED1*/*DBP1*. Growth shown for strains carrying a temperature-sensitive allele of the essential gene *DED1* and either: endogenous expression via the MoBY plasmid of *DED1* (*MoBY-DED1*) or *DBP1* (*MoBY-DBP1*), increased expression via 2µ plasmid of *DED1* (*2µ -DBP1*) or *DBP1* (*2µ -DBP1*), or the corresponding empty plasmids. Growth is at 37C, columns represent decreasing dilution (left to right). The resulting growth clearly demonstrates a rescue effect when *DBP1* is over-expressed. Although not depicted, identical results were observed for *GSP1/GSP2*.

3.4. Discussion

The noted lack of cell morbidity when deleting individual yeast genes with at least one paralog has lead to speculation that duplication is reserved for genes of limited functional importance[95]. By extension, WGD-resultant paralogs (which have generally less consequence upon single-gene deletion than SSD-resultant[59]), should be of even less functional importance than duplicates of other origin. However this absent decrease in phenotypic consequence upon deletion of a WGD-resultant duplicate can alternately be explained by the elevated capacity of this group to buffer deletions. The degree of epistasis among duplicates is inevitably a compromise between the maintenance of mechanisms that could both protect against gene loss and confer stress resilience, and the establishment of functional novelty. As I have witnessed here that epistasis is possible among even minimally co-maintained WGD paralogs, complete abolition of functional overlap would seem to have very little comparative benefit.

The frequency of WGD paralog pairs exhibiting phenotypic buffering under normal media conditions (~35% of those surveyed, or ~31% of all 457 WGD paralogs) observed in this study is similar to that previously determined through examination of a much smaller subset of duplicates (25 WGD- and 20 SSD-resultant) using a high-density genetic interaction map[154] and to two subsequently-published similar surveys[124,173]. However, while the previous study suggested that there was little additional evidence of functional redundancy in alternate environmental conditions, my results conversely suggest there may be limitless potential to reveal buffering.

While paralogs epistatic in solid minimal media and rich liquid media are notably more rigorously co-conserved and highly expressed, neither of these properties appear to

be essential for epistasis. Further, overlap in shared physical interaction partners, which serves as a proxy for overlapping function, showed no consistent difference between epistatic and non-epistatic paralogs. In a finding just as perplexing, presence of additional, non-WGD-resultant duplicates appeared to have little to no impact on phenotypic buffering. As the nature of epistasis among WGD paralogs is highly dependent on the environmental context, i.e. "*environmental robustness*"[153], and as only a small fraction of the potential stresses (as well as the endless permutations of stress magnitudes and combinations) and alternate environmental conditions were explored here, many non-epistatic WGD paralog pairs (25% of which are of unknown function) may yet have preserved an un-witnessed buffering mechanism over ~100 million years of evolution. This is reminiscent of what had previously been observed in comprehensive phenotypic studies of single-gene deletion strains in *S. cerevisiae* wherein loss of only ~19% of genes results in morbidity (the so-called essential genes[52]). Just as many single genes serve a function imperceptible under laboratory conditions but are required for viability specifically under cellular duress[52,174] or are suspected to function in non-laboratory growth conditions[175], certain WGD paralogs may retain a condition-specific buffering capacity.

The observations regarding concerted expression of buffering paralogs presented herein are surprising given previous findings. Specifically, it had been demonstrated that dispensable (i.e. buffering) paralogs generally have both a higher degree and more disparate expression patterns than non-buffering[92,97]. While I do confirm here that WGD paralogs epistatic under standard conditions have higher degree, they displayed no more disparity in expression than non-buffering (initially found to be significantly more

correlated, although this was a byproduct of the influence of ribosomal proteins), illustrating potential differences between the epistatic mechanisms of WGD-resultant paralogs and duplicates of other origin (i.e. SSD).

As the existence of duplicated genes can confound predictions of epistasis[153], a comprehensive understanding of the buffering capacity of paralogs is essential before extrapolating the knowledge gleaned form model organisms such as budding yeast to higher eukaryotes. Since the human genome has arguably been profoundly influenced by genome duplication events[29] it will ultimately be the understanding of adaptive epistatic mechanisms, evolved to cope with the abnormal states of stress and perturbation, that will lend the most insight into disease-related maladaptive processes. However as many of these relationships may not be obvious either under standard assay conditions or using standard techniques, new approaches are necessary to determine the full extent of epistasis and thus the full impact of WGD events on the genome.

Chapter 4

Discovering functional redundancies through triple-deletion SGA

Andrew Emili, Zhaolei Zhang, Charles Boone and Michael Costanzo supervised and advised the experimentation and I performed all relevant analysis. I also performed all of the described experimentation with the exception of the three *Hygromycin B* control screens which were run by JiYoung Youn. Also, I was assisted in microscope analysis by Franco Vizeacoumar and in scoring of SGA plates by Anastasia Baryshnikova and Chad Myers (University of Minnesota).

4.1. Introduction

Evidence collected in several laboratories and using varying methodologies (including my own work above) has demonstrated the frequency of synthetic genetic interactions to be as high as 35% among extant yeast duplicates[124,173], suggesting overlap in function to be substantial among this group. Further, my previous examination of condition-specific epistasis (confirming similar observations based on single-deletion profiles[53] and FBA[85]) suggested that observation of these interactions is highly dependent upon survey condition, and that many further paralog pairs may be epistatic in alternate conditions. As the condition space is infinite, an alternate means is necessary to assess redundant function in these (non-epistatic) paralogs. While not representing a change in extracellular environment, successive gene deletions perturb the cell in unique ways, and thus may be useful in understanding the coordinated response of duplicated genes.

Genetic interactions among three or more genes are more rarely surveyed than digenic interactions, but have been useful in understanding the nature of multi-component pathways such as DNA-damage[176,177], stress[178], and nutrient[179] response. Typically multi-gene interactions are used to determine impacts of successive deletions on a previously established phenotype. However as most pathways and complexes involve more than 2 members, and since complex diseases such as cancer can have etiologies that involve multiple simultaneous mutations[180], assays for large scale screening of multiple genes would be highly insightful. For this reason, screening of double-deletion query strains is thought to be the next frontier of large-scale genetic interaction assay[181]; however examples of double-deletion screens are currently lacking.

In the only previously published large-scale assay of the genetic interactions of double-deletion queries, Tong *et al* investigated the interactions resulting from Synthetic Genetic Array (SGA) screening of 3 mutant strains containing deletions of the pairs: *BIM1/BNI1*, *BNI1/KRE1* and the WGD-resultant paralog pair *CLN1/CLN2*. While many interactions were initially detected in both the *BIM1/BNI1* and *BNI1/KRE1* screens, a comparatively small number were determined to be specifically the result of the triple deletion during tetrad dissection (4 confirmed out of 171 detected for *BIM/BNI1* and 29 out of 156 for *BNI1/KRE1*), suggesting that a good deal of rigor is required to confirm triple-mutant interactions. The paralog pair *CLN1/CLN2* showed 36 interactions uniquely attributable to the triple deletion, indicating that SGA conducted using double-deletion queries (herein referred to as triple-deletion SGA) can be used to demonstrate unique interactions for non-epistatic paralog pairs. However, the extent of interactions detectable through triple-deletion SGA as well as the possible implications toward predicting epistasis or other functional associations remain unexplored.

4.1.1. Specific rationale and hypothesis

Useful both as a gauge of functional overlap and to assay for the presence of potential back-up mechanisms, aggravating synthetic genetic interactions occur when genes have a greater detriment to either cell viability or another observable phenotype when deleted within the same strain than would be predicted based on individual deletions. In an effort to further characterize any functional redundancies among non-epistatic paralog pairs I used the SGA protocol to facilitate screening using query strains carrying double-deletions of paralog pairs under the assumption that redundant functions

would uniquely be identified in the double-deletion SGA profile (see **Figure 4-1**). Through application of this technique I expect to demonstrate manifested instances of redundancy among paralog pairs with no other obvious indication of functional overlap.

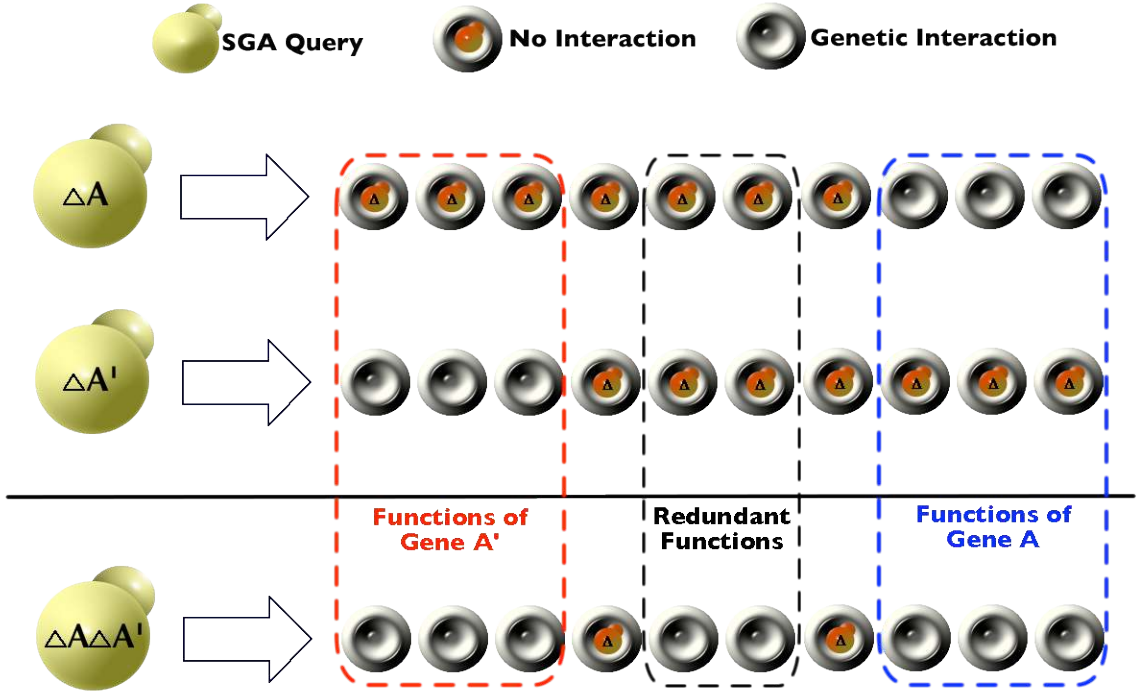
4.2. Methods

4.2.1. Construction of SGA double-deletion query strains

To facilitate double-deletion screening, resistance to *Hygromycin B* (*hyg*) was used as a selectable marker in addition to the *Nourseothricin* and *Kanamycin* resistance markers traditionally applied in SGA screening[119]. Since *Hygromycin B* inhibits similar processes as the already employed *Nourseothricin* and *Kanamycin* analogs (protein synthesis[182]), I reasoned that there should be very little additional toxicity. A three-step process was used to create double-deletion query strains for assay in SGA (see **Figure 4-2**). First, a diploid strain from the SGA deletion array (*kanMX*) containing the *Kanamycin* resistance gene in place of a gene of interest (referred to as *p1*; *MATa/MATα p1::kanR/P1*) as well as other auxototrophic markers used in the SGA process (*ura3Δ0 leu2Δ0 his3Δ1 met15Δ0*) was replaced with hygromycin resistance via a standard lithium acetate transformation protocol[183]. Switching involved transformation of a PCR fragment amplified from the *pFA6a-hphMX6* plasmid[184] using primers annealing directly upstream and downstream of the promoter and terminator sequence, respectively (5'-GACATGGAGGCCCAGAATAC-3' and 5'-TGGATGGCGGCGTTAGTATC-3').

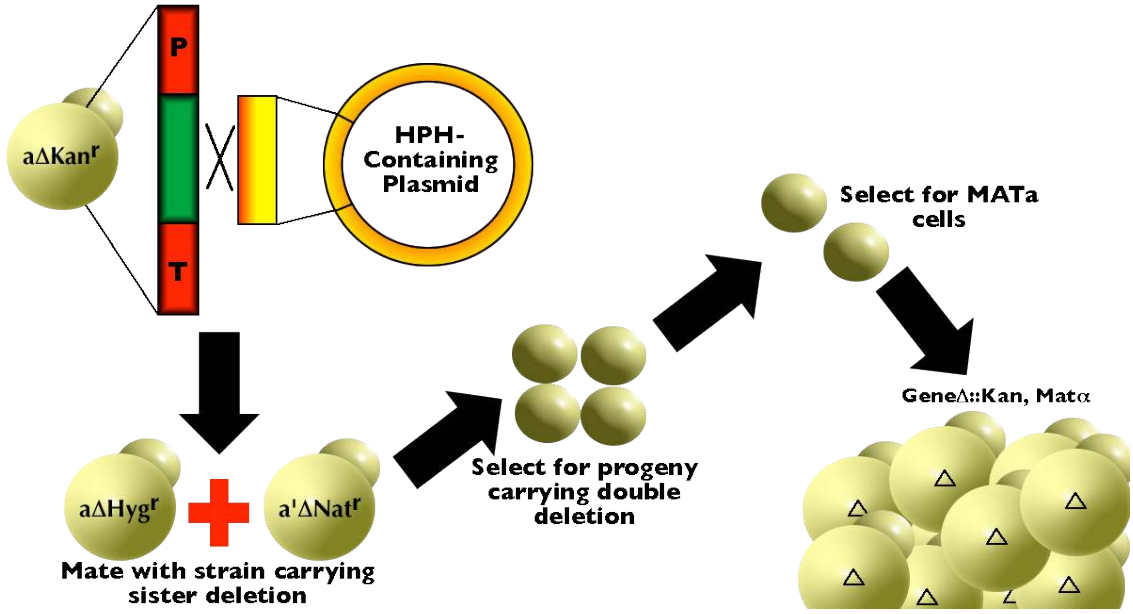
Next, *Hygromycin B* resistant strains were transferred to sporulation media and tetrads

Figure 4-1 Detection of redundant functions through triple-deletion SGA



The concept underlying triple-deletion SGA is that any interactions detected uniquely in the profile of the double-mutant (in this case ΔAΔA') were initially masked by redundant function and therefore not obvious in either single deletion screen. I expect interactions resulting from divergent (unique) functions will be obvious in the double-mutant screen as well.

Figure 4-2 Creation of query strains



Strains from the available *Kanamycin* resistance collection (with the *Kanamycin* resistance gene in place of a given paralog '**a**') undergo marker switching via transformation to contain *HPH*, the *Hygromycin B* resistance gene in place of gene '**a**'. The '**a**' deletion strain will then be mated with one containing *Nourseothricin* resistance in place of the sister paralog's gene (gene **a'** above). Appropriate progeny will be selected and mated against the SGA deletion collection. To ensure equal expression of markers, the same promoter and terminator sequences (indicated by **P** and **T** respectively above) will be used for all 3 deletion markers.

dissected after 5-7 days to confirm lack of potential aneuploidies and to select for proper mating type and appropriate reporters. Finally, selected strains (*MATa p1Δ::hygR*) were mated overnight with standard SGA query strains containing *Nourseothricin* resistance in place of corresponding paralog (*MATα p2Δ::natR*), sporulated, dissected and selected based on resistance and mating type to yield *MATα p1Δ::hygR p2Δ::natR* strains. These strains were mated against the SGA array strain collection in 1536-spot format.

4.2.2. SGA pinning protocol

Pinning and selection of mated strains was performed as previously described[120], except for the addition of *Hygromycin B* at 300ug/L (determined previously to be the optimal selective concentration[185]) in the diploid selection media (*YPD +G418 +ClonNAT +Hygromycin B*), and addition of a pinning step following selection of *Nourseothricin* and *Kanamycin* resistant haploids on *SD –arginine –lysine –hystidine +G418 +ClonNAT +Hygromycin B +Canavanine +Thiolysine*. Briefly, deletion strains were arrayed in quadruplicate in 1536-spot format using an automated Virtek colony arrayer (http://www.virtekbiotech.com). Query strains were grown in an even lawn overnight and mated to the deletion collection through robotic pinning. Following mating, appropriate haploids were selected through successive pinning steps and growth on appropriate solid media. After final pinning step plates were digitally imaged and growth scored.

4.2.3. Scoring of interactions and batch correction

Once pinned and photographed, images were digitally processed using an in-house software program and pixel counts corresponding to each colony obtained. Colony sizes measurements were then analyzed according to an established protocol (Baryshnikova *et al*; *in preparation for submission)*. Briefly, each colony size measurement was corrected for: plate-specific effects (based on plate-to-plate variability), spatial and position-related systematic effects, row/column effects, and batch survey effects. The purpose of correcting for batch-specific effects was to remove the confounding impact of similarities among plates being run by one individual at one time. However in this instance as all screens were run and photographed by myself using consistent machinery, all 10 surveyed screens were considered to be in the same batch. This should not only remove any consistent effects on behalf of the surveyor, but also any systematic interactions resulting from the modification of the SGA procedure (i.e. addition of *Hygromycin B*). To further ensure that any interactions were not due solely to the use of *Hygromycin B*, a strain carrying the *hph* gene inserted at a locus containing a phenotypically-null allele was screened three times by another investigator (Ji Young Youn) and array strains showing significant fitness defects in at least 2 of the 3 screens were removed from all future triple-deletion screens. Additionally, to remove linkage-effects[186] associated with gene deletions all double-mutant scores were manually examined by location and chromosomal regions displaying notable stretches of decreased colony size were removed from further assay.

Once the corrected colony scores were obtained, the fitness of each array strain was calculated through comparison with the median of all previous occurrences.

Variations in colony size for each strain were used to calculate standard deviations and p-values for inherent growth changes, and interactions calculated as:

$$\varepsilon = W_{AB} - (W_A \times W_B)$$

where ε represents the interaction score, W_A and W_B represent the query and array starting fitness scores, and W_{AB} the observed array score. Systematic analysis of the distribution of ε scores has previously indicated an optimal cutoff of -0.08 for aggravating genetic interactions (Costanzo *et al*; *submitted for publication*). Those strains producing an ε-value below this cutoff and with p-value less than 0.05 were taken to be aggravating genetic interactions.

4.2.4. Analysis of function

Functional enrichment analysis was performed on double-mutant interaction data using the BINGO plugin[162] for the Cytoscape software environment[187]. Where mentioned, correlation was performed using a Pearson's correlation of epsilon values resulting from SGA screening.

4.2.5. Microscopy

Log phase haploid deletion strains were fixed with formaldehyde, permeabilized using triton, and phalloidin-stained. Following staining cells were imaged using a DMI 6000B fluorescence microscope (Leica Microsystems) equipped with a spinning-disk head and argon laser (458, 488, and 514 nm; Quorum Technologies, Guelph, ON, Canada) coupled with an ImageEM-charge-coupled device camera (Hamamatsu

Photonics, Hamamatsu City, Japan). Images from the microscope were analyzed using Volocity software (Improvision, Coventry, United Kingdom).

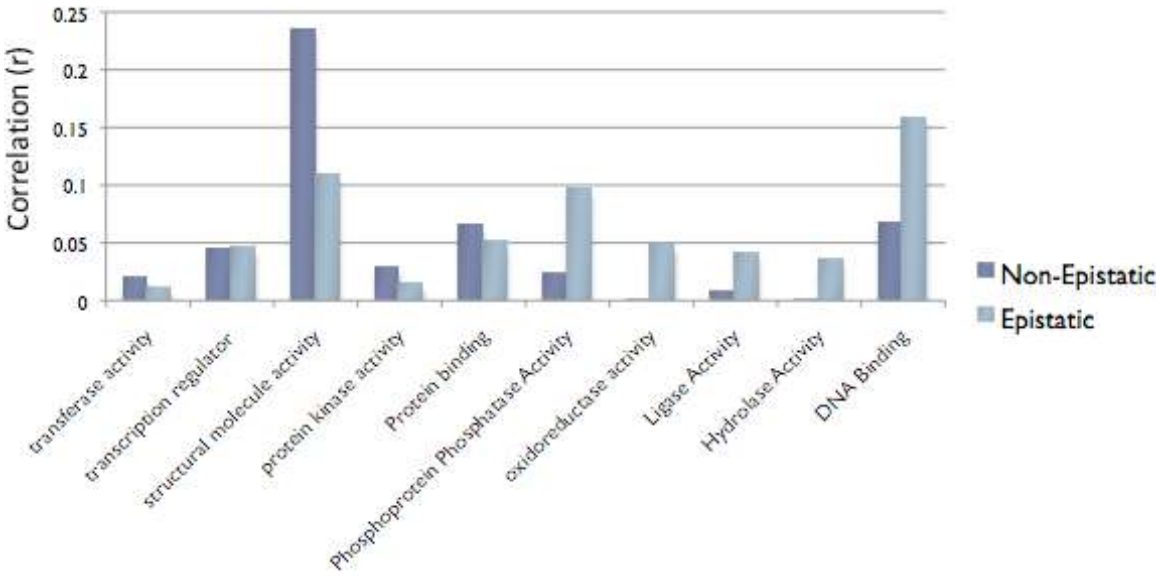
4.3. Results

4.3.1. Selection of strains for SGA screening

As previous results had suggested a lack of overlap in genetic interactions for duplicated genes[154], I sought to first test triple-deletion SGA on a body of paralogs that could directly demonstrate that this was the result of phenotypic masking. For this reason I sought to chose an initial starting set of paralog pairs with low similarity in SGA profiles. After dividing genes by functional category based on the broad associations in the GO[105] SLIM database (http://www.geneontology.org/GO.slims.shtml), I examined similarity in SGA results for paralogs in the various categories via correlation in SGA profiles (SGA data was from an ongoing full-genome screening effort, taken with permission from the Boone lab). This SGA data used consisted of over 2200 full-genome queries (over 6 million datapoints) and thus represents a more in-depth depiction of functional association than has been used in previous assessments.

Upon comparing correlation values for paralogs in the various functional categories, I found that in addition to having low similarity in profile, genes encoding kinases and transcription factors showed no marked increase in similarity when epistatic, as one might generally expect for genes that are functionally similar (see **Figure 4-3**). As paralog pairs in these gene categories are also depleted for pairwise interaction (see *Chapter 3*), a possible explanation for this finding is that not all genetic interactions

Figure 4-3 Correlation in SGA profile by functional category



Following division by functional category, similarity in SGA profile was compared for epistatic and non-epistatic WGD-resultant paralogs. Among those both displaying no general similarity in SGA interaction profile and no difference in similarity between epistatic and non-epistatic pairs are kinases and transcription factors.

between paralogous kinases and transcription factors have been observed. Thus to demonstrate not only phenotypic masking, but also any possible instances of manifested redundancy among these groups, an initial set of kinases and transcription factors was selected for triple-deletion SGA assay (see **Table 4-1**).

4.3.2. Modified SGA protocol shows reproducible results

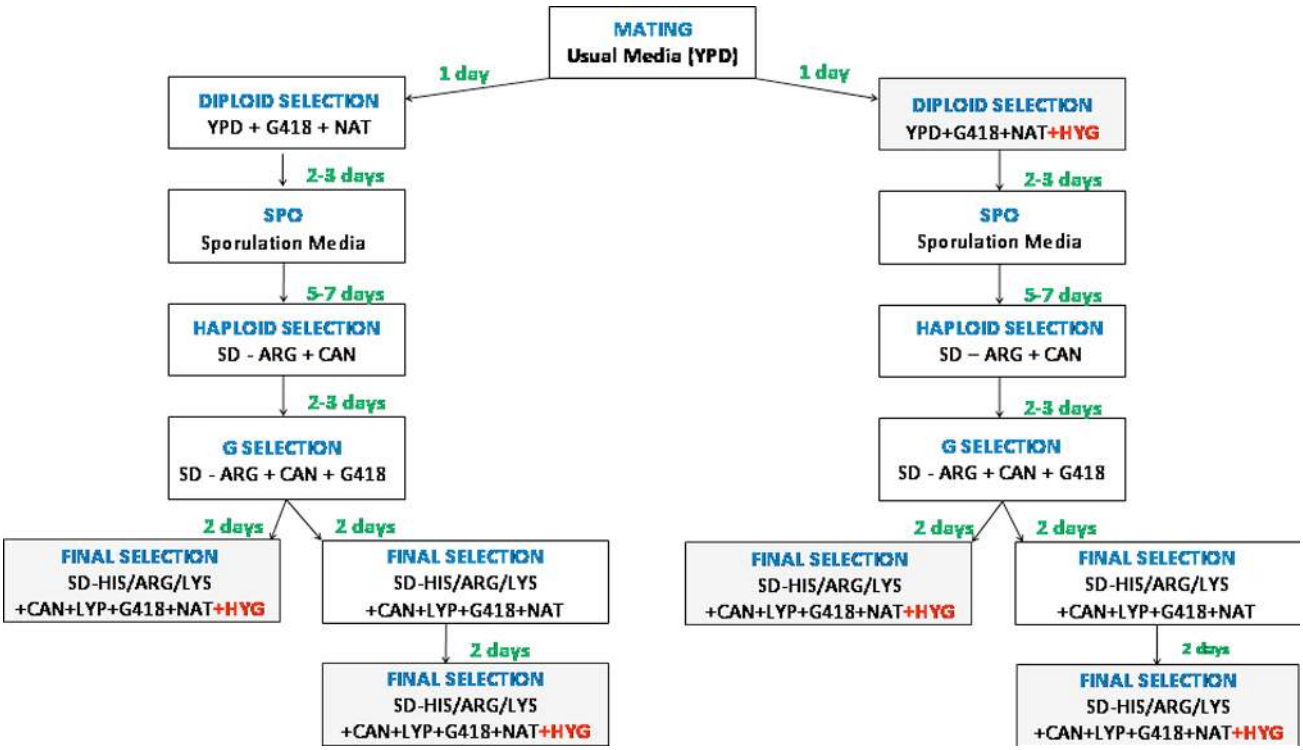
The standard SGA protocol implements 5 pinning steps after mating to select for appropriate haploid double-mutant strains[120]. To determine if additional pinning steps were necessary to select triple mutants, two initial screens were run in quadruplicate with *Hygromycin B* added to selection media (see **Figure 4-4**). Visual inspection of colonies indicated that the optimal protocol involved supplementing diploid selection media with *Hygromycin B*, and adding an additional final pinning step to select specifically for *Hygromycin B* resistance. However, despite having a slightly worsened phenotype when the final *Hygromycin B* selection step was excluded, correlation analysis indicated that the strains reported scores that were still highly similar, specifically in the range of scores that indicates aggravating genetic interactions (see **Figure 4-5**). This suggests that both the procedure and the scoring algorithm are robust to these variations in protocol, as well as demonstrating the reproducibility of screening results. However, to ensure that strains analyzed were at optimal health, all reported screens include the final *Hygromycin B* selection.

Table 4-1 Genes surveyed through triple-deletion SGA

Gene 1	Gene 2	Description 1	Description 2
VHS1	SKS1	Identified as a high-copy suppressor of the synthetic lethality of a sis2 sit4 double mutant, suggesting a role in G1/S phase progression	Involved in the adaptation to low concentrations of glucose independent of the SNF3 regulated pathway
ALK2	ALK1	Protein kinase; phosphorylated in response to DNA damage;	Belongs to the haspin family of kinases; contains a leucine zipper motif; may function in mitosis
CNA1	CMP2	Calcineurin A; one isoform (the other is CMP2) of the catalytic subunit of calcineurin,	Calcineurin A; one isoform (the other is CNA1) of the catalytic subunit of calcineurin,
HAA1	CUP2	Transcriptional activator involved in the transcription of TPO2, HSP30 and other genes encoding membrane stress proteins;	Activates transcription of the metallothionein genes CUP1-1 and CUP1-2 in response to elevated copper concentrations
KIN1	KIN2	Serine/threonine protein kinase involved in regulation of exocytosis;	Serine/threonine protein kinase involved in regulation of exocytosis;
SUT1	SUT2	Transcription factor of the Zn[II]2Cys6 family involved in sterol uptake; involved in induction of hypoxic gene expression	Putative transcription factor; multicopy suppressor of mutations that cause low activity of the cAMP/protein kinase A pathway;
PSK1	PSK2	Coordinately regulates protein synthesis and carbohydrate metabolism and storage in response to a unknown metabolite that reflects nutritional status	Regulates sugar flux and translation in response to an unknown metabolite
KIN82	FPK1	Putative serine/threonine protein kinase, most similar to cyclic nucleotide-dependent protein kinase subfamily and the protein kinase C subfamily	Putative protein kinase that, when overexpressed, interferes with pheromone-induced growth arrest;
NRG2	NRG1	Transcriptional repressor that mediates glucose repression and negatively regulates filamentous growth; has similarity to Nrg1p	mediates glucose repression and negatively regulates a variety of processes including filamentous growth and alkaline pH response

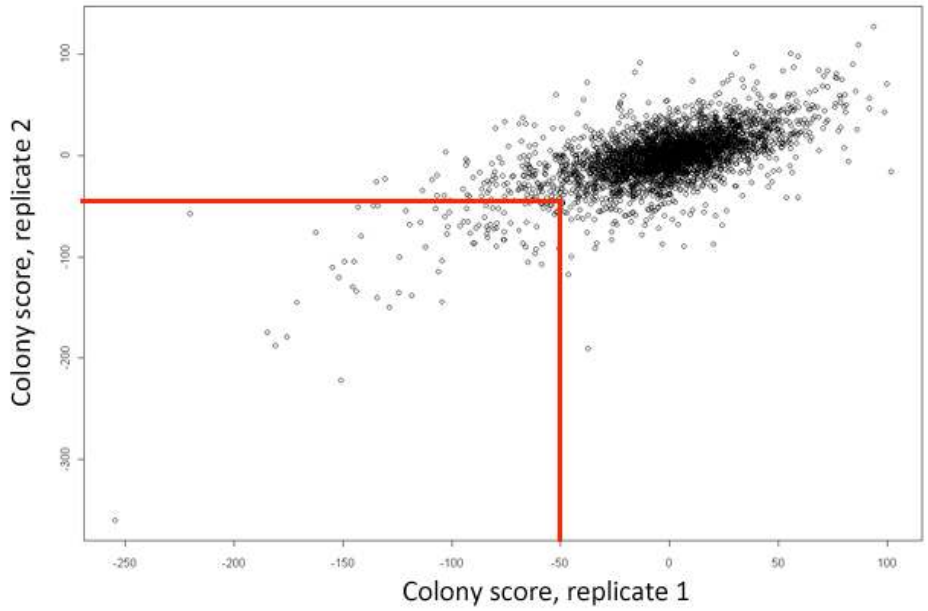
Analysis of triple-deletion SGA focused primarily on non-epistatic WGD-resultant paralog pairs that are annotated as being involved in cell signaling (i.e. kinases and transcription factors). One phosphatase pair was included for comparison, as were two control strains, which consisted of random pairings of genes from the same set.

Figure 4-4 Modifications in SGA protocol



Initial screens were conducted to investigate the effects of adding *Hygromycin B* to both the diploid and final selection media. Analysis of scores indicate that *Hygromycin B* is required in diploid selection media, and that strains should be pinned to media that selects for *Kanamycin* (G418) and *Nourseorthricin* (NAT) resistance before selecting for *Hygromycin B* resistance.

Figure 4-5 Similarity in interactions over replicates



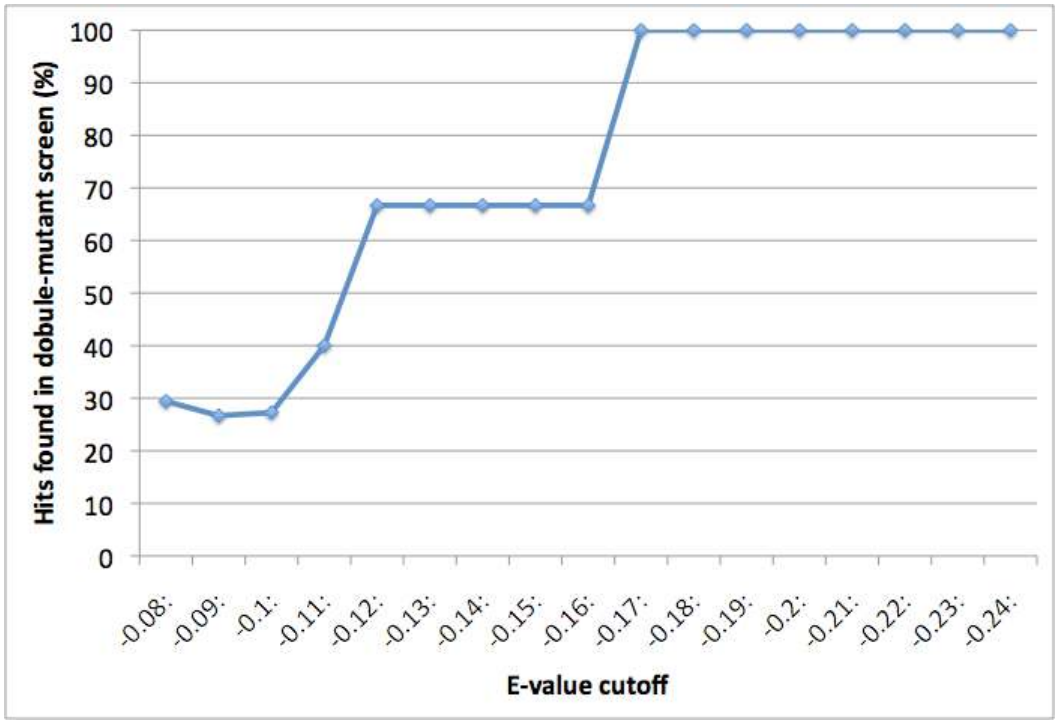
SGA scores resulting from two screens of an identical query strain either pinned directly to *Hygromcyin B*/*Nourseothricin*/*Kanamycin* selection media, or first pinned to *nourseothricin*/*G418* selection. Overall correlation of screens is r=0.65. Red lines indicate the region that would achieve scoring as an aggravating genetic interaction if significant at $p < 0.05$.

4.3.3. Triple-deletion SGA results require increased stringency in identification

At the standard ε-value cutoff score for aggravating genetic interaction using SGA (-0.08; Costanzo *et al*; *submitted for publication*), comparison of double-deletion screens with those of constitutive single deletions revealed an overlap of between 15-20% in shared interactions. While this number initially seems low, it is consistent with expectations based on the false-negative and false-positive rates associated with SGA (precision = 63%, recall=35%; Costanzo *et al*; *submitted for publication*). Further, since true interactions (taken as those detected independently by both single-deletion screens) are more likely to be detected in the double-deletion screen as stringency increases (**Figure 4-6**), the triple-deletion SGA method appears to be detecting valid interactions, albeit with low overlap with constitutive single-deletion screen results. Therefore I assumed that a high level of stringency would be required to confidently determine that a particular interaction was resulting from the triple-deletion and not any underlying pairwise gene combination.

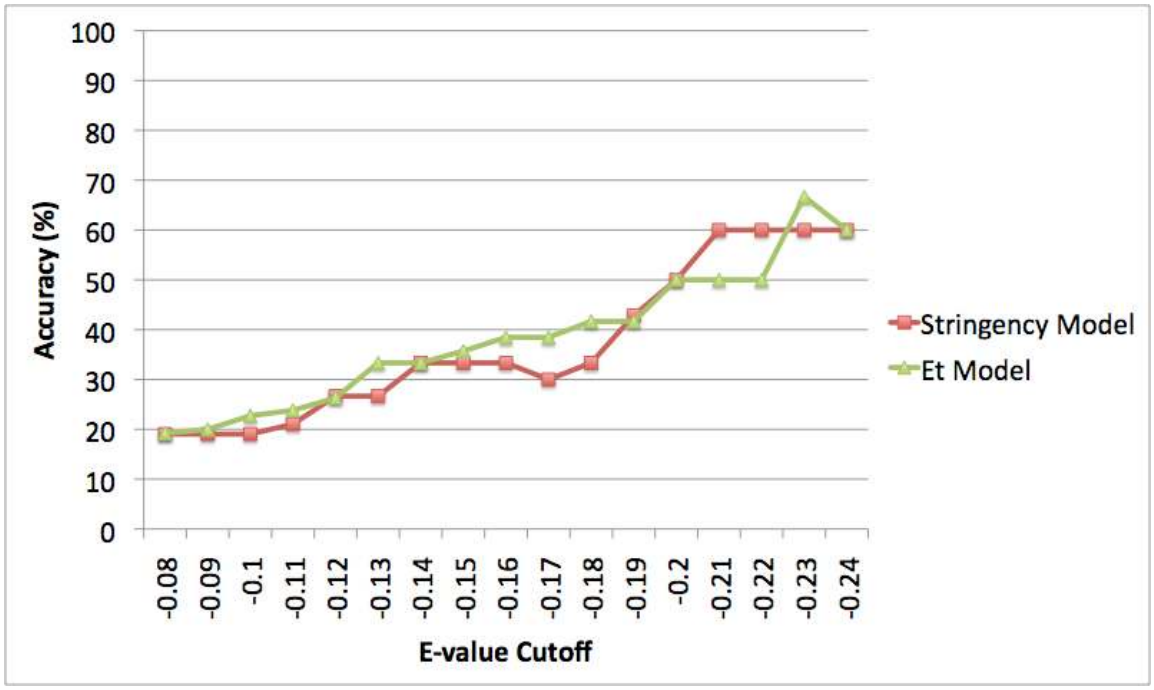
To determine the appropriate level of stringency necessary for confidently assessing trigenic interactions, a number of interactions were confirmed independently using tetrad dissection and results compared. This comparison revealed that bona fide triple-gene interactions were increasingly represented in the SGA screen at higher magnitudes of ε-value cutoff (see **Figure 4-7; red line**), hitting a plateau at approximately 60% accuracy at a cutoff of -0.2 and below. As this accuracy level is similar to that for the published digenic interaction dataset (Costanzo *et al*; *submitted for publication*), I felt that interactions scoring above this cutoff could be confidently

Figure 4-6 Single-deletion hits found in double-mutant screen



The number of 'true' interactions (i.e. those that are detected independently in single-deletion screens) detected in double-deletion screens shown for all 10 surveys. Increasing the cutoff for interaction detection in the single-deletion screens (indicated on the x-axis) results in interactions that occur more frequently in the double-mutant screens.

Figure 4-7 Accuracy of triple-interaction detection



The frequency of successfully detected trigenic interactions (as assessed through comparison with tetrad dissections) is indicated for varying ε-cutoff values. The 'stringency model' (red line) simply depicts the increase in accuracy as the ε-cutoff value increases, however the 'Et model' calculates ε by subtracting scores resulting from single-deletion screening from the double-deletion screen score. Both models give comparable accuracy, however the 'Et model' inherently incorporates the contribution to fitness of all 3 genes involved in the interaction, and is therefore more biologically accurate. Interactions from double-deletion screens having Et values less than -0.23 (at $p<0.05$) were taken to be valid.

reported. However, as presented the current ε model does not incorporate all digenic fitness values and therefore may be subject to error. For this reason a more appropriate score model was tested

To exclude the possibility that reported interactions were due to pairwise gene combinations I fit triple-gene interactions according to the following model:

$$\varepsilon_t = \varepsilon_{ijk} - \varepsilon_{ik} - \varepsilon_{jk}$$

By this definition, the interaction score ε_t is defined for an interaction among a query containing an initial deletion of genes *i* and *j* with potential interactor *k*. By definition of this model only those interactions with a relatively strong trigenic interaction score (as compared to constitutive digenic interactions) would meet a stringent ε-value cutoff. This model produced results similar to the overall ε cutoff model (see **Figure 4-7; green line**), leading to an overall accuracy of approximately 60% at cutoff of 0.23. All trigenic interactions subsequently presented have met this criterion and are thus of confidence comparable to existing SGA digenic data.

4.3.4. Trigenic interactions reveal insight towards overlapping functions

Using the criteria presented above all double-deletion interaction screens were found to have interactions not observable using either individual deletion screen, with degree ranging between 5 and 49 for the 10 surveyed pairs (see **Table 4-2**; also for full list of interactions see **Appendix Table 2**). Notably, while *NRG1*/*NRG2* had 49

Table 4-2 Degree of the surveyed paralog pairs

Gene 1	Gene 2	Group	# Ints
NRG2	NRG1	Transcription Factors	49
KIN82	FLK1	Kinases	32
CNA1	CMP2	Phosphatases	25
ALK2	ALK1	Kinases	19
NRG2	ALK1	Control	13
PSK1	PSK2	Kinases	10
KIN1	KIN2	Kinases	7
VHS1	SKS1	Transcription Factors	6
SUT1	SUT2	Transcription Factors	5

Surveyed double-mutant strains showed varying interactivity as SGA queries. The control pairing of *NRG2/ALK1* showed fewer interactions than the *ALK1/ALK2* and markedly less than the *NRG1/NRG2* screen. The degree of interactions for a double-mutant query could not be predicted based on individual interaction degree or similarity in SGA profile for the starting pair, however those pairs with high degree as double-deletion queries were also generally more co-expressed.

interactions, a control screen that paired *NRG1* with *ALK2* had only 13, emphasizing the redundancy between the bona fide paralogs. The full list of genes interacting with the 10 double-deletion screens (162 total) span a large range of functional categories, but notably contain many (~20) genes of uncharacterized function. This implies that such higher-order deletion screens may aid in further functionally categorizing the genes of *S. cerevisiae*.

While the surveyed set is a small group from which to draw general conclusions, I next sought to determine if any functional properties of the paralog pairs might be predictive of their degree as double-deletion queries. A moderate correlation was subsequently found between similarity in expression measured over multiple conditions[157] and degree ($r=0.61$, $p=0.052$, $n=8$). Further, dividing screened pairs into groups based on degree (low degree group having 10 interactions or less and a high degree group with more than 25) showed differences in expression similarity to be statistically significant ($p<0.005$). There were no appreciable differences in conservation (as assessed through shared sequence similarity) for low and high degree paralog pairs.

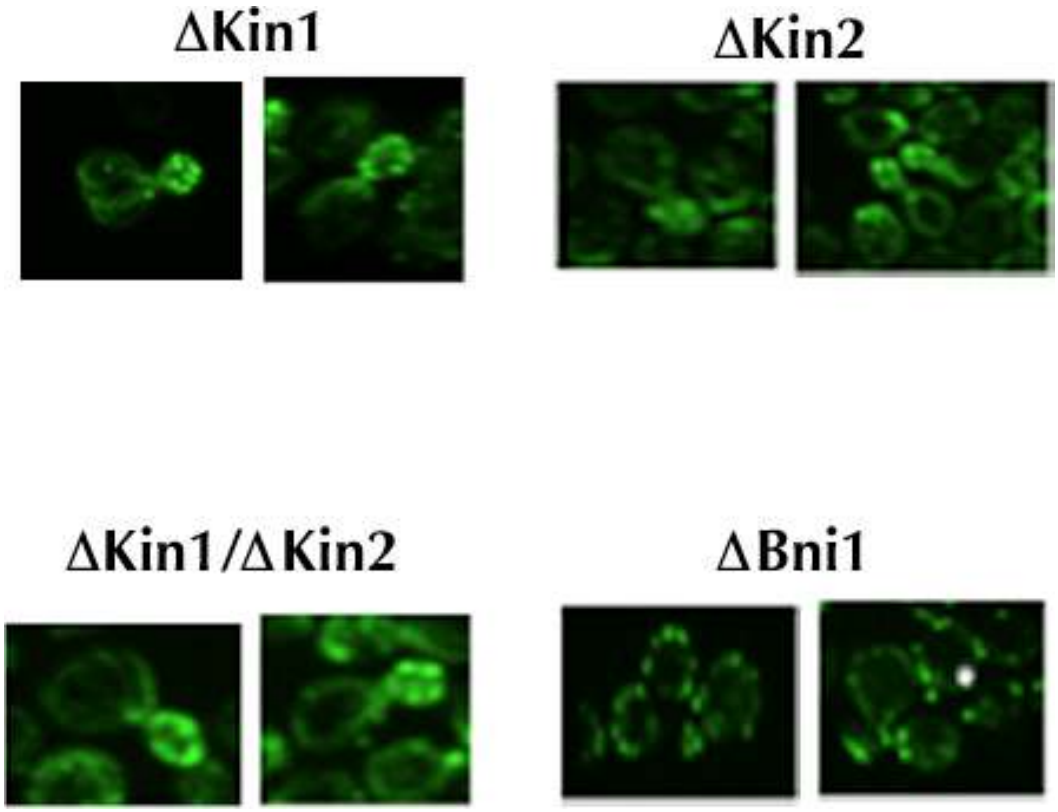
I next examined specific paralog pairs to determine if results from the double-deletion screens re-capitulated what was known about their function. As mentioned briefly above, *NRG1* and *NRG2* are transcriptional repressors that negatively regulate filamentous growth and are involved in alkaline pH response[188]. In keeping with their known cellular role, double-mutant hits were significantly enriched for genes annotated as being involved in 'cell differentiation' and 'response to pH' (neither single-mutant screen produced hits with the same functional enrichment). Also, the double-mutant uniquely showed interactions with the Rim complex, members of which are known to

repress *NRG1* during alkaline stress[189] (interactions with *Rim101*, *Rim20* and *Rim9* confirmed via tetrad dissection). Further, *SNF1*, a known antagonist of *NRG1* and *NRG2*[190] shows a strong anti-correlation in SGA profile with the *NRG1/NRG2* double-deletion screen (r=-0.2). Taken together these results re-iterate that NRG1/NRG2 act in concert to respond to changes in pH, and that this may have been the function of the single ancestral gene.

Another paralog pair, *KIN1*/*KIN2* are serine/threonine protein kinases involved in regulation of exocytosis[191]. While neither individual gene shows a noticeable phenotype upon deletion, deletion of the single ortholog of these two genes in *S. pombe* causes a distinct lack of cell polarity[192]. Two hits resulting from the *KIN1/KIN2* double-deletion screen also have known involvement in cell polarity (*BEM4*, *SLA1*), suggesting a redundant function for *KIN1/KIN2* in this regard. However, fluorescent actin staining revealed that there is no noticeable polarity defect in the double-deletion mutant (see **Figure 4-8**), suggesting that other genes may buffer the activity of *KIN1* and *KIN2*.

Lastly, the paralogs *CNA1* and *CMP2* (calcineurin) have well described functions related to their de-phosphorylation activity, participating in several biological processes. Appropriately the individual SGA screens for *CNA1* and *CMP2* correlate with a group of genes spanning multiple functions. Alternatively however, the double-deletion screen correlated mainly with genes that are enriched for nuclear transport. This provides (albeit speculative) evidence that ablation of this combination of catalytic subunits disrupts the nuclear-related function of calcineurin (the main target of calcineurin is a transcription

Figure 4-8 KIN1/KIN2 double deletion does not effect polarity



Images displayed are of haploid deletion mutants following phalloidin staining of actin. *KIN1*, *KIN2* and combined *KIN1/KIN2* deletion mutants show normal cell polarity (strands of actin spanning larger cells, localized patches in budded cells). In contrast, the *BNI1* deletion mutant strain has known polarity defects and shows only localized globules of actin.

factor that is targeted to the nucleus upon de-phosphorylation and modulates ion homeostasis[193,194]). As this relationship had been initially masked in single-deletion screens, it may represent the shared, ancestral function of these two genes.

4.4. Discussion

Although requiring an extra degree of rigor, these results indicate that inclusion of *Hygromycin B* as an additional marker is a viable avenue to pursue SGA screening of double-deletion query strains, and that trigenic interactions can subsequently be confidently identified. I submit that this protocol may have efficacy beyond the study of paralog pairs, for example in the determination of the function of heterodimers, isozymes, or any other genes with demonstratable functional redundancy.

Of the 10 pairs assayed through triple-deletion SGA, some paralog pairs proved particularly amenable to the process, notably *NRG1/NRG2* (which had nearly 50 unique interactions) and *CNA1/CMP2* (which showed functionally insightful correlations with previously conducted single-deletion screens). Although speculative, there appears to be very little regarding the properties of paralog that suggests *a priori* whether they may be more amenable to triple-deletion SGA. Notably however, results were generally less informative for some kinase paralog pairs, suggesting that they may be better surveyed for redundancy using alternate methods. Specifically, the kinase pair *ALK1/ALK2* had a degree that was only slightly greater than expected based on the *ALK1/NRG2* control pairing (19 versus 13), and while the paralog pair *KIN82/FLK1* had many interactions, it was hard to discern a unified redundant function. However, the demonstration that a small group of pairs with generally low sequence conservation and minimal overlap in

SGA profile as single deletion queries could demonstrate robust, unique interactions as triple mutants suggests that elements of redundancy may be pervasive and re-iterates the notion that functional redundancy among paralog pairs can not simply be predicted based on conservation. Further, the correlation between expression similarity in alternate conditions and degree as double-mutant queries for paralog pairs suggests that this method is useful in detecting condition-specific relationships, and I believe, would be useful in further investigating condition-specific epistasis among the remaining non-epistatic paralogs (see **Future Directions**).

My previous work had ultimately suggested that observation of epistasis may be incomplete in the limited conditions typically used for functional assay, and further that some functional categories of genes may be artificially under-represented for genetic interaction[123]. As all surveyed pairs had no existing evidence of epistasis, these results ultimately demonstrate the possibility that lack of similarity among some paralog pairs in SGA screening may be due to masked interaction, and further that detecting phenotypic manifestations of redundancy can require novel approaches.

Chapter 5

Thesis summary and future directions

5.1. Thesis summary

WGD events facilitate the acquisition of novel function by providing a tremendous and unique source of genomic material that is ultimately used to bypass normal evolutionary constraints. Since evidence does not tend to indicate that exact functional redundancy can be maintained between duplicates over long spans of evolutionary time, several models have been presented to describe the stable fixation of paralogous genes following their initial duplication. The two most popular models basically prescribe either the procurement of novel function through random mutation, or the partitioning of ancestral function, typically through alterations in patterns of expression. While these models use sporadic examples to illustrate their claims, systematic surveys of the overlapping function of paralogs have been lacking. Further, while the advantages of increased dosage may facilitate initial redundancy among WGD-resultant duplicates[195], logically some additional mechanism should facilitate the long-term preservation of functional overlap. To this end, recent observation of 'transcriptional back-up' mechanisms suggests that some element of retained redundancy is advantageous, contributing to the robustness of the genome towards mutation and potentially facilitating adaptation to alternate environmental states. With this work I sought to gain insight into the nature and extent of functional overlap among a large body of extant duplicates with the ultimate goal of discerning if functional redundancy is pervasive enough to suggest a selectable advantage.

Using two large physical interaction datasets (and subsequently several other datasets of comparable accuracy but generated using different experimental means) I demonstrated that the 457 WGD-resultant paralogs described by Kellis *et al*[17] had

substantially retained shared interactions. However, considering the nature of the result I am hesitant to state which model of functional divergence is supported. While the neo-functionalization model can be interpreted as suggesting complete functional dispersal, many clear instances of neo-functionalized paralogs still retain enough similarity as to potentially share physical associations. For example, mammalian Retinoic Acid Receptors (RARs) have clear evidence of neo-functionalization in both their targeting and binding sequences, however all still share interactions with a common substrate, suggesting retained functionality[196]. Similarly, while the sub-functionalization model dictates functional partitioning as opposed to dispersal, it is still entirely possible under the DDC model that even highly similar paralogous genes would not share interactions or complex membership (i.e. if they are expressed only under alternate circumstances).

Therefore while my results regarding shared interactions do appear superficially to support limited divergence in function for WGD paralogs, either model (or combination therein) may have been initially active. Ultimately data regarding the ancestral function is needed to truly resolve the issue of the dominant mode of retention, and at best we have only species variants separated by over 100 million years of evolution for comparison (see **Future Directions** below). However, regardless of whether or not they have divided the ancestral function, the extent of shared physical associations does suggest that some element of redundancy can be maintained over long spans of evolutionary time. Furthermore, variances in expression among co-complexed gene products lend support to the existence of transcriptional back-up circuitry as a means to maintain functional overlap.

Conceptually, maintaining variances in expression level could be a useful mechanism to compensate for ablation of a duplicate, however this introduces the question of why every other essential gene does not have a similar mechanism. An alternate explanation for the decreased expression of one paralog is that the low-abundance gene is not required for growth in surveyed condition, but reserves expression to help accommodate growth in alternate states. As the various post-WGD species would have encountered different environmental and selective forces following speciation, this could potentially explain why these species have retained alternate sets of duplicates. To test both these hypothesis I examined the extent of aggravating genetic interactions between surveyable WGD-resultant paralogs in both standard and stress conditions.

Assaying for aggravating genetic interactions between the 399 surveyed paralog pairs produced two significant findings. First, the frequency of genetic interactions was found to be higher than that previously noted for genes with identical functional annotation[120], implying that duplicated genes have a propensity for epistasis beyond that which could be explained by their shared function. Furthermore, the observation of some instances of epistasis solely in the assayed stress conditions suggested that epistatic relationships may be uniquely observable using additional stressors (approximately one-quarter of non-epistatic paralog pairs are of unknown function and thus may only have noticeable activity in an alternate state). These observations lend anecdotal support to the presence of transcriptional back-up mechanisms, although there is no direct evidence of their existence (more in **Future Directions** below) and confirm that instances of redundancy may be specifically preserved to accommodate growth in non-ideal cellular conditions. It is also worth mentioning that the function of those genes deemed epistatic

in alternate conditions were inherently related to the tested condition (e.g. epistasis among sodium transporters in high salt). Given the inherent relationship between observation of epistasis, experimental condition, and gene function, I am therefore hesitant to speculate at the extensibility of these findings to duplicates of other origin (specifically SSD-resultant duplicates which are known to be functionally distinct from those arising via WGD). I ultimately submit that SSD-resultant duplicates deserve independent consideration, and should be analyzed separately in future determinations of paralog evolution.

The second notable observation stemming from this line of experimentation was that epistatic paralog pairs could not generally be shown to have more shared functional overlap (as gauged by physical interactions) than comparable non-epistatic paralogs. Recently Ihmels *et al* showed that yeast paralog pairs have low overlap in genetic interaction profiles, suggesting that they may be capable of functional compensation without any obvious redundancy[154]. However, an alternate explanation is that not all epistatic relationships between paralogs have been revealed. To test this assertion I performed SGA screening of double-deletion queries, comparing results to the constitutive single-deletion strains. I focused experimentation on a small subset of paralogs that had low similarity in SGA profile in order to directly test the hypothesis that low overlap in SGA profile did not necessarily correspond to lack of redundancy. Notably, not all surveyed paralog pairs displayed a high degree of trigenic interactions, implying either that redundancy is ultimately not pervasive among extant duplicates, or that the triple-deletion SGA method was not appropriate to screen certain pairs. Potential alternate screening methods are discussed below (see **Future Directions**). However,

those paralog pairs with the most trigenic interactions (and thus arguably with the most redundant function) also displayed a notable correlation in expression across multiple conditions, seemingly supporting my assertion that redundancy may be retained to facilitate compensation in alternate environmental states.

The central theme of this thesis has been to analyze the nature of functional overlap among duplicated genes, and ultimately to determine whether retained redundancy among duplicates is pervasive enough to suggest a selectable advantage. Ultimately I present three findings that I believe to be substantial: (i) functional overlap is prevalent among WGD-resultant paralog pairs, (ii) compensatory mechanisms are also widespread among this same group of paralogs, even more so than would be predicted based on their strong functional associations, and (iii) functional compensation can not necessarily be predicted based on the conservation of duplicated pairs; direct assay of function is required. While results do strongly imply adaptation to alternate environmental conditions as a mechanism for the retention of functional overlap among gene duplicates, additional experiments would be required to prove this directly. Future experimental approaches that may be useful not only in answering this question but also in further characterizing the extent of functional overlap among gene duplicates are described below.

5.2. Future directions

5.2.1. Using further large-scale screening to identify functional overlap

5.2.1.1. Increased breadth of triple-deletion SGA screening

Increasing the number of WGD-resultant paralog pairs assayed through triple-deletion SGA would be useful on several fronts. First, the larger body of experimental data would provide a greater means with which to identify batch effects and other artifacts that arise not only from the addition of *Hygromycin B*, but also that may be generally associated with multiple gene deletions. Further, screening an additional bevy of control strains (meaning randomly paired genes) would allow better characterization of the range of interactions expected from non-redundant pairs, and therefore form a more appropriate reference against which I could identify true functional overlap. Systematic SGA screening would also allow more concrete identification of alleviating interactions, which may be specifically useful in determining the overlapping function of paralogs within pathways or complexes. Lastly, as the surveyed genes were selected based on annotation as kinases and transcription factors it would be worthwhile to use the triple-deletion SGA method to explore interactions among both the varying functional categories and pairs of unknown function. Of the 450 paralog pairs described by Kellis *et al*[17], approximately half have shown a notable growth defect either through my work or by DeLuna *et al*[124], or Dean *et al*[173]. The remaining paralog pairs should be thoroughly investigated using triple-deletion SGA. Condition space could also be further increased through chemogenomics[197].

In addition to screening for functional overlap among paralogous genes, triple-deletion SGA may be useful in understanding differences in epistasis across multiple

species. Evidence recently collected suggests that the majority of gene pairs found to be epistatic in *S. cerevisiae* (~70%) show no interaction in the distantly-related species *Schizosaccharomyces pombe* (*S. pombe*)[198]. Assay of orthologs representing pairs epistatic in the *S. pombe* strain but not in *S. cerevisiae* as double-deletion queries will determine if their combined loss of function effect was buffered by additional genes (possibly by genes retained uniquely in *S. cerevisiae* following the WGD event).

5.2.1.2. Determining functional overlap using other SGA-based techniques

While the synthetic lethal interactions screened via SGA identified many useful relationships between duplicated genes, ultimately there may exist alternate experimental means that may be helpful in characterizing functional overlap of yeast paralogs. Specifically, as these experiments have shown that genes functioning in growth-related processes have the most noticeable growth defects, perhaps assay of additional phenotypes would reveal functional relationships for those genes involved in cell signaling that were notably deficient for aggravating genetic interactions. Furthermore, experimentation performed focused on loss-of-function interactions, while contribution to phenotypes resulting from gain-of-function interactions went unexplored. For these reasons there are several large-scale assays that I feel would be particularly insightful in establishing redundant function for the remaining non-epistatic duplicated genes. Each of these methods adapts the SGA process and therefore would require very little in terms of query strain modifications.

Growth of microbial strains is often taken to be indicative of the health of an organism, however the possibility exists even for normally growing strains that there is

some other manifested phenotype that could otherwise alter fitness. The use of fluorescence microscopy to create and process morphological cell images on a large scale (often referred to as High Content Screening, or HCS[199]) can be applied to quantifiably characterize strain to strain variations in size/shape, cytoskeletal formation, and nuclear morphology[200]. Further, analysis indicates that functionally-related genes cause similar morphological differences upon deletion, and recent work suggests that high-throughput microscopy can be combined with the SGA pinning procedure to quantify such phenotypes on a genome-wide scale[201]. Therefore just as defects in growth had been used to derive an expectation for double-deletion fitness and subsequently assess epistasis based on a multiplicative growth model, any quantifiable phenotype could be used in the same way. HCS data suggest that approximately half of all deletion mutants have a corresponding quantifiable phenotypic defect[200], how these phenotypes may be buffered by duplicates remains to be explored.

Similar to the systematic gene deletion studies determining that approximately 20% of yeast genes are essential for cell viability[52] (i.e. have lethal phenotypes), over-expression surveys have found reduced growth among 15% of genes when transcript abundance is increased (gain of function phenotypes)[117]. Notably, many of the genes with gain of function phenotypes are involved in cellular signaling, suggesting a potential to be informative for the group of paralogs not found to be epistatic through RSA or GCA screening (see *Chapter 3*). Exploitation of the SGA platform has allowed systematic introduction of deletion mutants into strains systematically over-expressing yeast ORFs to facilitate large-scale analysis of interactions occurring upon over-expression of one gene and decreased expression of another (synthetic dosage lethality; SDL) and has been

particularly informative in mapping kinase-substrate interactions[202]. Analysis of the profiles corresponding to queries of paralogous kinases may indicate shared substrates and by extension overlapping function. Further, utilizing mutants containing double-deletions of paralogous kinase pairs and comparing to related single over-expression profiles (as in *Chapter 4*) may indicate to what extent paralogous kinases share substrate specificity, thus identifying potential redundancies among this large body of paralogs.

Lastly, two-color receptor screening developed by the Andrews lab at the University of Toronto utilizes the SGA deletion collection to assay responsiveness of genes to a promoter sequence of interest[203]. Basically, a query strain containing a target promoter sequence driving GFP expression as well as a reference, constitutively active promoter driving another fluorescent marker is mated to the deletion array. Following appropriate selection, the ratio of GFP to RFP fluorescence is used to determine the interactions between the various deleted genes and the promoter sequence. Investigation of the promoter sequences of paralogs could indicate the potential for duplicates to be co-expressed, as shared activation by transcription factors may indicate potential co-expression in alternate conditions.

5.2.2. Assaying for the presence of transcriptional back-up mechanisms

As mentioned above the high frequency of epistasis among WGD-resultant paralogs implies that presence of a duplicate largely buffers phenotypic effects associated with their deletion. The proposed mechanism for this occurrence involves modification of transcription, increasing the expression of a paralog upon deletion or inhibition of its sister gene. Naively, one might think this could be tested simply by monitoring the

expression of a given gene in a strain harboring a deletion of the corresponding paralog, however these changes are rarely observed in practice[204], potentially due to a diluted effect at time of assay (presumably because strains will have grown for generations with this deletion and the initial effect may be transient). A more direct means to test for the existence of transcriptional back-up mechanisms would be to replace the promoter sequence of a given paralog (i.e. give the paralog consistent expression), then subsequently delete the ORF of the sister gene. If the ORF deletion does not produce a phenotype with abnormal growth, then it is unlikely that the additional paralog is compensating through altered transcriptional activity.

Similarly, a simple switch of promoter sequences for epistatic paralogs may indicate if they have evolved according to the DDC model. If two genes are highly similar (i.e. varying only in their expression) switching their promoter sequences should switch their expression patterns, thus determining if one each gene is still capable of performing the other's function (i.e. that they have partitioned the ancestral function through altered expression). Assay of interactions should indicate whether the duplicates have substituted for one another.

5.2.3. Determining the properties of paralogs in other species

5.2.3.1. Examining the ancestral copies of epistatic paralogs

While experimental data can be used to speculate as to the nature of functional dispersal following duplication, ultimately these assertions cannot be proven without knowledge of the function of the ancestral gene. Hypothetically, obtaining the ancestral species would facilitate determination of function (i.e. through physical associations, and

co-expression, etc) and subsequent comparison with extant species could indicate what aspects have been retained by the corresponding paralogs. Also, knowledge of ancestral gene function could answer the question of whether essential genes in the ancestral species initially encoded paralogous gene pairs displaying an aggravating genetic interaction, and subsequently whether duplication is truly reserved for genes of limited functional importance (as suggested by He and Zhang[95]).

While no pre-WGD strain exists today, Kellis *et al*[17] established that *K. waltii* diverged from the *Saccharomyces* lineage soon before the WGD event making it one of the closest available proxies to the ancestral pre-WGD species. Obviously using *K. waltii* as the ancestral proxy poses some limitations as both *K. waltii* and *S. cerevisiae* have evolved for many millions of years since their respective speciation events, however these strains represent our closest available experimental approximations. Unfortunately there are also some potential technical difficulties, notably it is unknown how efficiently homologous recombination would occur in *K. waltii*, making creation of mutant strains difficult. Luckily, *K. waltii* has a characterized plasmid (analogous to the *S. cerevisiae* 2μ plasmid) which can be used to introduce foreign DNA sequences[205]. Further, *K. lactis* and *S. Kluyveri* (pre-WGD species having approximately the same divergence time from *S. cerevisiae* as *K. waltii*) also have available knockout protocols. Therefore analysis of both ancestral over-expression and deletion phenotypes would be possible, and could be combined with SDL (mentioned above) or SGA data to indicate what morphological defects may have been inherited from the ancestral gene, and which are buffered through duplication.

5.2.3.2. Conservation of epistatic relationships among post-WGD yeast species

While retention of duplicates in multiple post-WGD yeast species has been examined[36], it is less clear whether epistatic relationships among duplicates are maintained across species. Therefore, determining whether the same duplicate pair that contributes jointly to a phenotype in *S. cerevisiae*, does so in other species would indicate for the first time the extensibility of findings regarding the functional overlap of duplicated genes toward other species. Also, akin to determination of gene essentiality across species, investigation of epistasis may indicate how the various post-WGD species have modified their genetic network to accommodate adaptation to their unique environments.

Of the 5 post-WGD yeast species described in the YGOB (*S. cerevisiae, S. bayanus, C. glabrata, S. castelli* and *K. polysporus*) 2 have been shown to be amenable to gene deletion (*S. bayanus*[36] and *C. glabrata*[206]). Also, the recent discovery of inhibitory RNA (RNAi) in budding yeast (notably *S. castelli*[207]) introduces the possibility for gene knockdown assay not only in this species, but in other yeast not typically assayable through gene deletion assay (RNAi is a term generally used to describe a system in which short RNA sequences silence transcripts through complementary binding[208,209]). In addition to analysis of the conservation of epistasis across species, it would be worthwhile to determine whether genes existing in single-copy post-WGD in other species share a deletion phenotype consistent with double-deletions of paralogs in *S. cerevisiae*.

5.2.3.3. The impact of WGD on robustness of the human genome

At the time of my comparison of the protein complex membership of extant duplicates in *S. cerevisiae* (**Chapter 2**) there were scant interaction data in other species (notably human), and that which did exist focused on particular genes or processes and was therefore potentially biased. Further, there was limited, contested evidence of the presence of WGD-resultant paralogs in the human genome, so a confident list of duplicates could not be determined. Today interaction data are far more common for mammalian species (as of this writing the Human Protein Reference Database contains 38,806 human physical interactions; http://www.hprd.org), potentially facilitating re-visitation of functional overlap analysis using human duplicates. Further, copious amounts of expression data collected both in healthy tissues and in disease states (for example as found in the Gene Expression Omnibus; http://www.ncbi.nlm.nih.gov/geo/) would allow comparisons of the varying expression patterns of human duplicates, and subsequently determination of whether the noted disparity in expression for co-complexed yeast duplicates holds true in human. Also, sequencing of additional genomes has facilitated the determination of a set of human WGD-resultant paralogs[22,41], which could ultimately be used for such a comparison.

In addition to surveying function through analysis of expression and physical interactions, recent advances in the field of human gene inhibition have made the study of human genetic interactions more feasible than ever. Analysis of the genetic interactions of human duplicates would allow determination of the contribution of duplicates towards genomic robustness in human, which may in turn help elucidate susceptibility toward diseases with a genetic component.

While ablation of gene function via deletion is still not nearly as practical in cultured human cells as in yeast, the recent emergence of RNAi represents a stable, reproducible knockdown procedure. Creation of RNAi-containing vector libraries has facilitated the studying of cancer processes through large-scale gene disruption[210,211], and has spurred an initiative to construct a genome-wide library containing RNAi in lentiviral vectors capable of transfecting non-dividing cells[212]. While analysis in the nematode *C. elegans* indicates that genetic interactions are observable through the use of RNAi[213], and while analysis in mammalian cells demonstrates that both large-scale knockdown analysis through pooled RNAi libraries[212] and combinatorial addition of RNAi in human cancer cell lines is possible[214,215], systematic studies of double-gene knockdown are lacking. Paralogs could be selected for assay based on expression in a given cell type, and analyzed for interaction via addition of appropriate combinations of RNAi.

Appendix

Portions of this chapter have been reprinted or adapted from *Musso *et al*[103]

Generation of experimental data

The goal of any proteome-scale association assay is high-quality interaction data. The *S. cerevisiae* (budding yeast) proteome (which is relatively small in comparison to mammalian systems) has been investigated experimentally for over thirty years using low-throughput means[216,217]. Appropriately, the long-beheld gold-standard in protein complexes in yeast was generally regarded to be the curated set stored in the Munich Information center for Protein Sequences (MIPS)[104] database, which contains experimentally well-characterized protein complexes generated through low-throughput assay. However, increasingly comprehensive lists of protein interactions were only recently generated with the application of high-throughput assays such as Tandem Affinity Purification (TAP)[218] and Yeast-2-Hybrid (Y2H)[125] screening. Indeed, two recent global studies of yeast protein complexes published in 2006 by Gavin *et al*[48] and Krogan *et al*[49] each predicted the existence of over 350 alternate groupings of proteins based on clustering of the physical interaction data. Yet while high-resolution interaction detection methods may avoid some of the problems, such as the often high false-positive and false-negative rates, associated with their high-throughput counterparts[138,219], these low throughput interaction methods are not practical for proteome-scale studies.

Yeast-2-Hybrid

Y2H was first developed in the late 1980's[125] as a generalizable and highly sensitive method to screen for interactions among binary pairs of proteins, and is still frequently used both as a first pass screening tool and for genome-scale exploratory studies today[220-222]. Despite several design variations since its inception which have

resulted in improved assay efficiency[223-225], the basic principle of the Y2H assay remains the same. That being that Y2H takes advantage of the fact that the process of transcriptional activation (and thus expression of a suitable reporter gene) depends on the tethering of two distinct protein domains to a target promoter: first, a DNA-binding domain (BD), that binds to the upstream DNA element and, second, an activation domain (AD) that interacts with the general RNA polymerase machinery. In order to determine if an interaction occurs between two proteins, such as *x* and *y* (where *x* represents the bait protein, and *y* the possible interactor, or prey), protein x is expressed as a fusion to the DNA-binding domain, while the activation domain is likewise fused to protein *y*. If the re-engineered proteins are co-expressed and subsequently interact in the yeast nucleus, the jointly linked BD and AD will reconstitute an activator, leading to expression of a selectable reporter gene[125]. The presence or absence of a binary interaction can then be monitored on a large-scale by screening thousands of strains for the activated expression of selectable markers and by following the growth properties of viable yeast colonies. High throughput adaptability can be further enhanced by mating ordered arrays of yeast strains encoding distinct bait and prey in a 96-spot format[226].

The first adaptation of the Y2H method to genome interaction mapping was reported for the T7 *E.coli* bacteriophage, in which 25 interactions were identified among ~50 proteins[227]. The implications of this pioneering study were that the Y2H method could be applied to study interactions among the components encoded by a complex biological system or even an entire genome. In rapid succession, the interaction networks of even more complex organisms were surveyed over the course of the next few years using the Y2H method. In 2001, two separate initiatives compiled Y2H-based global

interaction maps in yeast[102,228], noting a combined set of 4000 putative PPIs. In 2004, an initial first-pass proteome-wide map was reported for the nematode *Caenorhabditis elegans*[229], the first study of its kind for a multi-cellular organism. Y2H screens have also been conducted in even more complex systems such as *Drosophila melanogaster* (fruit fly)[230,231], and most recently (although the data can be considered still preliminary), for human cells[232,233].

A major advantage of Y2H is that reformation of the transcription factor complex used to detect interactions can occur when assayed proteins only transiently interact[234], whereas comparable affinity-based purification methods (discussed below) have difficulty detecting transient interactions[235]. A major disadvantage of Y2H, however, is related to the often-elevated error rates. Analysis of large-scale datasets generated through Y2H tends to reveal low experimental overlap[138,228,236]. The most likely explanation for the lack of correlation between these two studies is a combination of both a high false-positive rate (estimated to be anywhere from 50%[138] to as to high as 90%[228,236]) and false-negative rate, wherein most biologically relevant interactions are presumed to be missed. These artifacts stem in part from the over-expression and forced co-localization of the candidate proteins in the yeast nucleus, leading to non-physiological context[223]. Consequently, while Y2H results are seen as a positive indication of a genuine protein interaction, the predictions benefit from additional supporting evidence.

Further limiting the applicability of Y2H in non-model organisms is its inability to survey interactions for gene products with incompletely defined coding sequences, as an appropriate vector must be created for each query protein containing the associated

gene and marker. This aspect limits implementation of comprehensive screens for mammalian systems where alternative splicing and incomplete knowledge of exons is common. Moreover, as Y2H is based on a binary interaction assay, it neglects interactions that involve 3 or more proteins[236], which is precisely the hallmark of many, if not most, cellular protein complexes.

Due to the fact that interacting proteins must co-localize to the nucleus, Y2H has also traditionally not been useful for surveying interactions among integral membrane proteins. However, a specialized variant of Y2H, the so-called split-ubiquitin assay[237], has been developed to tackle this missed opportunity and has shown increasing promise in recent years[238]. Briefly, one trans-membrane (TM)-domain containing protein is fused to the N-terminal half of ubiquitin, while the second TM protein is fused to the C-terminal half of ubiquitin and an adjoined transcription factor. Interaction of these proteins causes recognition of the complex by ubiquitin-recognizing proteases, thus releasing the transcription factor from its membrane anchor and thereby allowing subsequent activation of a reporter gene. Application of this method resulted in the identification of nearly 2000 total interactions involving 536 TM-bearing proteins in yeast (131 of the 2000 interactions were deemed to be high-quality based on a series of stringent criteria[238]).

Like most other Y2H-derived methods, this assay involves constitutively over-expressing the bait protein, often resulting in elevated (non-native) protein concentrations. Hence, the interactions captured by this approach may not occur at physiologic protein conditions, contributing to a high false-positive rate. Yet recent optimizations, such as integrating the tags into the target genome to achieve near-native

expressions, may further the applicability of this assay for investigating the physical makeup of otherwise scantly-characterized membrane-associated biochemical pathways.

Due to their ease of execution and scalability, Y2H-based binary assays remain prevalent in large-scale PPI surveys for many organisms despite their often-high associated error rates. As computational methods for data evaluation improve, biologists are becoming more adroit at reducing the number of false-positives, thereby increasing the practical utility of such methods in establishing probable protein-protein interactions. However, the true pitfall when using Y2H data alone when trying to deduce the subunit composition of protein complexes is the generally high false-negative rate, which results in sparse representation of the overall biological networks of interactions, and consequently, poor assessment of discrete biological modules. Several of the large-scale Y2H screens have been shown to result in a network topology thought to be inconsistent with true biological systems[101] (more about biological interaction networks and graph theory methods commonly applied to analyze them below).

Affinity purification

In an effort to study protein complexes specifically, and to circumvent the inherent false-negative and false-positive rate of Y2H, affinity purification was developed for large-scale interaction surveys. The underlying concept behind affinity purification is a consequence of what had been observed in biochemical and co-immunoprecipitation studies for decades[239]: by selectively retrieving a protein of interest from a cell extract through the use of a specific ligand or antibody, proteins stably bound to the query protein can usually be concomitantly retrieved. In affinity purification

studies, a universal epitope tag[240] is often systemically attached to the query proteins of interest which allows for routine bait capture along with any associated interactors via a single, well-defined and often commercially available tag-specific antibody. Proteins bound to the query protein are then usually identified through mass spectrometry, using either traditional gel-based methods or gel-free tandem mass spectrometry procedures (the former offers qualitative information regarding subunit stoichiometry, while the latter provides superior sensitivity).

Affinity purification offers three distinct advantages over Y2H methods. First, only the query proteins require tagging, allowing novel interactions to be discovered between the baits and one or more poorly characterized proteins. Second, entire protein complexes can be captured during a single purification, as opposed to the binary interaction format employed by Y2H. Third, while the purification procedure can be tedious to scale-up, the need to tag only one or two proteins in order to define a given complex reduces the number of experiments that need to be performed to achieve good proteomic coverage, compared to the multiple pair-wise permutations ($n \times n$ experiments) required in a Y2H screen.

Using a systematic method of affinity purification coupled with mass spectrometry, the first interaction map for yeast was published in 2002[241]. Reflecting a substantive increase in proteome coverage, the group released an interaction map of higher density than previous comparable-scale studies, consisting of 3617 putative protein-protein interactions for 493 tagged bait proteins. Importantly, the authors reported approximately 3-fold more interactions per protein which were curated in protein complex databases, suggesting a decreased false-negative rate[241]. However,

while seemingly more accurate, the results of affinity purification studies have the unfortunate disadvantage of being biased against detection of low abundance proteins, with results dominated by higher-abundance bait proteins and often many spurious interactions resulting from common 'housekeeping' contaminants.

Tandem affinity purification

In an effort to increase the sensitivity of affinity-purification to low-concentration proteins, the affinity purification process was further refined as Tandem-Affinity-Purification[218]. The principle behind the TAP procedure is to retrieve proteins bound to epitope-tagged proteins of interest through two successive steps of affinity chromatography: first, generally via binding of the tagged protein to IgG beads, second via attachment to calmodulin (or an alternative affinity-resin) beads[218]. Following the second elution, the proteins (bait and interacting partners) are typically identified by mass spectrometry. The TAP interaction survey method is now recognized as having the best coverage and accuracy of experimental high-throughput interaction detection methods[138], and has the substantive advantage of detecting interactions among proteins assembled into protein complexes under near-native physiological conditions.

The first adaptation of the TAP method to large-scale protein complex characterization was performed in yeast and reported in 2001[242]. By tagging approximately 1700 proteins, the authors were able to obtain data supporting 232 distinct functional interaction modules, and provide hints as to the possible biological roles of 344 uncharacterized proteins based on physical association with proteins of known function.

In the following 4 years, hundreds of studies were published identifying specific protein interactions and complexes not only in yeast, but also in *E. coli*[243], plant[244], *drosophila*[245], and human[246,247] (over 100 low-throughput studies as of 2004[248]). In 2006, two independent studies simultaneously published definitive, virtually comprehensive global surveys of stable soluble protein complexes for yeast[48,49]. Stringent data processing procedures were applied to the enormous raw datasets, seemingly eliminating false-positives. Yet, despite the rigor and sheer scale of the two studies, with ~50% overlap in the total number of proteins detected (1304 in common out of the 1993 reported in Gavin *et al*[48] study and the 2388 found in Krogan *et al*[9]), initial cross-comparisons revealed a surprisingly modest overlap in the respective interactions (<25%). Aside from minor differences in the respective screening procedures, the most likely cause for this seeming shortfall stems from differences in the computational algorithms used to ascertain the most likely protein interactions and interaction clusters[145], thus illustrating the impact of the increasingly sophisticated analytical methods used to interpret genome-scale interaction data. It should however be noted that although TAP is specifically designed for complex resolution, final determination of complexes has depended on algorithmic interpretation of determined confidence scores between pairs of interactors; thus the potential exists to misrepresent experimentally determined complexes. One other caveat to TAP screening is the functional interference potentially caused by introduction of the epitope tag or selection marker, which may perturb protein folding/function and mRNA stability/regulation, respectively. Despite being relatively innocuous, initial reports suggested that a tag may impair function in as much as 18% of all targeted proteins[242]. It is worth noting however, that one third[48] to

nearly one half[49] of interacting proteins were identified as untagged preys for other tagged proteins, and thus could still be surveyed.

Text mining

As individual researchers typically gather information for proteins of interest through examining publications, text mining remains among the most established method of gathering PPI data. However, since the number of publications available today is expanding explosively for virtually every sub-specialty of the biological sciences, comprehensive text interpretation has become too demanding. To accommodate this data overload, computer programs have been developed to systematically process and parse out interaction data from large bodies of published literature in an automated manner.

Algorithms that scan the literature for PPI have two tasks: first, to recognize conclusively instances of mentioned proteins (which is challenging because proteins often have multiple names and abbreviations), and, second, to define the biophysical context in which these proteins are being discussed[249]. Due to the complexity of the English language, and of the varied nature of protein interactions themselves, defining context is no trivial task. More importantly, the algorithm must be able to accurately distinguish a genuine interaction from a coincidental occurrence, which may present with nearly the same lexical syntax. Effective algorithms must be trained to recognize appropriate English sentence features (including verbal form, presence/absence of a noun), and scored against manual curation to evaluate performance, before finally being used to parse large bodies of literature[250]. To improve accuracy, recent computational approaches have integrated information from sentences both preceding and following the

mention of protein/gene names to improve the accuracy of the general approach[251,252] (for more in-depth review, see Jensen *et al*[249] and Hirschman *et al*[253]). Today there are several publicly available software packages for performing automated literature-mining of protein interactions (MEDSYNDIKATE[254], CONAN[255]).

Text mining can be an effective data collection method for several reasons. First, the data collected is often from multiple experimental sources, resulting in a collective set of interactions less subject to any specific experimental bias. Second, as the data retrieved by text mining algorithms is vastly beyond what is published in any single (even high-throughput) experiment, there is increased potential for cross-validation. For example, there are over 80,000 yeast-specific protein interactions in the last release of the BioGRID[256] database. If one filters and reduces this set to only the most accurate 1% of interactions, an even larger subset of putative interactions is generated than obtained for either the Krogan *et al*[49] or Gavin *et al*[48] published datasets. For these reasons, text mining, in combination with manual curation, has been used to populate public databases such as preBIND[257] and BioGRID[256], which are invaluable for computational biologists, as they tend to represent the largest sources of PPI data for every species. The accuracy of interactions housed within these databases significantly increases, however, when literature retrieval algorithms are coupled with manual curation[257], so experts can comb through and remove as many spurious interactions as possible. Interaction networks created from literature-mined protein interactions also exhibit topology similar to networks generated by high-throughput screening alone, although with better coverage[139].

As far as the study of mammalian protein interactions, a potential disadvantage of text mining is the prevalence of literature for commonly studied proteins, such as those

associated with cancer or other widely studied processes or diseases. The increased representation skews the resulting interaction map. Consequently, the accuracy of text mining algorithms also improves as more experimental data (and hence publications) is generated for a given organism.

Clustering of interaction data

After obtaining interaction data (either experimentally or computationally), the next challenge then becomes that of assigning proteins into individual complexes. Computational partitioning of interaction networks into highly connected clusters has been used to impressive effect in large-scale yeast[48,49,258] and human[259,260] studies.

Although varied, clustering algorithms usually define sub-groups of proteins that exhibit higher similarity amongst themselves than with other sub-groups[261]. In defining interaction clusters, which are posited to represent protein complexes, there are several algorithms that can be used, the selection of which will depend on the nature of the desired outcome. Some algorithms produce in individual (exclusive) clusters with non-shared members; others will allow shared members between clusters. Non-exclusivity (i.e. clusters can have shared members) can be viewed as being more biologically accurate, as many proteins show promiscuity in terms of complex membership. However complexes with non-shared members ease post-analysis of results and facilitate functional categorizations. Additionally, some algorithms are capable of incorporating biological or functional data.

The lack of overlap[145] in corresponding yeast protein interactions reported by Krogan *et al*[49] and Gavin *et al*[48] points to the importance of selecting a standardized

computational assessment procedure. Recent follow-up studies[262] (also S. Pu & S. Wodak; *personal communication*) demonstrate that the application of a unified clustering method results in similar clusters for both datasets. Moreover, an additional caveat is that while disparate algorithms can decipher alternate interconnected groups of proteins, the results serve only as an approximation of the actual physical complexes present within the cell. Many of the algorithms described operate based on properties of graph theory (see below).

Clustering algorithms

K-means is one of the simplest clustering algorithms in application today. For a set of *X* clusters, *X* centroid values will be determined equally covering the range of the inputted set. Data points are then individually assigned to the centroid that they are closest in value to. Unfortunately, in order to use *k*-means clustering, the number of clusters must be anticipated in advance. This represents a major disadvantage when studying novel interactomes, as the number of complexes present is impossible to pre-determine.

Commonly depicted using a dendrogram, hierarchical clustering is famously used in biology to classify species based on phenotypic or phylogenetic properties. More recently, hierarchical clustering has been applied to PPI data, finding proteins with highly similar expression patterns[263]. There are several variations of hierarchical clustering; however, all commonly applied in interaction cluster analysis consist of the same steps. Each node in the set being analyzed begins the process as its own cluster. From there, similarity between any two nodes in terms of properties such as minimal path length[264], is

computed using one of many different measures (i.e. Spearman's Rank). The two nodes that are the most similar are moved into the same cluster, and distances to all other nodes are re-computed. This is continued until the entire tree structure has been established. Examining proteins grouped together at one level of the hierarchy allows one to draw finite protein clusters.

In a purely computational study, Krause *et al*[265] applied three variations of hierarchical clustering to a yeast affinity-purification dataset, ultimately concluding that more interaction data is required for an accurate complexosome description. Several years later, Gavin *et al*[48] built on Krause's results and used a similar approach to draw their PPI clusters based on experimental data in the yeast proteome. This method also integrated functional data when deriving clusters, and was able to group proteins into either stable protein 'modules' comprising the functional core of unified protein complexes, and extended promiscuous associators; designations the established authors felt to be truly indicative of the state of the complexosome.

One method that is gaining increased attention for PPI clustering is the Markov clustering algorithm (MCL)[144]. Once a network graph of proteins has been generated, random 'walks' are created *in silico* (wherein nodes are picked at random and a pre-determined number of PPI "edges" is traversed). Through an iterative process of many such walks, the algorithm splits the proteins into exclusive groups based on the relative flow across highly traversed regions (high connectivity indicates clusters). In a recent comparison of biologically applied clustering algorithms[266], MCL was shown to be remarkably resilient to spurious graph perturbations. Appropriately, MCL was used to

describe many novel protein complexes within the yeast proteome based on large-scale TAP experimental data[49].

Another algorithm, known as MCODE[267], used for detecting protein complexes among PPI networks similarly divides interaction data into clusters based on regions of high connectivity. This algorithm is freely available as a plug-in for the Cytoscape[187] software package, which allows for ready viewing of the agglomerative results.

Introduction to graph theory

Graphs with proteins or genes depicted as nodes and interactions as the edges between them are frequently used to represent both protein and genetic interaction networks. Using graphs of this type to depict interaction networks allows analysis by a certain set of mathematical formulae, those pertaining to graph theory.

Many naturally occurring graphs exhibit a topology in which the majority of constituent nodes have only a few associations, and just a few nodes have many associations (i.e. the number of associations per protein follows a power-law distribution, with corresponding graphs referred to as being scale-free). Scale-free graphs exhibit 'small-world' characteristics, as most nodes can be linked to one another by following a short path. Models of this type are generally noted in real-world networks, in fact the so-called 'small-world' problem[268] was originally described when Stanley Milgram noticed shorter than expected path lengths among social networks (later the basis for the famous 'six degrees of separation' hypothesis). It is tempting to believe that scale-free topology exists in biological networks as well, as it would offer greater protection against random deletions than other types of associative graph structures.

Whether the PPI network truly follows a scale-free pattern remains controversial, with compelling evidence presented both supporting[148,269] and negating[270,271] the claim. Regardless, proteins with higher connectivity in the network (often called 'hubs') tend to be more essential to cell function[147,148]. Similarly, graph properties such as betweenness and closeness (literally measuring the relative proximities of two nodes on a network graph) are often used to describe how functionally related two proteins or genes are[272,273]. There are several freely available software packages used in the analysis of PPI networks, the most common of which are Pajek[144], and Cytoscape[187] with Cytoscape having many useful plug-ins for biological analysis.

Appendix Table 1 Epistacy detected using RSA and GCA

ORF 1 (KAN)	Gene 1	ORF 2 (NAT)	Gene 2	RSA	GCA
YAL023C	PMT2	YOR321W	PMT3	1	1
YBL039C	URA7	YJR103W	URA8	1	1
YBL085W	BOI1	YER114C	BOI2	1	1
YBL087C	RPL23A	YER117W	RPL23B	1	1
YBR009C	HHF1	YNL030W	HHF2	1	1
YBR048W	RPS11B	YDR025W	RPS11A	1	1
YBR082C	UBC4	YDR059C	UBC5	1	1
YBR169C	SSE2	YPL106C	SSE1	1	1
YBR189W	RPS9B	YPL081W	RPS9A	1	1
YBR191W	RPL21A	YPL079W	RPL21B	1	1
YCR031C	RPS14A	YJL191W	RPS14B	1	1
YDL061C	RPS29B	YLR388W	RPS29A	1	1
YDL134C	PPH21	YDL188C	PPH22	1	1
YDL161W	ENT1	YLR206W	ENT2	1	1
YDL175C	AIR2	YIL079C	AIR1	1	1
YDR098C	GRX3	YER174C	GRX4	1	1
YDR213W	UPC2	YLR228C	ECM22	1	1
YDR253C	MET32	YPL038W	MET31	1	1
YDR312W	SSF2	YHR066W	SSF1	1	1
YDR358W	GGA1	YHR108W	GGA2	1	1
YDR385W	EFT2	YOR133W	EFT1	1	1
YDR447C	RPS17B	YML024W	RPS17A	1	1
YDR450W	RPS18A	YML026C	RPS18B	1	1
YDR502C	SAM2	YLR180W	SAM1	1	1
YER031C	YPT31	YGL210W	YPT32	1	1
YER062C	HOR2	YIL053W	RHR2	1	1
YFR053C	HXK1	YGL253W	HXK2	1	1
YGL076C	RPL7A	YPL198W	RPL7B	1	1
YGL135W	RPL1B	YPL220W	RPL1A	1	1
YGR010W	NMA2	YLR328W	NMA1	1	1
YGR032W	GSC2	YLR342W	FKS1	1	1
YGR038W	ORM1	YLR350W	ORM2	1	1
YGR108W	CLB1	YPR119W	CLB2	1	1
YGR209C	TRX2	YLR043C	TRX1	1	1
YGR214W	RPS0A	YLR048W	RPS0B	1	1
YHR021C	RPS27B	YKL156W	RPS27A	1	1
YHR135C	YCK1	YNL154C	YCK2	1	1
YHR203C	RPS4B	YJR145C	RPS4A	1	1
YIL095W	PRK1	YNL020C	ARK1	1	1
YIL105C	SLM1	YNL047C	SLM2	1	1
YIL133C	RPL16A	YNL069C	RPL16B	1	1
YIR033W	MGA2	YKL020C	SPT23	1	1
YJL129C	TRK1	YKR050W	TRK2	1	1
YJL138C	TIF2	YKR059W	TIF1	1	1

YKL043W	PHD1	YMR016C	SOK2	1	1
YLR028C	ADE16	YMR120C	ADE17	1	1
YLR441C	RPS1A	YML063W	RPS1B	1	1
YLR450W	HMG2	YML075C	HMG1	1	1
YMR183C	SSO2	YPL232W	SSO1	1	1
YNL096C	RPS7B	YOR096W	RPS7A	1	1
YNL098C	RAS2	YOR101W	RAS1	1	1
YNL302C	RPS19B	YOL121C	RPS19A	1	1
YPR052C	NHP6A	YBR089C-A	NHP6B	1	1
YBR010W	HHT1	YNL031C	HHT2	1	0
YBR118W	TEF2	YPR080W	TEF1	1	0
YBR210W	ERV15	YGL054C	ERV14	1	0
YDL022W	GPD1	YOL059W	GPD2	1	0
YDL138W	RGT2	YDL194W	SNF3	1	0
YDR436W	PPZ2	YML016C	PPZ1	1	0
YER081W	SER3	YIL074C	SER33	1	0
YGR124W	ASN2	YPR145W	ASN1	1	0
YGR192C	TDH3	YJR009C	TDH2	1	0
YGR254W	ENO1	YHR174W	ENO2	1	0
YJL098W	SAP185	YKR028W	SAP190	1	0
YJL133W	MRS3	YKR052C	MRS4	1	0
YKL032C	IXR1	YMR072W	ABF2	1	0
YKL129C	MYO3	YMR109W	MYO5	1	0
YMR186W	HSC82	YPL240C	HSP82	1	0
YOR226C	ISU2	YPL135W	ISU1	1	0
YAL053W	FLC2	YOR365C	YOR365C	0	1
YBL027W	RPL19B	YBR084C-A	RPL19A	0	1
YBL067C	UBP13	YER098W	UBP9	0	1
YBL068W	PRS4	YER099C	PRS2	0	1
YBL072C	RPS8A	YER102W	RPS8B	0	1
YBL079W	NUP170	YER105C	NUP157	0	1
YBL089W	AVT5	YER119C	AVT6	0	1
YBL106C	SRO77	YPR032W	SRO7	0	1
YBR052C	RFS1	YDR032C	PST2	0	1
YBR078W	ECM33	YDR055W	PST1	0	1
YBR145W	ADH5	YOL086C	ADH1	0	1
YBR161W	CSH1	YPL057C	SUR1	0	1
YBR284W	YBR284W	YJL070C	YJL070C	0	1
YCL024W	KCC4	YDR507C	GIN4	0	1
YDL021W	GPM2	YOL056W	GPM3	0	1
YDL042C	SIR2	YOL068C	HST1	0	1
YDL048C	STP4	YLR375W	STP3	0	1
YDL075W	RPL31A	YLR406C	RPL31B	0	1
YDL088C	ASM4	YMR153W	NUP53	0	1
YDL130W-A	STF1	YDL181W	INH1	0	1
YDL224C	WHI4	YNL197C	WHI3	0	1
YDR066C	YDR066C	YER139C	YER139C	0	1

YDR069C	DOA4	YER144C	UBP5	0	1
YDR099W	BMH2	YER177W	BMH1	0	1
YDR251W	PAM1	YPL032C	SVL3	0	1
YDR264C	AKR1	YOR034C	AKR2	0	1
YDR300C	PRO1	YHR033W	YHR033W	0	1
YDR348C	YDR348C	YHR097C	YHR097C	0	1
YDR351W	SBE2	YHR103W	SBE22	0	1
YDR379W	RGA2	YOR127W	RGA1	0	1
YDR389W	SAC7	YOR134W	BAG7	0	1
YDR409W	SIZ1	YOR156C	NFI1	0	1
YDR418W	RPL12B	YEL054C	RPL12A	0	1
YDR463W	STP1	YHR006W	STP2	0	1
YDR480W	DIG2	YPL049C	DIG1	0	1
YDR497C	ITR1	YOL103W	ITR2	0	1
YDR505C	PSP1	YLR177W	YLR177W	0	1
YEL041W	YEF1	YJR049C	UTR1	0	1
YER037W	PHM8	YGL224C	SDT1	0	1
YER059W	PCL6	YIL050W	PCL7	0	1
YER131W	RPS26B	YGL189C	RPS26A	0	1
YFL004W	VTC2	YPL019C	VTC3	0	1
YFR040W	SAP155	YGL229C	SAP4	0	1
YGL031C	RPL24A	YGR148C	RPL24B	0	1
YGL049C	TIF4632	YGR162W	TIF4631	0	1
YGL063W	PUS2	YPL212C	PUS1	0	1
YGL084C	GUP1	YPL189W	GUP2	0	1
YGL133W	ITC1	YPL216W	YPL216W	0	1
YGR056W	RSC1	YLR357W	RSC2	0	1
YGR070W	ROM1	YLR371W	ROM2	0	1
YGR085C	RPL11B	YPR102C	RPL11A	0	1
YGR092W	DBF2	YPR111W	DBF20	0	1
YGR109C	CLB6	YPR120C	CLB5	0	1
YGR121C	MEP1	YPR138C	MEP3	0	1
YGR188C	BUB1	YJL013C	MAD3	0	1
YGR256W	GND2	YHR183W	GND1	0	1
YGR279C	SCW4	YMR305C	SCW10	0	1
YHL003C	LAG1	YKL008C	LAC1	0	1
YHR030C	SLT2	YKL161C	YKL161C	0	1
YHR115C	DMA1	YNL116W	DMA2	0	1
YHR117W	TOM71	YNL121C	TOM70	0	1
YHR123W	EPT1	YNL130C	CPT1	0	1
YIL131C	FKH1	YNL068C	FKH2	0	1
YIL135C	VHS2	YNL074C	MLF3	0	1
YIL149C	MLP2	YKR095W	MLP1	0	1
YJL082W	IML2	YKR018C	YKR018C	0	1
YJL139C	YUR1	YKR061W	KTR2	0	1
YJL164C	TPK1	YKL166C	TPK3	0	1
YJL165C	HAL5	YKL168C	KKQ8	0	1

YJR148W	BAT2	YHR208W	BAT1	0	1
YKL126W	YPK1	YMR104C	YPK2	0	1
YKL127W	PGM1	YMR105C	PGM2	0	1
YKR072C	SIS2	YOR054C	VHS3	0	1
YKR077W	YKR077W	YOR066W	YOR066W	0	1
YKR089C	TGL4	YOR081C	TGL5	0	1
YKR106W	YKR106W	YCL073C	YCL073C	0	1
YLL010C	PSR1	YLR019W	PSR2	0	1
YML100W	TSL1	YMR261C	TPS3	0	1
YML109W	ZDS2	YMR273C	ZDS1	0	1
YMR194W	RPL36A	YPL249C-A	RPL36B	0	1
YMR198W	CIK1	YPL253C	VIK1	0	1
YMR199W	CLN1	YPL256C	CLN2	0	1
YMR233W	TRI1	YOR295W	UAF30	0	1
YMR242C	RPL20A	YOR312C	RPL20B	0	1
YMR243C	ZRC1	YOR316C	COT1	0	1
YNL087W	TCB2	YOR086C	TCB1	0	1
YNL293W	MSB3	YOL112W	MSB4	0	1
YNL298W	CLA4	YOL113W	SKM1	0	1
YNL299W	TRF5	YOL115W	TRF4	0	1
YOR233W	KIN4	YPL141C	YPL141C	0	1
YPR159W	KRE6	YGR143W	SKN1	0	1

Genes detected as epistatic using either Random Spore Analysis (RSA) or Growth Curve Analysis (GCA). Epistasis using either technique is indicated with a ‘1’ in their respective columns. Those genes with a genetic interaction detectible using both RSA and GCA were termed the ‘intersect group’ for subsequent analysis, and those detectible through either RSA or GCA the ‘union group’.

Appendix Table 2 Detected trigenic interactions

Gene 1	Gene 2	Hit	Et
ALK2	ALK1	CEM1	-0.473
ALK2	ALK1	SNC2	-0.415
ALK2	ALK1	NTG2	-0.393
ALK2	ALK1	NCE102	-0.36
ALK2	ALK1	YOL131W	-0.342
ALK2	ALK1	YVC1	-0.334
ALK2	ALK1	MMS22	-0.322
ALK2	ALK1	CRZ1	-0.307
ALK2	ALK1	YLL058W	-0.295
ALK2	ALK1	PIN2	-0.294
ALK2	ALK1	KES1	-0.275
ALK2	ALK1	ASF1	-0.258
ALK2	ALK1	HMO1	-0.258
ALK2	ALK1	CKA1	-0.256
ALK2	ALK1	CMK2	-0.256
ALK2	ALK1	YPR096C	-0.245
CNA1	CMP2	RPS16B	-0.46
CNA1	CMP2	YUR1	-0.423
CNA1	CMP2	SEC28	-0.422
CNA1	CMP2	EAF1	-0.39
CNA1	CMP2	SMI1	-0.366
CNA1	CMP2	CSG2	-0.349
CNA1	CMP2	KEX1	-0.342
CNA1	CMP2	ALG3	-0.33
CNA1	CMP2	YCR102C	-0.326
CNA1	CMP2	RPN4	-0.302
CNA1	CMP2	BST1	-0.298
CNA1	CMP2	MGA2	-0.297
CNA1	CMP2	YPS7	-0.289
CNA1	CMP2	ERV14	-0.288
CNA1	CMP2	PTC1	-0.285
CNA1	CMP2	YJL068C	-0.273
CNA1	CMP2	PEX10	-0.271
CNA1	CMP2	YCR016W	-0.268
CNA1	CMP2	CWC15	-0.26
CNA1	CMP2	VPS74	-0.252
CNA1	CMP2	ROT2	-0.244
CNA1	CMP2	ETR1	-0.243
CNA1	CMP2	TRP4	-0.243
KIN1	KIN2	OAR1	-0.39

KIN1	KIN2	VPS51	-0.323
KIN1	KIN2	SLA1	-0.274
KIN1	KIN2	FCY2	-0.272
KIN1	KIN2	BEM4	-0.264
KIN1	KIN2	FMC1	-0.254
KIN1	KIN2	YDJ1	-0.244
KIN82	FPK1	CHK1	-0.673
KIN82	FPK1	GAS4	-0.598
KIN82	FPK1	MAL12	-0.526
KIN82	FPK1	INO4	-0.504
KIN82	FPK1	MSN1	-0.502
KIN82	FPK1	EAF1	-0.501
KIN82	FPK1	ADY4	-0.482
KIN82	FPK1	DUG2	-0.419
KIN82	FPK1	GYP1	-0.416
KIN82	FPK1	ARL3	-0.376
KIN82	FPK1	YOR304C-A	-0.358
KIN82	FPK1	YPR022C	-0.358
KIN82	FPK1	UBC11	-0.338
KIN82	FPK1	YGR122W	-0.332
KIN82	FPK1	PMP3	-0.327
KIN82	FPK1	CAM1	-0.32
KIN82	FPK1	YCL049C	-0.305
KIN82	FPK1	YPR097W	-0.29
KIN82	FPK1	NKP2	-0.276
KIN82	FPK1	OYE3	-0.271
KIN82	FPK1	YPL260W	-0.268
KIN82	FPK1	MRH1	-0.263
KIN82	FPK1	PTC2	-0.263
KIN82	FPK1	FMP46	-0.262
KIN82	FPK1	AXL1	-0.257
KIN82	FPK1	PEX13	-0.254
KIN82	FPK1	DSD1	-0.253
KIN82	FPK1	YDR179W-A	-0.249
KIN82	FPK1	RPL43A	-0.242
NRG2	ALK1	UBC7	-0.407
NRG2	ALK1	MDM12	-0.322
NRG2	ALK1	SAC1	-0.307
NRG2	ALK1	HMO1	-0.269
NRG2	ALK1	REC114	-0.269
NRG2	ALK1	TRI1	-0.268
NRG2	ALK1	RPS16A	-0.264
NRG2	ALK1	TMA108	-0.262

NRG2	ALK1	YML037C	-0.255
NRG2	ALK1	PMP3	-0.247
NRG2	ALK1	VPS74	-0.246
NRG2	ALK1	ERG3	-0.244
NRG2	NRG1	ADY4	-0.773
NRG2	NRG1	TRM44	-0.481
NRG2	NRG1	RAD1	-0.4
NRG2	NRG1	YKL061W	-0.382
NRG2	NRG1	RIM101	-0.378
NRG2	NRG1	CHS5	-0.362
NRG2	NRG1	RIM13	-0.358
NRG2	NRG1	PEX30	-0.354
NRG2	NRG1	RIM9	-0.348
NRG2	NRG1	VRP1	-0.347
NRG2	NRG1	IME2	-0.344
NRG2	NRG1	GDH1	-0.34
NRG2	NRG1	IZH2	-0.336
NRG2	NRG1	CYB2	-0.332
NRG2	NRG1	CNB1	-0.324
NRG2	NRG1	YLR278C	-0.324
NRG2	NRG1	FRE1	-0.32
NRG2	NRG1	SKG3	-0.32
NRG2	NRG1	DSE1	-0.312
NRG2	NRG1	IRC8	-0.307
NRG2	NRG1	RIM20	-0.301
NRG2	NRG1	DFG16	-0.3
NRG2	NRG1	LSM1	-0.3
NRG2	NRG1	YLR287C	-0.295
NRG2	NRG1	AIM43	-0.283
NRG2	NRG1	ATF1	-0.282
NRG2	NRG1	PXA2	-0.28
NRG2	NRG1	YHR080C	-0.28
NRG2	NRG1	MCM22	-0.276
NRG2	NRG1	HMX1	-0.272
NRG2	NRG1	YCR102C	-0.271
NRG2	NRG1	YMR031C	-0.269
NRG2	NRG1	GCN1	-0.265
NRG2	NRG1	CAF20	-0.262
NRG2	NRG1	REV1	-0.26
NRG2	NRG1	MSH2	-0.258
NRG2	NRG1	PLB2	-0.254
NRG2	NRG1	ACN9	-0.249
NRG2	NRG1	PTP2	-0.249

NRG2	NRG1	PNG1	-0.248
NRG2	NRG1	CLN2	-0.246
NRG2	NRG1	RAV1	-0.246
NRG2	NRG1	YGL159W	-0.245
NRG2	NRG1	YKL187C	-0.242
NRG2	NRG1	YFR045W	-0.241
PSK1	PSK2	MEP2	-0.416
PSK1	PSK2	YNL144C	-0.382
PSK1	PSK2	PDR12	-0.313
PSK1	PSK2	VPS41	-0.29
PSK1	PSK2	SYH1	-0.253
PSK1	PSK2	MKT1	-0.245
PSK1	PSK2	RPL13A	-0.242
SUT1	SUT2	PRS5	-0.525
SUT1	SUT2	GCY1	-0.317
SUT1	SUT2	ARO1	-0.283
SUT1	SUT2	SUR4	-0.265
SUT1	SUT2	BUB3	-0.259
VHS1	SKS1	SUL1	-0.414
VHS1	SKS1	COQ10	-0.359
VHS1	SKS1	TIR2	-0.324
VHS1	SKS1	FYV12	-0.287
VHS1	SKS1	IES4	-0.262

Genetic interactions detected using double-deletion queries. First two columns represent the deletions present in the initial query strain, and third column the corresponding hits. Et column shows the corresponding ε values, only statistically significant interactions are shown.

References

1. Haldane, J.B.S. *The causes of evolution*, (Princeton University Press, 1932).

2. Ohno, S. *Evolution by Gene Duplication*, (Springer-Verlag, Berlin, 1970).

3. Lynch, M. & Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151-5.

4. Hurles, M. (2004) Gene duplication: the genomic trade in spare parts. *PLoS Biol* **2**, E206.

5. Kratz, E., Dugas, J.C. & Ngai, J. (2002) Odorant receptor gene regulation: implications from genomic organization. *Trends Genet* **18**, 29-34.

6. Thompson, L.H. & Schild, D. (2001) Homologous recombinational repair of DNA ensures mammalian chromosome stability. *Mutat Res* **477**, 131-53.

7. Zhang, J. (2003) Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18**, 292-8.

8. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403-10.

9. Campbell, N.A. & Reece, J.B. *Biology*, 1231 p. (Pearson, San Francisco, CA, 2005).

10. Conrad, B. & Antonarakis, S.E. (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annual review of genomics and human genetics* **8**, 17-35.

11. Eichler, E.E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* **17**, 661-9.

12. Zhang, L., Lu, H.H., Chung, W.Y., Yang, J. & Li, W.H. (2005) Patterns of segmental duplication in the human genome. *Mol Biol Evol* **22**, 135-41.

13. Long, M., Betran, E., Thornton, K. & Wang, W. (2003) The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**, 865-75.

14. Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A. & Kaessmann, H. (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* **3**, e357.

15. Kihara, H. & Ono, T. (1926) Chromosomenzahlen und systematische gruppierung der rumex-arten. *Cell and Tissue Research*.

16. Chain, F.J.J., Ilieva, D. & Evans, B.J. (2008) Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evol Biol* **8**, 43.

17. Kellis, M., Birren, B.W. & Lander, E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**, 617-24.

18. Otto, S.P. & Whitton, J. (2000) Polyploid incidence and evolution. *Annu Rev Genet* **34**, 401-437.

19. Spring, J. (1997) Vertebrate evolution by interspecific hybridisation--are we polyploid? *FEBS Lett* **400**, 2-8.

20. Masterson, J. (1994) Stomatal Size in Fossil Plants: Evidence for Polyploidy in Majority of Angiosperms. *Science* **264**, 421-424.

21. Holland, P.W. & Takahashi, T. (2005) The evolution of homeobox genes: Implications for the study of brain development. *Brain Res Bull* **66**, 484-90.

22. Holland, L.Z., Albalat, R., Azumi, K., Benito-Gutierrez, E., Blow, M.J., Bronner-Fraser, M., Brunet, F., Butts, T., Candiani, S., Dishaw, L.J., Ferrier, D.E., Garcia-Fernandez, J., Gibson-Brown, J.J., Gissi, C., Godzik, A., Hallbook, F., Hirose, D., Hosomichi, K., Ikuta, T., Inoko, H., Kasahara, M., Kasamatsu, J., Kawashima, T., Kimura, A., Kobayashi, M., Kozmik, Z., Kubokawa, K., Laudet, V., Litman, G.W., McHardy, A.C., Meulemans, D., Nonaka, M., Olinski, R.P., Pancer, Z., Pennacchio, L.A., Pestarino, M., Rast, J.P., Rigoutsos, I., Robinson-Rechavi, M., Roch, G., Saiga, H., Sasakura, Y., Satake, M., Satou, Y., Schubert, M., Sherwood, N., Shiina, T., Takatori, N., Tello, J., Vopalensky, P., Wada, S., Xu, A., Ye, Y., Yoshida, K., Yoshizaki, F., Yu, J.K., Zhang, Q., Zmasek, C.M., de Jong, P.J., Osoegawa, K., Putnam, N.H., Rokhsar, D.S., Satoh, N. & Holland, P.W. (2008) The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res* **18**, 1100-11.

23. Andalis, A.A., Storchova, Z., Styles, C., Galitski, T., Pellman, D. & Fink, G.R. (2004) Defects arising from whole-genome duplications in Saccharomyces cerevisiae. *Genetics* **167**, 1109-21.

24. Mayer, V.W. & Aguilera, A. (1990) High levels of chromosome instability in polyploids of Saccharomyces cerevisiae. *Mutat Res* **231**, 177-86.

25. Wolfe, K.H. & Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708-13.

26. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. (2000) Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell* **12**, 1093-101.

27. Jeffreys, A.J., Wilson, V., Wood, D., Simons, J.P., Kay, R.M. & Williams, J.G. (1980) Linkage of adult alpha- and beta-globin genes in X. laevis and gene duplication by tetraploidization. *Cell* **21**, 555-64.

28. Christoffels, A., Koh, E.G., Chia, J.M., Brenner, S., Aparicio, S. & Venkatesh, B. (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol* **21**, 1146-51.

29. Dehal, P. & Boore, J.L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**, e314.

30. Postlethwait, J.H., Woods, I.G., Ngo-Hazelett, P., Yan, Y.L., Kelly, P.D., Chu, F., Huang, H., Hill-Force, A. & Talbot, W.S. (2000) Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res* **10**, 1890-902.

31. Wolfe, K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* **2**, 333-41.

32. Koszul, R., Caburet, S., Dujon, B. & Fischer, G. (2004) Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *Embo J* **23**, 234-43.

33. Piskur, J. (2001) Origin of the duplicated regions in the yeast genomes. *Trends Genet* **17**, 302-3.

34. Skrabanek, L. & Wolfe, K.H. (1998) Eukaryote genome duplication - where's the evidence? *Curr Opin Genet Dev* **8**, 694-700.

35. Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.M., Beyne, E., Bleykasten, C., Boisrame, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.F., Straub, M.L., Suleau, A., Swennen, D., Tekaia, F., Wesolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P. & Souciet, J.L. (2004) Genome evolution in yeasts. *Nature* **430**, 35-44.

36. Byrne, K.P. & Wolfe, K.H. (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15**, 1456-61.

37. Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S. & Wolfe, K.H. (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341-5.

38. Wang, H., Xu, Z., Gao, L. & Hao, B. (2009) A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evol Biol* **9**, 195.

39. Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biemont, C., Skalli, Z., Cattolico, L., Poulain, J., De Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K.J., McEwan, P., Bosak, S., Kellis, M., Volff, J.N., Guigo, R., Zody, M.C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quetier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E.S., Weissenbach, J. & Roest Crollius, H. (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-57.

40. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815.

41. Putnam, N.H., Butts, T., Ferrier, D.E., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.K., Benito-Gutierrez, E.L., Dubchak, I., Garcia-Fernandez, J., Gibson-Brown, J.J., Grigoriev, I.V., Horton, A.C., de Jong, P.J., Jurka, J., Kapitonov, V.V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L.A., Salamov, A.A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin, I.T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L.Z., Holland, P.W., Satoh, N. & Rokhsar, D.S. (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-71.

42. Bassett, D.E., Jr., Boguski, M.S., Spencer, F., Reeves, R., Kim, S., Weaver, T. & Hieter, P. (1997) Genome cross-referencing and XREFdb: implications for the identification and analysis of genes mutated in human disease. *Nat Genet* **15**, 339-44.

43. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S.G. (1996) Life with 6000 genes. *Science* **274**, 546, 563-7.

44. Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. & Davis, R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**, 65-73.

45. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. & Herskowitz, I. (1998) The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705.

46. Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K. & Weissman, J.S. (2003) Global analysis of protein expression in yeast. *Nature* **425**, 737-41.

47. Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S. & O'Shea, E.K. (2003) Global analysis of protein localization in budding yeast. *Nature* **425**, 686-91.

48. Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B. & Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631-6.

49. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrin-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. & Greenblatt, J.F. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637-43.

50. Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. & Botstein, D. (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26**, 73-9.

51. Sherman, F. (2002) Getting started with yeast. *Meth Enzymol* **350**, 3-41.

52. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A.P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.D., Flaherty, P., Foury, F., Garfinkel, D.J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J.H., Hempel, S., Herman, Z., Jaramillo, D.F., Kelly, D.E., Kelly, S.L., Kotter, P., LaBonte, D., Lamb, D.C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S.L., Revuelta, J.L., Roberts, C.J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D.D., Sookhai-Mahadeo, S., Storms,

R.K., Strathern, J.N., Valle, G., Voet, M., Volckaert, G., Wang, C.Y., Ward, T.R., Wilhelmy, J., Winzeler, E.A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J.D., Snyder, M., Philippsen, P., Davis, R.W. & Johnston, M. (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-91.

53. Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., Davis, R.W. & Li, W.H. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63-6.

54. Guan, Y., Dunham, M.J. & Troyanskaya, O.G. (2007) Functional Analysis of Gene Duplications in Saccharomyces cerevisiae. *Genetics* **175**, 933-43.

55. Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. & Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* **102**, 5454-9.

56. Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S. & Van de Peer, Y. (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* **7**, R43.

57. Taylor, J.S. & Raes, J. (2004) Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet* **38**, 615-43.

58. Conant, G.C. & Wagner, A. (2002) GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res* **30**, 3378-86.

59. Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G. & Robertson, D.L. (2007) All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol* **8**, R209.

60. Blanc, G. & Wolfe, K.H. (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16**, 1679-91.

61. Davis, J.C. & Petrov, D.A. (2005) Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet* **21**, 548-51.

62. Li, L., Huang, Y., Xia, X. & Sun, Z. (2006) Preferential duplication in the sparse part of yeast protein interaction network. *Mol Biol Evol* **23**, 2467-73.

63. Kim, P.M., Lu, L.J., Xia, Y. & Gerstein, M.B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314**, 1938-41.

64. Conant, G.C. & Wolfe, K. (2006) Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biol* **4**, e109.

65. Brookfield, J. (1992) Can genes be truly redundant? *Curr Biol* **2**, 553-4.

66. Fisher, R.A. (1935) The sheltering of lethals. *American Naturalist*.

67. Ferris, S.D. & Whitt, G.S. (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J Mol Evol* **12**, 267-317.

68. Lynch, M. *Genetics and analysis of quantitative traits, Michael Lynch, Bruce Walsh*, (1997).

69. Kimura, M. & Ota, T. (1974) On some principles governing molecular evolution. *Proc Natl Acad Sci U S A* **71**, 2848-52.

70. Allendorf, F.W. & Utter, F.M. (1976) Gene duplication in the family Salmonidae. III. Linkage between two duplicated loci coding for aspartate aminotransferase in the cutthroat trout (Salmo clarki). *Hereditas* **82**, 19-24.

71. Bisbee, C.A., Baker, M.A., Wilson, A.C., Haji-Azimi, I. & Fischberg, M. (1977) Albumin phylogeny for clawed frogs (Xenopus). *Science* **195**, 785-7.

72. Whitkus, R., Doebley, J. & Lee, M. (1992) Comparative genome mapping of Sorghum and maize. *Genetics* **132**, 1119-30.

73. Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. & Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531-45.

74. Hughes, A.L. (1994) The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**, 119-24.

75. Dermitzakis, E.T. & Clark, A.G. (2001) Differential selection after duplication in mammalian developmental genes. *Mol Biol Evol* **18**, 557-62.

76. Lynch, M. & Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459-73.

77. van Hoof, A. (2005) Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics* **171**, 1455-61.

78. MacCarthy, T. & Bergman, A. (2007) The limits of subfunctionalization. *BMC Evol Biol* **7**, 213.

79. Francino, M.P. (2005) An adaptive radiation model for the origin of new gene functions. *Nat Genet* **37**, 573-7.

80. Dorus, S., Gilbert, S.L., Forster, M.L., Barndt, R.J. & Lahn, B.T. (2003) The CDY-related gene family: coordinated evolution in copy number, expression profile and protein sequence. *Hum Mol Genet* **12**, 1643-50.

81. Des Marais, D. & Rausher, M. (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, 5.

82. Hittinger, C.T. & Carroll, S.B. (2007) Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**, 677-81.

83. Scannell, D.R. & Wolfe, K.H. (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Research* **18**, 137-47.

84. Nowak, M.A., Boerlijst, M.C., Cooke, J. & Smith, J.M. (1997) Evolution of genetic redundancy. *Nature* **388**, 167-71.

85. Papp, B., Pal, C. & Hurst, L.D. (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**, 661-4.

86. Kuepfer, L., Sauer, U. & Blank, L.M. (2005) Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome Res* **15**, 1421-30.

87. Pereira-Leal, J.B. & Teichmann, S.A. (2005) Novel specificities emerge by stepwise duplication of functional modules. *Genome Res* **15**, 552-9.

88. Wagner, A. (2000) Robustness against mutations in genetic networks of yeast. *Nat Genet* **24**, 355-61.

89. Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D.P., Zipperlen, P. & Ahringer, J. (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-7.

90. Su, Z. & Gu, X. (2008) Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. *J Mol Evol* **67**, 705-9.

91. Gu, Z., Nicolae, D., Lu, H.S. & Li, W.H. (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* **18**, 609-13.

92. Kafri, R., Bar-Even, A. & Pilpel, Y. (2005) Transcription control reprogramming in genetic backup circuits. *Nat Genet* **37**, 295-9.

93. Kafri, R., Levy, M. & Pilpel, Y. (2006) The regulatory utilization of genetic redundancy through responsive backup circuits. *Proc Natl Acad Sci U S A* **103**, 11653-8.

94. Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J. & Barkai, N. (2005) Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* **309**, 938-40.

95. He, X. & Zhang, J. (2006) Higher duplicability of less important genes in yeast genomes. *Mol Biol Evol* **23**, 144-51.

96. He, X. & Zhang, J. (2006) Transcriptional reprogramming and backup between duplicate genes: is it a genomewide phenomenon? *Genetics* **172**, 1363-7.

97. Kafri, R., Dahan, O., Levy, J. & Pilpel, Y. (2008) Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *Proc Natl Acad Sci U S A* **105**, 1243-8.

98. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54-61.

99. Musso, G., Zhang, Z. & Emili, A. (2007) Retention of protein complex membership by ancient duplicated gene products in budding yeast. *Trends Genet* **23**, 266-269.

100. Baudot, A., Jacq, B. & Brun, C. (2004) A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein-protein interaction network. *Genome Biol* **5**, R76.

101. Wagner, A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* **18**, 1283-92.

102. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J.M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7.

103. Musso, G.A., Zhang, Z. & Emili, A. (2007) Experimental and computational procedures for the assessment of protein complexes on a genome-wide scale. *Chem Rev* **107**, 3585-600.

104. Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J. & Ruepp, A. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32**, D41-4.

105. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9.

106. Maltsev, N., Glass, E.M., Ovchinnikova, G. & Gu, Z. (2005) Molecular mechanisms involved in robustness of yeast central metabolism against null mutations. *J Biochem* **137**, 177-87.

107. Marland, E., Prachumwat, A., Maltsev, N., Gu, Z. & Li, W.H. (2004) Higher gene duplicabilities for metabolic proteins than for nonmetabolic proteins in yeast and E. coli. *J Mol Evol* **59**, 806-14.

108. Bateson, W. *Mendel's Principles of Heredity*, 460 (1907).

109. Phillips, P.C. (2008) Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**, 855-67.

110. Fisher, R.A. (1918) The correlation between relatives on the supposition of Mendelian inhereitance. *Philos Trans R Soc Edin* **52**, 399-433.

111. Wright, S. (1932) The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proc Sixth Intl Cong Genetics* **1**, 356-66.

112. Roth, F.P., Lipshitz, H.D. & Andrews, B.J. (2009) Q&A: epistasis. *J Biol* **8**, 35.

113. Novick, P., Osmond, B.C. & Botstein, D. (1989) Suppressors of yeast actin mutations. *Genetics* **121**, 659-74.

114. Hartman, J.L.t., Garvik, B.M. & Hartwell, L.H. (2001) Principles for the buffering of genetic variation. *Science* **291**, 1001-4.

115. St Onge, R.P., Mani, R., Oh, J., Proctor, M., Fung, E., Davis, R.W., Nislow, C., Roth, F.P. & Giaever, G. (2007) Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* **39**, 199-206.

116. Mani, R., St Onge, R.P., Hartman, J.L.t., Giaever, G. & Roth, F.P. (2008) Defining genetic interaction. *Proc Natl Acad Sci U S A* **105**, 3461-6.

117. Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S.G., Cyert, M., Hughes, T.R., Boone, C. & Andrews, B. (2006) Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* **21**, 319-30.

118. Jones, S., Jedd, G., Kahn, R.A., Franzusoff, A., Bartolini, F. & Segev, N. (1999) Genetic interactions in yeast between Ypt GTPases and Arf guanine nucleotide exchangers. *Genetics* **152**, 1543-56.

119. Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., Andrews, B., Tyers, M. & Boone, C. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364-8.

120. Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D.S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan,

N., Li, Z., Levinson, J.N., Lu, H., Menard, P., Munyana, C., Parsons, A.B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A.M., Shapiro, J., Sheikh, B., Suter, B., Wong, S.L., Zhang, L.V., Zhu, H., Burd, C.G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F.P., Brown, G.W., Andrews, B., Bussey, H. & Boone, C. (2004) Global mapping of the yeast genetic interaction network. *Science* **303**, 808-13.

121. Pan, X., Yuan, D.S., Ooi, S.L., Wang, X., Sookhai-Mahadeo, S., Meluh, P. & Boeke, J.D. (2007) dSLAM analysis of genome-wide genetic interactions in Saccharomyces cerevisiae. *Methods* **41**, 206-21.

122. Collins, S.R., Miller, K.M., Maas, N.L., Roguev, A., Fillingham, J., Chu, C.S., Schuldiner, M., Gebbia, M., Recht, J., Shales, M., Ding, H., Xu, H., Han, J., Ingvarsdottir, K., Cheng, B., Andrews, B., Boone, C., Berger, S.L., Hieter, P., Zhang, Z., Brown, G.W., Ingles, C.J., Emili, A., Allis, C.D., Toczyski, D.P., Weissman, J.S., Greenblatt, J.F. & Krogan, N.J. (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**, 806-10.

123. Musso, G., Costanzo, M., Huangfu, M., Smith, A.M., Paw, J., San Luis, B.J., Boone, C., Giaever, G., Nislow, C., Emili, A. & Zhang, Z. (2008) The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res* **18**, 1092-9.

124. Deluna, A., Vetsigian, K., Shoresh, N., Hegreness, M., Colón-González, M., Chao, S. & Kishony, R. (2008) Exposing the fitness contribution of duplicated genes. *Nat Genet* **40**, 676-681.

125. Fields, S. & Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-6.

126. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* **17**, 1030-2.

127. Alberts, B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291-4.

128. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. (1999) From molecular to modular cell biology. *Nature* **402**, C47-52.

129. Ge, H., Liu, Z., Church, G.M. & Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29**, 482-6.

130. Jansen, R., Greenbaum, D. & Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12**, 37-46.

131. Carmi, S., Levanon, E.Y., Havlin, S. & Eisenberg, E. (2006) Connectivity and expression in protein networks: proteins in a complex are uniformly expressed. *Phys Rev E Stat Nonlin Soft Matter Phys* **73**, 031909.

132. Batada, N.N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hurst, L.D. & Tyers, M. (2006) Stratus not altocumulus: A new view of the yeast protein interaction network. *PLOS Biology* **4**, 1-12.

133. Sprinzak, E., Altuvia, Y. & Margalit, H. (2006) Characterization and prediction of protein-protein interactions within and between complexes. *Proc Natl Acad Sci U S A* **103**, 14718-23.

134. Li, B., Vilardell, J. & Warner, J.R. (1996) An RNA structure involved in feedback regulation of splicing and of translation is critical for biological fitness. *Proc Natl Acad Sci U S A* **93**, 1596-600.

135. Veitia, R.A. (2002) Exploring the etiology of haploinsufficiency. *Bioessays* **24**, 175-84.

136. Papp, B., Pál, C. & Hurst, L.D. (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194-7.

137. Freeling, M. & Thomas, B.C. (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research* **16**, 805-14.

138. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. & Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.

139. Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hon, G.C., Myers, C.L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., Stark, C., Ho, Y., Botstein, D., Andrews, B., Boone, C., Troyanskya, O.G., Ideker, T., Dolinski, K., Batada, N.N. & Tyers, M. (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol* **5**, 11.

140. de Lichtenberg, U., Jensen, L.J., Brunak, S. & Bork, P. (2005) Dynamic complex formation during the yeast cell cycle. *Science* **307**, 724-7.

141. Greenbaum, D., Jansen, R. & Gerstein, M. (2002) Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* **18**, 585-96.

142. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-6.

143. Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-7.

144. Batagelj, V., Mrvar, A (1998) Pajek - Program for large network analysis. *Conncections* **21**, 47-57.

145. Goll, J. & Uetz, P. (2006) The elusive yeast interactome. *Genome Biol* **7**, 223.

146. Yates, J., Eng, JK, Clauser, KR, Burlingame, AL (1996) Search of sequence databases with uninterpreted high-energy collision-induces dissociation spectra of peptides. *Journal of the American Society of Mass Spectrometry* **7**, 1089-1098.

147. Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. & Feldman, M.W. (2002) Evolutionary rate in the protein interaction network. *Science* **296**, 750-2.

148. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41-2.

149. Sharp, P.M. & Li, W.H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281-95.

150. Pal, C., Papp, B. & Hurst, L.D. (2001) Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927-31.

151. Koonin, E.V. (2005) Paralogs and mutational robustness linked through transcriptional reprogramming. *Bioessays* **27**, 865-8.

152. Lynch, M., O'Hely, M., Walsh, B. & Force, A. (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**, 1789-804.

153. Harrison, R., Papp, B., Pal, C., Oliver, S.G. & Delneri, D. (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci U S A* **104**, 2307-12.

154. Ihmels, J., Collins, S.R., Schuldiner, M., Krogan, N.J. & Weissman, J.S. (2007) Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. *Mol Syst Biol* **3**, 86.

155. Ho, C.H., Magtanong, L., Barker, S.L., Gresham, D., Nishimura, S., Natarajan, P., Koh, J.L., Porter, J., Gray, C.A., Andersen, R.J., Giaever, G., Nislow, C., Andrews, B., Botstein, D., Graham, T.R., Yoshida, M. & Boone, C. (2009) A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. *Nat Biotechnol* **27**, 369-77.

156. Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bähler, J., Wood, V., Dolinski, K. & Tyers, M.

(2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36**, D637-40.

157. Tirosh, I. & Barkai, N. (2007) Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol* **8**, R50.

158. Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I.H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics* **20**, 2479-81.

159. Pierce, S.E., Fung, E.L., Jaramillo, D.F., Chu, A.M., Davis, R.W., Nislow, C. & Giaever, G. (2006) A unique and universal molecular barcode array. *Nat Methods* **3**, 601-3.

160. Gu, Z., Cavalcanti, A., Chen, F.C., Bouman, P. & Li, W.H. (2002) Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Mol Biol Evol* **19**, 256-62.

161. Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85-94.

162. Maere, S., Heymans, K. & Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448-9.

163. McAlister, L. & Holland, M.J. (1982) Targeted deletion of a yeast enolase structural gene. Identification and isolation of yeast enolase isozymes. *J Biol Chem* **257**, 7181-8.

164. Yashiroda, H., Kaida, D., Toh-e, A. & Kikuchi, Y. (1998) The PY-motif of Bul1 protein is essential for growth of Saccharomyces cerevisiae under various stress conditions. *Gene* **225**, 39-46.

165. Estruch, F. & Carlson, M. (1993) Two homologous zinc finger genes identified by multicopy suppression in a SNF1 protein kinase mutant of Saccharomyces cerevisiae. *Mol Cell Biol* **13**, 3872-81.

166. Stucka, R., Dequin, S., Salmon, J.M. & Gancedo, C. (1991) DNA sequences in chromosomes II and VII code for pyruvate carboxylase isoenzymes in Saccharomyces cerevisiae: analysis of pyruvate carboxylase-deficient strains. *Mol Gen Genet* **229**, 307-15.

167. Griffioen, G., Swinnen, S. & Thevelein, J.M. (2003) Feedback inhibition on cell wall integrity signaling by Zds1 involves Gsk3 phosphorylation of a cAMP-dependent protein kinase regulatory subunit. *J Biol Chem* **278**, 23460-71.

168. Wang, C.W., Hamamoto, S., Orci, L. & Schekman, R. (2006) Exomer: A coat complex for transport of select membrane proteins from the trans-Golgi network to the plasma membrane in yeast. *J Cell Biol* **174**, 973-83.

169. Mendoza, I., Quintero, F.J., Bressan, R.A., Hasegawa, P.M. & Pardo, J.M. (1996) Activated calcineurin confers high tolerance to ion stress and alters the budding pattern and cell morphology of yeast cells. *J Biol Chem* **271**, 23061-7.

170. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. & Natale, D.A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.

171. Jamieson, D.J. & Beggs, J.D. (1991) A suppressor of yeast spp81/ded1 mutations encodes a very similar putative ATP-dependent RNA helicase. *Mol Microbiol* **5**, 805-12.

172. Belhumeur, P., Lee, A., Tam, R., DiPaolo, T., Fortin, N. & Clark, M.W. (1993) GSP1 and GSP2, genetic suppressors of the prp20-1 mutant in Saccharomyces cerevisiae: GTP-binding proteins involved in the maintenance of nuclear organization. *Mol Cell Biol* **13**, 2152-61.

173. Dean, E.J., Davis, J.C., Davis, R.W. & Petrov, D.A. (2008) Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet* **4**, e1000113.

174. Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D. & Brown, P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**, 4241-57.

175. Pena-Castillo, L. & Hughes, T.R. (2007) Why are there still over 1000 uncharacterized yeast genes? *Genetics* **176**, 7-14.

176. Keszenman, D.J., Salvo, V.A. & Nunes, E. (1992) Effects of bleomycin on growth kinetics and survival of Saccharomyces cerevisiae: a model of repair pathways. *J Bacteriol* **174**, 3125-32.

177. Montelone, B.A., Hoekstra, M.F. & Malone, R.E. (1988) Spontaneous mitotic recombination in yeast: the hyper-recombinational rem1 mutations are alleles of the RAD3 gene. *Genetics* **119**, 289-301.

178. Izawa, S., Maeda, K., Sugiyama, K., Mano, J., Inoue, Y. & Kimura, A. (1999) Thioredoxin deficiency causes the constitutive activation of Yap1, an AP-1-like transcription factor in Saccharomyces cerevisiae. *J Biol Chem* **274**, 28459-65.

179. Cook, J.G., Bardwell, L., Kron, S.J. & Thorner, J. (1996) Two novel targets of the MAP kinase Kss1 are negative regulators of invasive growth in the yeast Saccharomyces cerevisiae. *Genes Dev* **10**, 2831-48.

180. Loeb, L.A., Loeb, K.R. & Anderson, J.P. (2003) Multiple mutations and cancer. *Proc Natl Acad Sci U S A* **100**, 776-81.

181. Fiedler, D., Braberg, H., Mehta, M., Chechik, G., Cagney, G., Mukherjee, P., Silva, A.C., Shales, M., Collins, S.R., van Wageningen, S., Kemmeren, P., Holstege, F.C., Weissman, J.S., Keogh, M.C., Koller, D., Shokat, K.M. & Krogan, N.J. (2009) Functional organization of the S. cerevisiae phosphorylation network. *Cell* **136**, 952-63.

182. Pittenger, R.C., Wolfe, R.N., Hoehn, P.N., Daily, W.A. & McGuire, J.M. (1953) Hygromycin I. Preliminary studies in the production and biologic activity on a new antibiotic. *Antibiot. Chemother.* **3**, 1268-78.

183. Tong, A.H. & Boone, C. (2006) Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods Mol Biol* **313**, 171-92.

184. Hentges, P., Van Driessche, B., Tafforeau, L., Vandenhaute, J. & Carr, A.M. (2005) Three novel antibiotic marker cassettes for gene disruption and marker switching in Schizosaccharomyces pombe. *Yeast* **22**, 1013-9.

185. Goldstein, A.L. & McCusker, J.H. (1999) Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. *Yeast* **15**, 1541-53.

186. Jorgensen, P., Nelson, B., Robinson, M.D., Chen, Y., Andrews, B., Tyers, M. & Boone, C. (2002) High-resolution genetic mapping with ordered arrays of Saccharomyces cerevisiae deletion mutants. *Genetics* **162**, 1091-9.

187. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504.

188. Berkey, C.D., Vyas, V.K. & Carlson, M. (2004) Nrg1 and nrg2 transcriptional repressors are differently regulated in response to carbon source. *Eukaryotic Cell* **3**, 311-7.

189. Lamb, T.M. & Mitchell, A.P. (2003) The transcription factor Rim101p governs ion tolerance and cell differentiation by direct repression of the regulatory genes NRG1 and SMP1 in Saccharomyces cerevisiae. *Mol Cell Biol* **23**, 677-86.

190. Kuchin, S., Vyas, V.K. & Carlson, M. (2002) Snf1 protein kinase and the repressors Nrg1 and Nrg2 regulate FLO11, haploid invasive growth, and diploid pseudohyphal differentiation. *Mol Cell Biol* **22**, 3994-4000.

191. Elbert, M., Rossi, G. & Brennwald, P. (2005) The yeast par-1 homologs kin1 and kin2 show genetic and physical interactions with components of the exocytic machinery. *Mol Biol Cell* **16**, 532-49.

192. Levin, D.E. & Bishop, J.M. (1990) A putative protein kinase gene (kin1+) is important for growth polarity in Schizosaccharomyces pombe. *Proc Natl Acad Sci U S A* **87**, 8272-6.

193. Matheos, D.P., Kingsbury, T.J., Ahsan, U.S. & Cunningham, K.W. (1997) Tcn1p/Crz1p, a calcineurin-dependent transcription factor that differentially regulates gene expression in Saccharomyces cerevisiae. *Genes Dev* **11**, 3445-58.

194. Stathopoulos, A.M. & Cyert, M.S. (1997) Calcineurin acts through the CRZ1/TCN1-encoded transcription factor to regulate gene expression in yeast. *Genes Dev* **11**, 3432-44.

195. Seoighe, C. & Wolfe, K. (1999) Yeast genome evolution in the post-genome era. *Curr Opin Microbiol* **2**, 548-54.

196. Escriva, H., Bertrand, S., Germain, P., Robinson-Rechavi, M., Umbhauer, M., Cartry, J., Duffraisse, M., Holland, L., Gronemeyer, H. & Laudet, V. (2006) Neofunctionalization in vertebrates: the example of retinoic acid receptors. *PLoS Genet* **2**, e102.

197. Hillenmeyer, M.E., Fung, E., Wildenhain, J., Pierce, S.E., Hoon, S., Lee, W., Proctor, M., St Onge, R.P., Tyers, M., Koller, D., Altman, R.B., Davis, R.W., Nislow, C. & Giaever, G. (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362-5.

198. Dixon, S.J., Fedyshyn, Y., Koh, J.L., Prasad, T.S., Chahwan, C., Chua, G., Toufighi, K., Baryshnikova, A., Hayles, J., Hoe, K.L., Kim, D.U., Park, H.O., Myers, C.L., Pandey, A., Durocher, D., Andrews, B.J. & Boone, C. (2008) Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc Natl Acad Sci U S A* **105**, 16653-8.

199. Taylor, D.L. (2007) Past, present, and future of high content screening and the field of cellomics. *Methods Mol Biol* **356**, 3-18.

200. Ohya, Y., Sese, J., Yukawa, M., Sano, F., Nakatani, Y., Saito, T.L., Saka, A., Fukuda, T., Ishihara, S., Oka, S., Suzuki, G., Watanabe, M., Hirata, A., Ohtani, M., Sawai, H., Fraysse, N., Latge, J.P., Francois, J.M., Aebi, M., Tanaka, S., Muramatsu, S., Araki, H., Sonoike, K., Nogami, S. & Morishita, S. (2005) High-dimensional and large-scale phenotyping of yeast mutants. *Proc Natl Acad Sci U S A* **102**, 19015-20.

201. Vizeacoumar, F.J., Chong, Y., Boone, C. & Andrews, B.J. (2009) A picture is worth a thousand words: genomics to phenomics in the yeast Saccharomyces cerevisiae. *FEBS Lett* **583**, 1656-61.

202. Sopko, R., Papp, B., Oliver, S.G. & Andrews, B.J. (2006) Phenotypic activation to discover biological pathways and kinase substrates. *Cell Cycle* **5**, 1397-402.

203. Kainth, P., Sassi, H.E., Pena-Castillo, L., Chua, G., Hughes, T.R. & Andrews, B. (2009) Comprehensive genetic analysis of transcription factor pathways using a dual reporter gene system in budding yeast. *Methods* **48**, 258-64.

204. Wong, S.L. & Roth, F.P. (2005) Transcriptional compensation for gene loss plays a minor role in maintaining genetic robustness in Saccharomyces cerevisiae. *Genetics* **171**, 829-33.

205. Chen, X.J., Cong, Y.S., Wesolowski-Louvel, M., Li, Y.Y. & Fukuhara, H. (1992) Characterization of a circular plasmid from the yeast Kluyveromyces waltii. *J Gen Microbiol* **138**, 337-45.

206. Willins, D.A., Shimer, G.H., Jr. & Cottarel, G. (2002) A system for deletion and complementation of Candida glabrata genes amenable to high-throughput application. *Gene* **292**, 141-9.

207. Drinnenberg, I.A., Weinberg, D.E., Xie, K.T., Mower, J.P., Wolfe, K.H., Fink, G.R. & Bartel, D.P. (2009) RNAi in Budding Yeast. *Science* **326**, 544-50.

208. Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. & Mello, C.C. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-11.

209. Hannon, G.J. & Rossi, J.J. (2004) Unlocking the potential of the human genome with RNA interference. *Nature* **431**, 371-8.

210. Kolfschoten, I.G., van Leeuwen, B., Berns, K., Mullenders, J., Beijersbergen, R.L., Bernards, R., Voorhoeve, P.M. & Agami, R. (2005) A genetic screen identifies PITX1 as a suppressor of RAS activity and tumorigenicity. *Cell* **121**, 849-58.

211. Westbrook, T.F., Martin, E.S., Schlabach, M.R., Leng, Y., Liang, A.C., Feng, B., Zhao, J.J., Roberts, T.M., Mandel, G., Hannon, G.J., Depinho, R.A., Chin, L. & Elledge, S.J. (2005) A genetic screen for candidate tumor suppressors identifies REST. *Cell* **121**, 837-48.

212. Moffat, J., Grueneberg, D.A., Yang, X., Kim, S.Y., Kloepfer, A.M., Hinkle, G., Piqani, B., Eisenhaure, T.M., Luo, B., Grenier, J.K., Carpenter, A.E., Foo, S.Y., Stewart, S.A., Stockwell, B.R., Hacohen, N., Hahn, W.C., Lander, E.S., Sabatini, D.M. & Root, D.E. (2006) A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283-98.

213. Lehner, B., Crombie, C., Tischler, J., Fortunato, A. & Fraser, A.G. (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet* **38**, 896-903.

214. Sancak, Y., Peterson, T.R., Shaul, Y.D., Lindquist, R.A., Thoreen, C.C., Bar-Peled, L. & Sabatini, D.M. (2008) The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science* **320**, 1496-501.

215. Yoo, J.W., Kim, S. & Lee, D.K. (2008) Competition potency of siRNA is specified by the 5'-half sequence of the guide strand. *Biochem Biophys Res Commun* **367**, 78-83.

216. Trewyn, R.W., Nakamura, K.D., O'Connor, M.L. & Parks, L.W. (1973) An interaction between S-adenosyl-L-methionine and pyridoxal 5'-phosphate, and its effect on Saccharomyces cerevisiae. *Biochim Biophys Acta* **327**, 336-44.

217. Penninckx, M. (1975) Interaction between arginase and L-ornithine carbamoyltransferase in Saccharomyces cerevisiae. The regulatory sites of arginase. *Eur J Biochem* **58**, 533-8.

218. Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M. & Seraphin, B. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**, 218-29.

219. Bader, G.D. & Hogue, C.W. (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol* **20**, 991-7.

220. Landazuri, M.O., Vara-Vega, A., Viton, M., Cuevas, Y. & del Peso, L. (2006) Analysis of HIF-prolyl hydroxylases binding to substrates. *Biochem Biophys Res Commun* **351**, 313-20.

221. Kim, J.S., Rho, B., Lee, T.H., Lee, J.M., Kim, S.J. & Park, J.H. (2006) The interaction of hepatitis B virus X protein and protein phosphatase type 2 Calpha and its effect on IL-6. *Biochem Biophys Res Commun* **351**, 253-8.

222. Solaz-Fuster, M.C., Gimeno-Alcaniz, J.V., Casado, M. & Sanz, P. (2006) TRIP6 transcriptional co-activator is a novel substrate of AMP-activated protein kinase. *Cell Signal* **18**, 1702-12.

223. Vidal, M. & Legrain, P. (1999) Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res* **27**, 919-29.

224. Colland, F., Jacq, X., Trouplin, V., Mougin, C., Groizeleau, C., Hamburger, A., Meil, A., Wojcik, J., Legrain, P. & Gauthier, J.M. (2004) Functional proteomics mapping of a human signaling pathway. *Genome Res* **14**, 1324-32.

225. Obrdlik, P., El-Bakkoury, M., Hamacher, T., Cappellaro, C., Vilarino, C., Fleischer, C., Ellerbrok, H., Kamuzinzi, R., Ledent, V., Blaudez, D., Sanders, D., Revuelta, J.L., Boles, E., Andre, B. & Frommer, W.B. (2004) K+ channel interactions detected by a genetic system optimized for systematic studies of membrane protein interactions. *Proc Natl Acad Sci U S A* **101**, 12242-7.

226. Walhout, A.J. & Vidal, M. (2001) High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* **24**, 297-306.

227. Bartel, P.L., Roecklein, J.A., SenGupta, D. & Fields, S. (1996) A protein linkage map of Escherichia coli bacteriophage T7. *Nat Genet* **12**, 72-7.

228. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-74.

229. Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., Goldberg, D.S., Li, N., Martinez, M., Rual, J.F., Lamesch, P., Xu, L., Tewari, M., Wong, S.L., Zhang, L.V., Berriz, G.F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H.W., Elewa, A., Baumgartner, B., Rose, D.J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S.E., Saxton, W.M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K.C., Harper, J.W., Cusick, M.E., Roth, F.P., Hill, D.E. & Vidal, M. (2004) A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540-3.

230. Stanyon, C.A., Liu, G., Mangiola, B.A., Patel, N., Giot, L., Kuang, B., Zhang, H., Zhong, J. & Finley, R.L., Jr. (2004) A Drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biol* **5**, R96.

231. Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., Jacq, B., Arpin, M., Bellaiche, Y., Bellusci, S., Benaroch, P., Bornens, M., Chanet, R., Chavrier, P., Delattre, O., Doye, V., Fehon, R., Faye, G., Galli, T., Girault, J.A., Goud, B., de Gunzburg, J., Johannes, L., Junier, M.P., Mirouse, V., Mukherjee, A., Papadopoulo, D., Perez, F., Plessis, A., Rosse, C., Saule, S., Stoppa-Lyonnet, D., Vincent, A., White, M., Legrain, P., Wojcik, J., Camonis, J. & Daviet, L. (2005) Protein interaction mapping: a Drosophila case study. *Genome Res* **15**, 376-84.

232. Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P. & Vidal, M. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-8.

233. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. & Wanker, E.E. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957-68.

234. Causier, B. & Davies, B. (2002) Analysing protein-protein interactions with the yeast two-hybrid system. *Plant Mol Biol* **50**, 855-70.

235. Bauch, A. & Superti-Furga, G. (2006) Charting protein complexes, signaling pathways, and networks in the immune system. *Immunol Rev* **210**, 187-207.

236. Ito, T., Ota, K., Kubota, H., Yamaguchi, Y., Chiba, T., Sakuraba, K. & Yoshida, M. (2002) Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol Cell Proteomics* **1**, 561-6.

237. Johnsson, N. & Varshavsky, A. (1994) Split ubiquitin as a sensor of protein interactions in vivo. *Proc Natl Acad Sci U S A* **91**, 10340-4.

238. Miller, J.P., Lo, R.S., Ben-Hur, A., Desmarais, C., Stagljar, I., Noble, W.S. & Fields, S. (2005) Large-scale identification of yeast integral membrane protein interactions. *Proc Natl Acad Sci U S A* **102**, 12123-8.

239. Phizicky, E.M. & Fields, S. (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol Rev* **59**, 94-123.

240. Fritze, C.E. & Anderson, T.R. (2000) Epitope tagging: general method for tracking recombinant proteins. *Methods Enzymol* **327**, 3-16.

241. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. & Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-3.

242. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7.

243. Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J. & Emili, A. (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433**, 531-7.

244. Rubio, V., Shen, Y., Saijo, Y., Liu, Y., Gusmaroli, G., Dinesh-Kumar, S.P. & Deng, X.W. (2005) An alternative tandem affinity purification strategy applied to Arabidopsis protein complex isolation. *Plant J* **41**, 767-78.

245. Veraksa, A., Bauer, A. & Artavanis-Tsakonas, S. (2005) Analyzing protein complexes in Drosophila with tandem affinity purification-mass spectrometry. *Dev Dyn* **232**, 827-34.

246. Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P.O., Bergamini, G., Croughton, K., Cruciat, C., Eberhard, D., Gagneur, J., Ghidelli, S., Hopf, C., Huhse, B., Mangano, R., Michon, A.M., Schirle, M., Schlegl, J., Schwab, M., Stein, M.A., Bauer, A., Casari, G., Drewes, G., Gavin, A.C., Jackson, D.B., Joberty, G., Neubauer, G., Rick, J., Kuster, B. & Superti-Furga, G. (2004) A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat Cell Biol* **6**, 97-105.

247. Brajenovic, M., Joberty, G., Kuster, B., Bouwmeester, T. & Drewes, G. (2004) Comprehensive proteomic analysis of human Par protein complexes reveals an interconnected protein network. *J Biol Chem* **279**, 12804-11.

248. Dziembowski, A. & Seraphin, B. (2004) Recent developments in the analysis of protein complexes. *FEBS Lett* **556**, 1-6.

249. Jensen, L.J., Saric, J. & Bork, P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* **7**, 119-29.

250. Sugiyama, K., Hatano, K., Yoshikawa, M. & Uemura, S. (2003) Extracting information on protein-protein interactions from biological literature based on machine learning approaches. *Genome Informatics* **14**, 699-700.

251. Ding, J., Berleant, D., Nettleton, D. & Wurtele, E. (2002) Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomput*, 326-37.

252. Hahn, U., Romacker, M. & Schulz, S. (2002) Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. *Pac Symp Biocomput*, 338-49.

253. Hirschman, L., Park, J.C., Tsujii, J., Wong, L. & Wu, C.H. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics* **18**, 1553-61.

254. Hahn, U., Romacker, M. & Schulz, S. (2002) MEDSYNDIKATE--a natural language system for the extraction of medical information from findings reports. *Int J Med Inform* **67**, 63-74.

255. Malik, R., Franke, L. & Siebes, A. (2006) Combination of text-mining algorithms increases the performance. *Bioinformatics* **22**, 2151-7.

256. Breitkreutz, B.J., Stark, C. & Tyers, M. (2003) The GRID: the General Repository for Interaction Datasets. *Genome Biol* **4**, R23.

257. Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K., Pawson, T. & Hogue, C.W. (2003) PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* **4**, 11.

258. Schwikowski, B., Uetz, P. & Fields, S. (2000) A network of protein-protein interactions in yeast. *Nat Biotechnol* **18**, 1257-61.

259. Jonsson, P.F. & Bates, P.A. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics*.

260. Ramani, A.K., Bunescu, R.C., Mooney, R.J. & Marcotte, E.M. (2005) Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol* **6**, R40.

261. Backer, E. & Jain, A. (1981) A clustering performance measure based on fuzzy set decomposition. *IEEE Trans. Pattern Anal. Mach. Intell* **PAMI-3**, 66-75.

262. Collins, S.R., Kemmeren, P., Zhao, X.C., Greenblatt, J.F., Spencer, F., Holstege, F.C., Weissman, J.S. & Krogan, N.J. (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* **6**, 439-50.

263. Krogan, N.J., Peng, W.T., Cagney, G., Robinson, M.D., Haw, R., Zhong, G., Guo, X., Zhang, X., Canadien, V., Richards, D.P., Beattie, B.K., Lalev, A., Zhang, W., Davierwala, A.P., Mnaimneh, S., Starostine, A., Tikuisis, A.P., Grigull, J., Datta, N., Bray, J.E., Hughes, T.R., Emili, A. & Greenblatt, J.F. (2004) High-definition macromolecular composition of yeast RNA-processing complexes. *Mol Cell* **13**, 225-39.

264. Arnau, V., Mars, S. & Marin, I. (2005) Iterative cluster analysis of protein interaction data. *Bioinformatics* **21**, 364-78.

265. Krause, R., von Mering, C. & Bork, P. (2003) A comprehensive set of protein complexes in yeast: mining large scale protein-protein interaction screens. *Bioinformatics* **19**, 1901-8.

266. Brohee, S. & van Helden, J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**, 488.

267. Bader, G.D. & Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2.

268. Milgram, S. (1967) The small world problem. *Psychology Today* **2**, 60-67.

269. Maslov, S. & Sneppen, K. (2002) Specificity and stability in topology of protein networks. *Science* **296**, 910-3.

270. Przulj, N., Corneil, D.G. & Jurisica, I. (2004) Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508-15.

271. Khanin, R. & Wit, E. (2006) How scale-free are biological networks. *J Comput Biol* **13**, 810-8.

272. Wunderlich, Z. & Mirny, L.A. (2006) Using the topology of metabolic networks to predict viability of mutant strains. *Biophys J* **91**, 2304-11.

273. Cusick, M.E., Klitgord, N., Vidal, M. & Hill, D.E. (2005) Interactome: gateway into systems biology. *Hum Mol Genet* **14 Spec No. 2**, R171-81.