Going Fast and Getting Lost: Gene Duplication in Yeast

by Devin Scannell

A thesis submitted to The University of Dublin for the degree of Doctor of Philosophy

Supervised by Professor Kenneth H. Wolfe

Department of Genetics
Trinity College
University of Dublin

2007

Declaration

This thesis is submitted by the undersigned for the degree of Doctor of Philosophy at the University of Dublin. It has not been submitted as an exercise for a degree in any other university. Apart from the advice, assistance, and joint effort mentioned in the acknowledgements and in the text, this thesis is entirely my own work. I agree that the library may lend or copy this thesis freely upon request.

Devin Scannell.

December, 2006.

Acknowledgements

Many thanks to everyone in the Wolfe lab past and present: Kevin, Jon, Simon, Avril, Gavin, Brian, Marie, Meg, Nora, Matt, Ailís, Greg and the anyone else I may have forgotten. You made it a lot of fun and it was a pleasure to work with all of you.

Many thanks also to Mike Lynch and everyone in Bloomington, Indiana: Desiree, Doug, Sarah, Melania, Nicole, Jhin, Michelle and Matt. I had a great time in the Midwest and learned more in six in six weeks than I have done in a long time.

Also thanks to all my collaborators along the way, on the various projects that did work out, did not work out and will work out in the future: Kasper Møller, Mark Johnston, Jeff Sabina and Barry Williams. A special thanks to Andrew Lloyd for encouraging me to go for it with the *TUB2* project – much appreciated.

Thanks to Ken Wolfe for being a great supervisor and mentor – critical, insightful and the kind of person that is easy to learn from. And there was the trip to the Galapagos… Thanks for all the opportunities to travel and meet people as well as to learn.

Finally, I want to dedicate this to my parents. They always believed that education was the way to a happier life and always backed me whatever I wanted to do. Many thanks for everything.

Summary

In this thesis I study how duplicate gene pairs created by a whole-genome duplication in an ancestor of several yeast species were resolved. I show that gene duplication may lead not just to the emergence of new gene functions, but also to the emergence of new species. I used comparative genomics between ten hemiascomycete yeasts to study both the process of gene loss that caused over 4000 genes to be rapidly lost from the *S. cerevisiae* genome and the altered molecular evolution of those genes that have been retained in duplicate. Among the genomes I studied was that of the non-model hemiascomycete yeast *Kluyveromyces polysporus*, which was sequenced, annotated and analyzed during the course of this thesis.

Three major findings arise from this work.

First, I show for the first time that both members of duplicate gene pairs experience a burst of protein sequence evolution in the immediate aftermath of duplication (Chapter 4). Following this burst, purifying selection is rapidly restored on one duplicate while the other continues to evolve rapidly for at least 100 Myr. Because gene duplication is often associated with the emergence of new biological functions, the altered evolutionary dynamics of duplicate genes identified in this work may be the molecular signature of evolutionary innovation.

Second, the work presented in Chapter 3 paints the most complete picture yet of gene loss in any organism. I show that when duplicate gene pairs are returned to single-copy the "choice" of which copy to lose is not random – as duplicate genes diverge in sequence, one member becomes favoured and will preferentially be retained, while the other is more likely to be lost. By contrast, for very young duplicate genes or those that are involved in highly conservative biological processes, selection cannot differentiate between the two copies and both are equally likely to be lost in independent lineages. The observation that natural selection can distinguish between copies of some duplicate gene pairs but not others suggests an analogy with the Nearly Neutral Theory, in which random genetic drift determines the fate of alleles whose selective coefficients are similar but natural selection is the dominant force when one allele confers a significant advantage over the other. A "nearly equal" theory of duplicate gene resolution may describe the process of gene loss after duplication.

Finally, I have provided the first evidence for a model of speciation in which ancestrally duplicated loci that have undergone reciprocal gene loss between a pair of species behave as Dobzhansky–Muller incompatibilities and contribute to reproductive isolation (Chapter 2). Because some spores produced after a hybridization between two lineages that have fixed null alleles at alternative copies of an ancestrally duplicated locus may inherit only these null gene copies and will be inviable assuming the gene is essential, gene duplication followed by gene loss can be a significant barrier to gene flow. Indeed, reciprocal gene loss gene loss at just 16 unlinked ancestrally duplicated loci is sufficient to reduce spore viability to ~1% but I show that hundreds of reciprocal gene losses separate all the major lineages that emerged after the WGD in yeast.

Abbreviations

WGD	whole-genome duplication
YGOB	(the) Yeast Gene Order Browser
RGL	reciprocal gene loss
DGP	duplicate gene pair
Myr	millions of years
ORF	open reading frame

Chapter 1. Introduction

In this chapter I provide general information on the hemiascomycete yeasts, in particular, the model organism *S. cerevisiae*. I also summarize what is known about the non-model yeast *K. polysporus*. In addition, I review the literature on gene duplication and provide some background on the genetics of speciation.

1.1 The True Yeasts

The hemiascomycete phylum spans an evolutionary range as great as the vertebrates (Dujon *et al.*, 2004) and includes at least nineteen species whose genomes have been sequenced to high-quality draft level or better (Chapter 3; Wolfe, 2006). Amongst these are multiple *Saccharomyces* yeasts that are very closely related to the model yeast, *S. cerevisiae*, several more distantly related *Kluyveromyces* and *Candida* species and one member of the genus *Yarrowia,* which is very diverged from *S. cerevisiae.* To put these in context, the protein sequence divergence between *S. cerevisiae* and *Y. lipolytica* is comparable to that between human and the sea squirt, *Ciona intestinalis*, while *S. cerevisiae* and *S. bayanus* are roughly as diverged as human and mouse. However, whereas the genomes of human and mouse are very large (>2.5Gb), repeat rich (>40%) and at least moderately rearranged with respect to one another (>295 rearrangements; Waterston *et al.*, 2002), the genomes of these two *Saccharomyces* are compact (~14Mb), almost repeat free (<1%) and virtually collinear (three inversions and five reciprocal translocations; Kellis *et al.*, 2003). The significance of these differences for the study of genome evolution can hardly be overstated.

Hemiascomycete yeasts are also of interest for practical reasons. As noted above, several species of *Candida* have been sequenced and four of these – *C. albicans*, *C. tropicalis, C. dubliniensis* and *C. glabrata* – are important or emerging pathogens of humans (Logue *et al.*, 2005). Surprisingly however, these yeasts are not monophyletic with respect to the *Saccharomyces*. *C. glabrata* is much more closely related to *S. cerevisiae* than to the other *Candida* species (Barns *et al.*, 1991), suggesting that it may be relatively easy for yeasts to become pathogenic. Consistent with this, a recent study showed that the loss of the kynureine pathway (*BNA* genes) from *C. glabrata* (it is present in *S. cereisiae*) contributes significantly to virulence in this species (Domergue *et al.*, 2005). It is worth noting however, that yeasts have also proven to be of considerable medical benefit. *S. cerevisae* is a useful model for human diseases including Parkinson's disease (Willingham *et al.*, 2003), HIV-1 drug resistance (Nissley *et al.*, 1998), and potentially cancer (Pfeiffer *et al.*, 2001). In addition, becasue the core gene complement is so well conserved between yeast and humans it has even been possible to use comparative genomics between different hemiascomycete yeasts to track down human mitochondrial disease genes (Nussbaum, 2005).

Hemiascomycete yeasts are also of interest for a wide range of industrial purposes, of which three stand out: Biomolecular synthesis, environmental applications and food production. The first category includes both straight forward 'bulk' protein production (Moller *et al.*, 2004, Moller *et al.*, 2001b) and the use of heavily engineered yeasts strains to perform sophisticated molecular synthesis. The recent production of the anti-malarial drug precursor artemisinic acid in *S. cerevisiae* is a notable example of the latter (Ro *et al.*, 2006). Amongst the environmental applications are bioremediation (*e.g. Debaryomyces* and other yeasts can potentially be used to clean up oil spills (see refs in Wong and Wolfe, 2006) and the production of fuel grade ethanol from xylose using either evolved strains of *S. cerevisiae* (Sherlock *et al.*, 2006) or non-model hemiascomycetes such as *Pichia stipitis* (Jeffries, 2006). It is worth pointing out that since many of these applications rely on identifying genomic differences (such as segmental duplications; Bond *et al.*, 2004) between natural and artificially evolved yeast strains, or making inferences about natural non-model yeasts (Woolfit *et al.*, 2006), that comparative genomics is likely to play an increasingly significant role in future applied research.

Finally, hemiascomycete yeasts are central to the production of bread and alcoholic beverages such as wine and beer (Hansen and Piskur, 2003). Indeed, the ability of yeast to ferment sugars to alcohol is almost certainly the original reason for the domestication of *S. cerevisiae* (Fay and Benavides, 2005a) and the subsequent study of brewing strains by researchers such as Winge and Lindegren (reviewed in Mortimer, 1993b and Mortimer, 1993a) ultimately led to the adoption of *S. cerevisiae* as a model organism. In this regard, and given the focus of this thesis on gene duplication, it seems appropriate to point out that one of the key genomic changes underlying the fermentative lifestyle of modern *S. cerevisiae* is a gene duplication (Thomson *et al.*, 2005). The duplication of the ancestral alcohol dehydrogenase gene to produce *ADH1* (which favors the production of ethanol) and *ADH2* (which favours the reverse reaction) allowed the ancestor of *S. cerevisiae* to outcompete its competitors by rapidly converting available sugars to ethanol via *ADH1* and later consuming the ethanol via *ADH2* (Thomson *et al.*, 2005, Pfeiffer *et al.*, 2001). Since *ADH1* and *ADH2* were not created by the whole-genome duplication they are not amongst the duplicate gene pairs studied during the course of this research, but, as perhaps the best example of neofunctionalization in the hemiascomycete yeasts, they will be discussed again later (Section 1.2.2.3).

1.1.1 Phylogenetics

1.1.1.1 Phylogenetic position of the hemiascomycete yeasts

Eukaryotes are often divided into four kingdoms, animals, plants, fungi and protists. Although abundant evidence indicates that protists are paraphyletic the concentration of model organism research in the remaining three kingdoms (which are monophyletic) makes this a useful simplification (Hedges, 2002). The opisthokonta hypothesis, which states that fungi and animals are more closely related to each other than either are to plants, is typically taken to describe the relationships among these three kingdoms and is well supported by molecular data (Baldauf *et al.*, 2000, James *et al.*, 2006). It also accords with the taxonomic view, which distinguishes autotrophic plants from heterotrophic fungi (and animals) and distinguishes fungi from animals on the basis that that the former absorb (rather than ingest) food (Ingold and Hudson, 1961). Unlike either plants or animals, fungi have cell walls composed of varying proportions of chitin and β-glucan (Ingold and Hudson, 1961).

The number of extant fungal species is thought to be in the millions although only about 80,000 have been described (Hedges, 2002). The known species are typically divided into five phyla (*Ascomycota, Basidiomycota, Glomeromycota, Zygomycota* and *Chytridomycota*) although some of these may not be monophyletic (James *et al.*, 2006). The largest phylum, the ascomycetes or ‘sac’ fungi, is defined by the production of a specialized structure – the ascus - that surrounds the spores formed during meiosis (James *et al.*, 2006). The monophyly of the ascomycetes is well supported by molecular data (Galagan *et al.*, 2005b). The ascomycetes were traditionally further divided into euascomycetes (hyphal fungi such as *Neurospora crassa*) and hemiascomycetes (yeasts such as *Saccharomyces cerevisiae*) but molecular evidence showing that *Schizosaccharomyces pombe* is an outgroup to both of these taxa has prompted the proposal of a third ascomycete class (Heckman *et al.*, 2001), the archiascomycetes (Figure 1.1). Additional molecular sequence data have supported the novel classification (James *et* *al.*, 2006, Galagan *et al.*, 2005b), indicating that unicellular yeasts have evolved from multicellular hyphal progenitors more than once. Nevertheless, the work in this thesis focuses exclusively on the study of the hemiascomycete or ‘true’ yeasts.

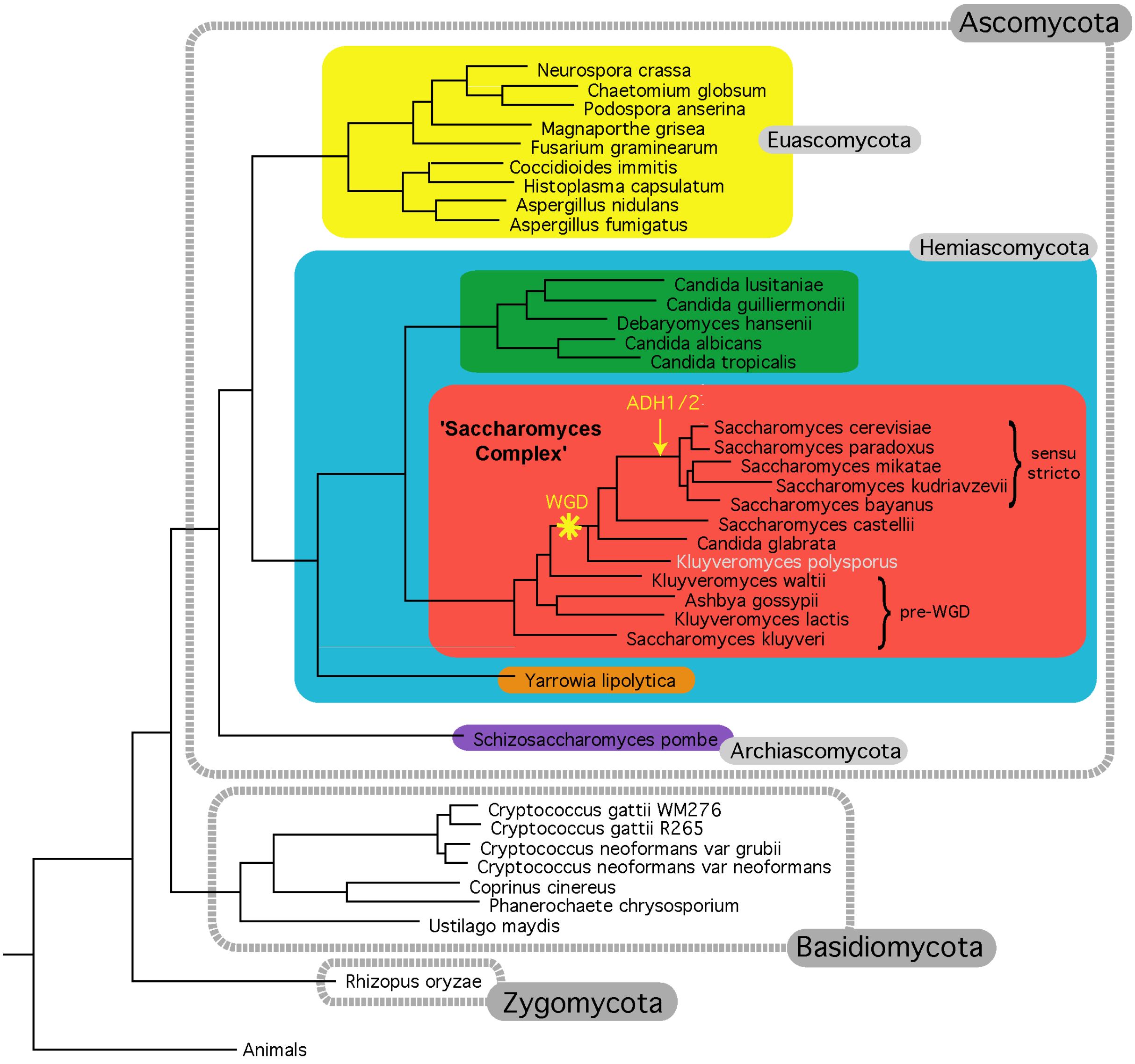


Figure 1.1 Phylogenetic tree including several of the fungal species whose genomes have either been completely sequenced or are in the process of being sequenced. Adapted from http://fungal.genome.duke.edu/ (Jason Stajich). The original tree was reconstructed using the protein sequences of 165 genes for which putative orthologs could be found in all the species shown (except *K. polysporus*; see below) as well as ten animal species and the slime mold *Dictyostelium discoideum* (which was used to root the tree). The approximate position of the whole-genome duplication (WGD; yellow asterisk), the ADH1/2 duplication (yellow arrow) and the divergence of *Kluyveromyces polysporus* (based on data presented in Chapter 3) were added by the author. Fungal phyla are labeled in dark grey and ascomycete classes are labeled in light grey. The section in red has been labeled the '*Saccharomyces* complex' because it overlaps with the tree of the '*Saccharomyces* complex' (Kurtzman and Robnett, 2003) that is reproduced in Appendix VI.

1.1.1.2 Phylogenetic relationships amongst the hemiascomycete yeasts

Although over 700 species of hemiascomycetes have been described, all of the genomes sequenced thus far (with the exception of *Y. lipolytica*) are members of the family *Saccharomycetaceae* and fall into one of two clusters. The *Candida* cluster consists primarily of *Candida* species but also includes yeasts such as *D. hansenii*, while the second cluster is comprised primarily of species from the *Saccharomyces* and *Kluyveromyces* genera (green and red backgrounds respectively in Figure 1.1). For consistency with previous authors the second cluster is referred to as the "*Saccharomyces* complex" (Kurtzman and Robnett, 2003). Because the separate branches leading to the *Saccharomyces* complex and the *Candida* cluster are very well supported (Galagan *et al.*, 2005b), sequences from the *Candida* cluster have been used to root phylogenetic trees of sequences from the *Saccharomyces* complex throughout this thesis.

The major division within the *Saccharomyces* complex is between those yeasts whose common ancestor underwent a whole-genome duplication (WGD; Section 1.2.1.2) and those that diverged prior to this event (Figure 1.1; yellow asterisk). We term these post-WGD and pre-WGD yeasts respectively. The relationshisp among the sequenced post-WGD yeasts have been well studied (Rokas *et al.*, 2003, Scannell *et al.*, 2006a) but it is useful to distinguish the yeasts that (like *S. cerevisiae*) possess the *ADH1/2* duplication noted previously (Figure 1.1) from those that do not. The former are capable of rapid anaerobic growth, preferentially ferment glucose in the presence of oxygen and are referred to as *Saccharomyces* sensu stricto yeasts. As well as being phenotypically very similar to *S. cerevisiae,* these yeasts have almost collinear genomes and are partially inter-fertile (Kellis *et al.*, 2003, Cliften *et al.*, 2003, Greig *et al.*, 2002b). By contrast, the lineages that diverged from one another between the WGD and the time of the *ADH1/2* duplication (represented by *K. polysporus*, *S. castellii, C. glabrata* and *S. cerevisiae*; Figure 1.1) are phenotypically, genomically and reproductively diverged. In this thesis I focus on these lineages, which diverged from one another in the aftermath of the WGD, and compare how duplicate genes pairs created by the whole-genome duplication were resolved in each lineage. However, I also make extensive use of the sequenced pre-WGD yeast genomes and the sensu stricto yeast *S. bayanus*.

1.1.2 Genomics

1.1.2.1 Genome structure of hemiascomycete yeasts

S. cerevisiae was the first eukaryote to have both a whole chromosome (Oliver *et al.*, 1992) and its entire genome sequenced (Goffeau *et al.*, 1996) in large part because its genome was estimated to be smaller and simpler than those of most other model eukaryotes (Dujon, 1996). This has proven to be correct and the sequenced genomes of other ascomycetes have served both to emphasize how similar are all hemiascomycete genomes (Dujon *et al.*, 2004) and how different they are to the genomes of other eukaryotes, including other ascomycetes (Galagan *et al.*, 2005a). For instance, the average hemiascomycete genome contains only half as many genes (5,000 *versus* 10,000) and one third as much DNA (14Mb *versus* 40Mb) as the euascomycete *N. crassa* (Galagan *et al.*, 2003). Similarly, it also has less than one tenth as many introns as the archiascomycete *S. pombe* (in 4% of genes *versus* in 40% of genes) and purportedly simpler gene regulation (Wood *et al.*, 2002). It is possible that the unusual compactness and relative simplicity of hemiascomycete genomes are a consequence of selection for rapid growth (Dujon *et al.*, 2004).

If we accept *S. cerevisiae* as a representative example then in total around 70% of the yeast genome is likely to be protein-coding (approximately one gene every 2 Kb) with an additional 15% being transcribed into RNA (David *et al.*, 2006). These additional sequences include UTRs, RNAs with familiar functions (rRNAs, tRNAs, snRNAs, *TLC1* etc) and RNAs involved in transcriptional interference (Martens *et al.*, 2004) or other processes reminiscent of the complex transcriptional architecture of metazoans (Katayama *et al.*, 2005, Carninci *et al.*, 2005). The remainder of the genome is made up of structural elements (telomeres, centromeres and ARSs; Hirschman *et al.*, 2006), approximately 50 retrotransposons (Ty elements; Kim *et al.*, 1998), a small number of elements associated with mating-type switching and silencing (the recombination enhancer RE and the I and E silencer elements at the HM loci) (Haber, 1998) and intergenic sequences. Although several studies (Kellis *et al.*, 2003, Cliften *et al.*, 2003) have used comparative genomics to identify 'dictionaries' of cis-regulatory elements that occur in intergenic regions (the most recent of which defined ~300 cis-regulatory elements that occur approximately 30,000 times in total; Wang and Stormo, 2005), it remains unknown whether selection operates on a specified number of discrete cis-regulatory elements in each promoter (perhaps with spacing requirements; Sudarsanam *et al.*, 2002) or whether it imposes more diffuse constraints on the entire promoter sequence as has recently been proposed (Tanay, 2006). The recent demonstration that nucleosome positioning is at least partly determined by DNA sequences (Segal *et al.*, 2006, Ioshikhes *et al.*, 2006) supports the view that promoters may contain less freely evolving DNA than previously suspected (Halligan and Keightley, 2006) and it is not inconceivable that the vast majority of the yeast genome is functionally constrained to some degree.

In addition to selection on genome content there is evidence of selection on gene order in the *S. cerevisiae* genome. For instance, it has been shown that (nonhomologous) genes that operate in the same biological pathway tend to be much closer together in the genome than expected by chance (Lee and Sonnhammer, 2003, Teichmann and Veitia, 2004). A recent study that used comparative genomics to trace the assembly of the six-gene *DAL* cluster implicates natural selection in the assembly of these clusters but also suggested a role for recombination in maintaining them (Wong and Wolfe, 2005). Despite being close to a sub-telomeric region (normally associated with high recombination rates) the *DAL* cluster is one of the least recombinogenic regions in the entire yeast genome. It has been proposed that selection for linkage also underlies the over-representation of essential genes in low recombination regions of the genomes (Pal and Hurst, 2003). However, it is possible that this is a consequence of another process that causes essential genes to be linked to centromeres (Taxis *et al.*, 2005). Finally, it has also been shown that genes with similar expression patterns cluster in the yeast genome. It has long been known that this was the case for neighbouring genes (Cohen *et al.*, 2000) but a recent study indicates that expression domains extending up to thirty genes may also exist (Lercher and Hurst, 2006). This is consistent with reporter studies indicating that the position of genes within the nucleus is a major determinant of expression (Taddei *et al.*, 2006). The relationships between these different factors (recombination, function and expression) and the precise forces determining gene order are unlikely to be unraveled soon but the remarkable conservation of synteny among yeasts in the *Saccharomyces* complex implies that the yeast genome is a highly ordered place.

1.1.2.2 Genome content of hemiascomycete yeasts

Although the compaction and order of hemiascomycete genomes may not be typical of all eukaryotes (Semon and Duret, 2005), their gene content undoubtedly is. Hemiascomycete genomes encode representatives of most of the signature eukaryotic gene families (cytoskeletal proteins, ubiquitin ligases etc; Rubin *et al.*, 2000) and at least 40% of yeast genes have homologues in humans (Lander *et al.*, 2001). Indeed, subunits of many multi-protein complexes such as the COPII vesicle complex (the main component of one of the three essential membrane trafficking systems in eukaryotes) exhibit a one-to-one orthologous relationship between *S. cerevisiae* and humans (Kirchhausen, 2000). Perhaps the most important contribution of yeast however is as a model for understanding core biological processes such as DNA replication, DNA damage repair and recombination, which are conserved throughout eukaryotes (Rubin *et al.*, 2000).

Some differences must exist between yeast and other eukaryotes, however and there are both genes families that are present in a diverse range of other eukaryotes that are absent in yeast and *vice versa*. For instance, many gene families that are important in epigenetic silencing and animal development (such as the *polycomb* genes) are unsurprisingly absent from yeast (Rubin *et al.*, 2000), although in this particular case the study of yeast may yet prove to be informative: the *Drosophila* homolog of *SIR2*, which is involved in epigenetic silencing in yeast, interacts genetically with *polycomb* genes in *Drosophila* (Chopra and Mishra, 2005). Amongst the fungal specific gene families than are those such as the ‘zinc cluster’ transcription factor family (of which GAL4 is the most famous member; MacPherson *et al.*, 2006) and there are also hemiascomycete specific sub-families of more widely distributed gene families such as the YAP (yeast activator protein) family (Fernandes *et al.*, 1997), which is related to *AP-1* in humans.

Given the modest number of genes in yeast genomes, it is perhaps unsurprising that they contain proportionately fewer detectable duplicate genes than animals (Gu *et al.*, 2002). In the case of *S. cerevisiae*, most authors agree that there are around 1800 genes with detectable homology to at least one other gene in the genome (Rubin *et al.*, 2000, Gu *et al.*, 2003) and when whole-genome duplicates are excluded (discussed below; not all of the *circa* 550 pairs can be detected by BLAST (Wolfe, 2004)) it seems likely that around 1000 detectable duplicates remain. This is very similar to the number of genes assigned to families of size two or greater in the pre-WGD yeast *K. lactis* by Dujon et al. (2004). The same comparison of five hemiascomycete yeasts also highlights the fact that most gene families in hemiascomycetes are small, with fewer than 20% of the defined gene families in *K. lactis* having more than four members. This is similar to the situation in *S. cerevisiae* where the few large gene families are found in sub-telomeric regions (e.g. PAU, COS and FLO in S. cerevisiae; up to 27 copies; Fabre *et al.*, 2005, Gu *et al.*, 2002). These families tend to be highly variable and have been proposed to be important for adaptation to novel environments (Landry *et al.*, 2006, Liti *et al.*, 2005).

1.1.2.3 The yeast whole-genome duplication

A feature of the *S. cerevisiae* genome is the preponderance of small gene families relative to many other hemiascomycete yeasts (Dujon *et al.*, 2004). Indeed, the existence of many pairs of apparently functionally redundant duplicate genes had been noted by yeast geneticists and the conservation of gene order among duplicate gene pairs in unlinked regions of the genome was recognized by some to indicate an evolutionary relationship between pairs of chromosomal regions (Melnick and Sherman, 1993). The sequencing of the *S. cerevisiae* genome permitted Wolfe and Shields to show that these duplicated ("sister") chromosomal regions were the product of a single polyploidization event (Wolfe and Shields, 1997b) as envisioned by Susumu Ohno (Ohno, 1970), rather than a series of independent segmental duplications (Llorente *et al.*, 2000). Having used BLAST homology between proteins to identify 55 pairs of sister regions that spanned at least three pairs of duplicate genes, they showed both that triplicated regions were underrepresented (some level of re-duplication is expected if 55 independent segmental duplications occur) and that orientation with respect to the centromere tended to be conserved between sister regions (Wolfe and Shields, 1997b). This is not predicted by a model of random segmental duplication (Wolfe, 2001). The conclusion that the 55 pairs of sister regions were most likely created by a whole-genome duplication event was supported by additional map-based (Wong *et al.*, 2002) and clock-based analyses (Friedman and Hughes, 2001).

The conclusive proof that the distribution of duplicate gene pairs in the yeast genome is primarily the result of an ancient polyploidization event however awaited the sequencing of a yeast species that diverged from the *S. cerevisiae* lineage prior to the whole genome duplication event. Wolfe and Shields had noted on the basis of limited sequence data that it was common for single-copy genes in opposite sister regions to be neighbors in *K. lactis* (Wolfe and Shields, 1997b). This suggested that opposite members of ancestrally duplicated gene pairs had been lost between the two sister regions and predicted that even pairs of sister regions without surviving duplicates would show an "interleaved" pattern of gene loss relative to an appropriate outgroup. This is precisely what was observed when the genomes of both *A. gossypii* (Dietrich *et al.*, 2004) and *K. waltii* (Kellis *et al.*, 2004) were sequenced and analyzed. As can be seen in Figure 1.2 although no duplicate gene pairs have survived between the left arms of *S. cerevisiae* chromosomes 4 and 14, the regions are clearly homologous to a single chromosomal region in *K. waltii* and *A. gossypii* and, by inference, descended from a single chromosomal region in the common ancestor of all of these species. In Chapter 2 and Chapter 3 we show that this same pattern of "interleaving" is observed when gene order in other post-WGD yeasts such as *S. bayanus* (Kellis *et al.*, 2003), *C. glabrata* (Dujon *et al.*, 2004), *S. castellii* (Cliften *et al.*, 2003) and *K. polysporus* (Scannell *et al.*, 2006b) is compared to that in pre-WGD yeast species.

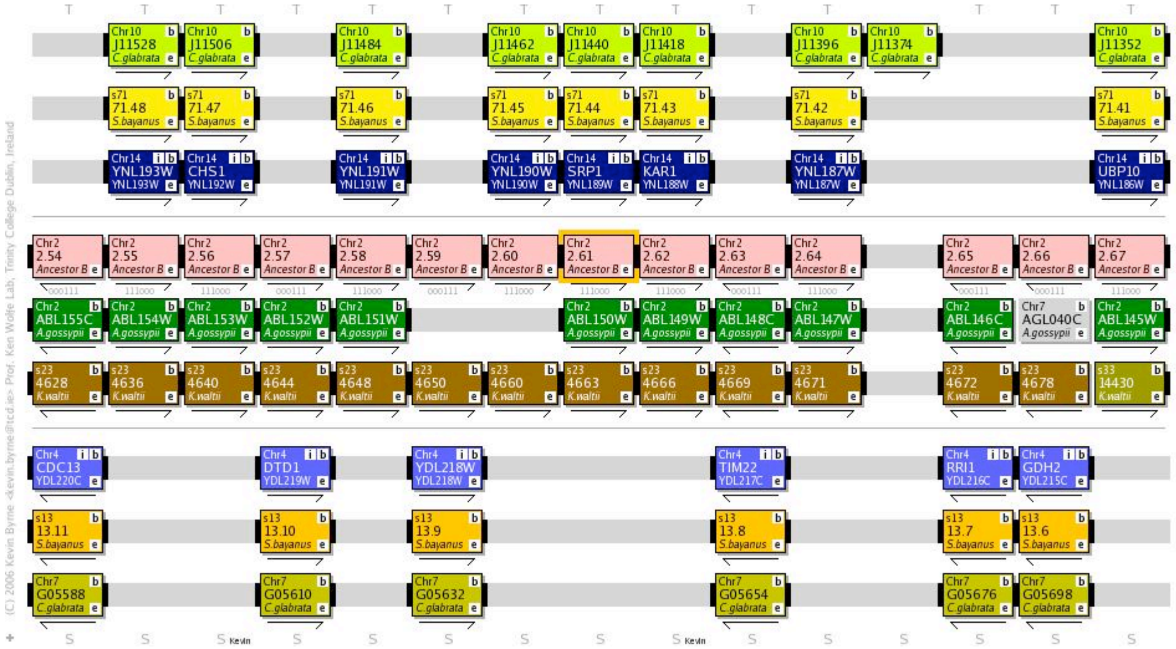


Figure 1.2 Interleaving of single-copy *S. cerevisiae* (blue), *S. bayanus* (yellow) and *C. glabrata* (light green) genes between two sister chromosomal regions when compared to the pre-WGD yeasts *K. waltii* (brown) and *A. gossypii* (dark green). The hypothetical ancestral gene order is also shown (pink). Screenshot from the Yeast Gene Order Browser (Byrne and Wolfe, 2005).

The genome of modern *S. cerevisiae* is dominated by the changes wrought by the WGD and the subsequent diploidization (Wolfe, 2001). The most obvious structural change is the doubling of the number of chromosomes relative to the ancestral pre-WGD yeast (Wolfe, 2006). In addition, the loss of one member of most of the previously duplicated gene pairs and the resulting "interleaving" (Figure 1.2) of single-copy genes between duplicated regions means that around half of all neighboring gene relationships have been altered. It has recently been shown that following a genome duplication event in *Arabidopsis* (Blanc *et al.*, 2003), that the pattern of duplicate loss between sister regions was not random but resulted in the production of "gene rich" and "gene poor" regions (Thomas *et al.*, 2006). Although there is no evidence that this occurred after the WGD in yeast, the orientation bias of neighboring genes has been altered, resulting in an amelioration of the excess of convergently and divergently arrayed neighboring gene pairs (as opposed to tandems) seen in pre-WGD yeasts (Byrnes *et al.*, 2006). This appears to have had an effect on the correlation in expression of neighbouring genes (Byrnes *et al.*, 2006) and raises the obvious question of whether the chromosomal clustering of coexpressed genes in *S. cerevisiae* (or genes involved in the same biological process; Section 1.1.2.1) was affected by the reorganization of neighbouring gene relationships after the WGD.

Whether or not these changes in the organization of the yeast genome turn out to be important, the changes in gene content are likely to have had a significant impact on the biology of *S. cerevisiae*, in particular, in facilitating its adaptation to anaerobic growth and its preference for fermentation of glucose in the presence of oxygen (Wolfe, 2004). For instance, very many of the duplicate gene pairs that have been retained by *S. cerevisiae* are involved in carbohydrate metabolism (Seoighe and Wolfe, 1999a) and the members of many of these pairs are differentially expressed in response to either oxygen (Kwast *et al.*, 2002) or glucose (DeRisi *et al.*, 1997) availability. Experiments on *S. kluyveri* suggest that the ancestral pre-WGD yeast may have possessed a limited ability to grow anaerobically (Moller *et al.*, 2001a) and it will be interesting to see to what extent the functions of *S. cerevisiae* duplicate gene pairs can be complemented by their single-copy orthologs in *S. kluyveri* (van Hoof, 2005) and to what extent they represent true evolutionary innovations. Other genes retained in duplicate since the WGD are discussed in Section 1.2.1.

1.1.3 Lifecycle

1.1.3.1 Lifecycle of *S. cerevisiae*

Yeasts have traditionally been divided into anamorphic (asexual) or teleomorphic (sexual) yeasts and the latter were then further described as being either homothallic (self fertile) or heterothallic (self sterile) depending on whether colonies derived from a single spore could undergo mating. Although these phenotypic designations are being superseded by direct analysis of genomic data (Hull and Johnson, 1999, Tzung *et al.*, 2001, Wong *et al.*, 2003) and quantitative descriptions of the lifestyles of yeasts (e.g. recognition that many “anamorphic” yeasts simply mate rarely; Hull *et al.*, 2000, Miller and Johnson, 2002), by these criteria *S. cerevisiae* is a homothallic teleomorph: Haploids of opposite mating type (see below) mate readily and any one of the haploid spores produced by such a cross may be used to found further diploid colonies. It is worth noting however that whereas lab strains are often maintained as vegetatively growing haploids, wild isolates of *S. cerevisiae* are almost always diploid (Mortimer, 2000). This difference occurs because most lab strains have defective alleles of the *HO* mating-type switching gene. Nevertheless, a recent comparison of the genomes of three sequenced strains of *S. cerevisiae* (S288C, YJM789 and RM11-1a) indicates that the rate of recombination amongst these strains has been on the order of once every 50,000 generations (Ruderfer *et al.*, 2006). Although this leaves open the possibility that the rate of recombination (and hence sporulation and mating) within strains is high, it does suggest that wild strains typically propagate as mitotically dividing diploids. When sporulation does occur (such as following starvation), it is thought that intra-ascus mating follows and that the resulting diploids revert to vegetative reproduction (Taxis *et al.*, 2005). It should be noted also that under nitrogen starvation conditions *S. cerevisiae* diploids can be induced to undergo unipolar budding (as opposed to the bipolar budding typical of diploids and the axial budding typical of haploids). Cells then grow away from the colony as long thin structures known as pseudohyphae in an attempt to forage for food (Gimeno *et al.*, 1992).

Mating in *S. cerevisiae* is controlled by a single locus called the *MAT* locus, which in a haploid may express either of two idiomorphs, **a** or α (Figure 1.3A). **a** and α cells may mate to produce diploids which then possess one idiomorph of each type (they are obligate heterozygotes at the *MAT* locus) and cannot mate but may sporulate. Mating occurs because **a** and α cells express a-specific and α-specific genes respectively. These sets of genes include cell-type specific mating pheromones (a-factor and α-factor respectively), transporters for the export of the relevant mating-factors and receptors for mating factors of the opposite mating type (Johnson, 1995). Thus, **a** cells secrete a-factor which is detected by an a-factor receptor expressed on the surface of α cells and *vice versa*. Once mating factors are detected, the cell cycle is arrested, shmoos (mating projections) are produced by the mating cells and cytogamy is initiated. The later stages of this process are shared between the two haploid cell types and are regulated by a set of haploid-specific genes that are expressed in both **a** and α cells but repressed in diploids.

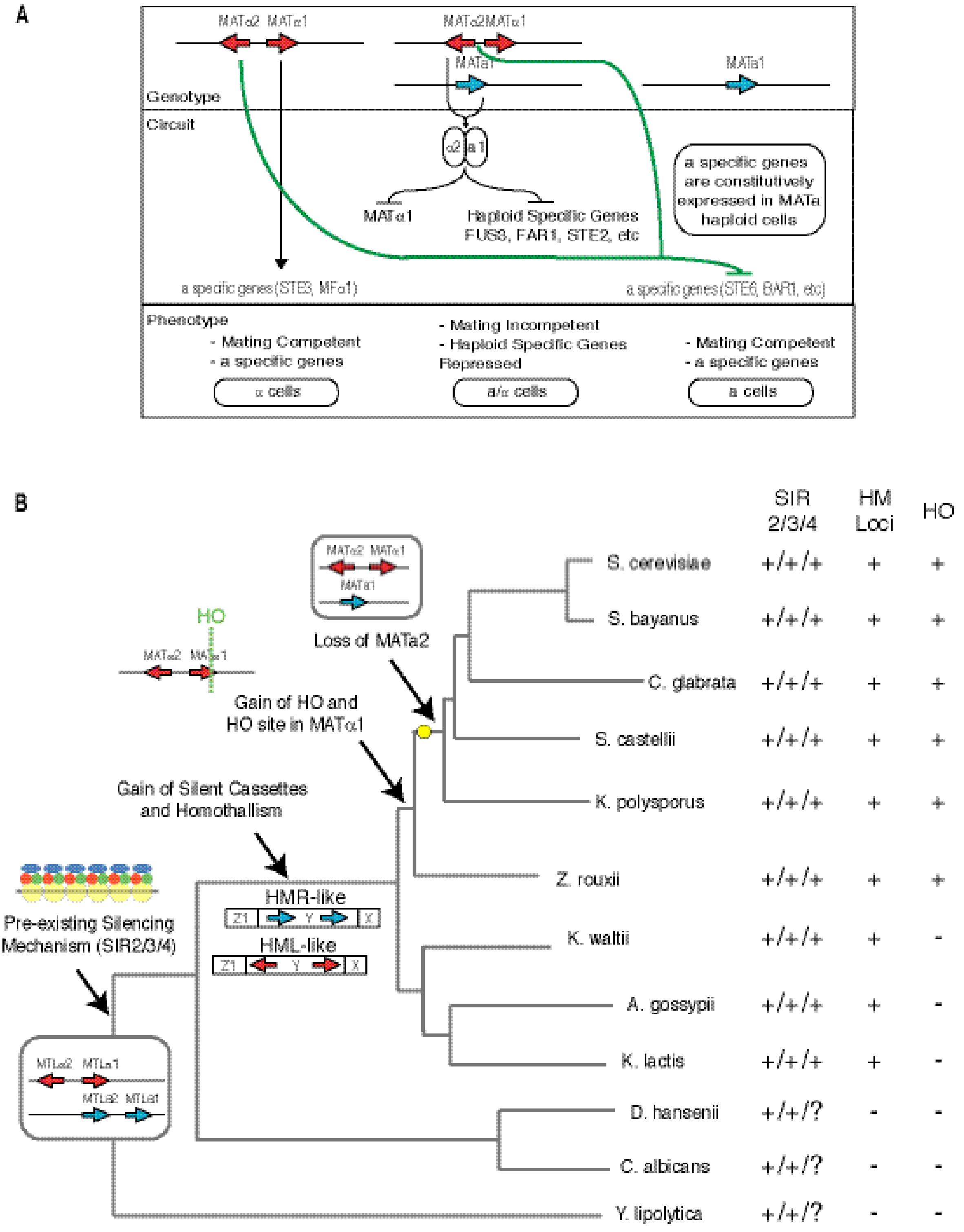


Figure 1.3 The mating system of *S. cerevisiae* (A) and the evolution of mating-type switching in hemiascomycete yeasts (B). Red arrows represent *MAT* (or *MTL*) **α** genes and blue arrows represent *MAT* (or *MTL*) **a** genes. **(A)** Genotypes and phenotypes of the three naturally occurring combinations of *MAT* idiomorphs. Modified from (Scannell and Wolfe, 2004). **(B)** Tree modified from (Scannell *et al.*, 2006a) to include *D. hansenii*, *Y. lipolytica* and *Z. rouxii*. Branch lengths may not be reliable for these taxa. The yellow circle indicates the time of the WGD. Additional data based on (Butler *et al.*, 2004), (Fabre *et al.*, 2005), Gordon and Wolfe (pers. comm.) and results of

BLASTP using *K. waltii* syntenic orthologs of *S. cerevisiae SIR2/3/4* genes against the Genolevures database (http://cbi.labri.fr/Genolevures/blast.php).

The expression of a-specific and α-specific genes in *S. cerevisiae* is regulated by genes at the *MAT* locus. The α idiomorph encodes two genes and in α cells both of these are expressed. α1 operates as an activator of α-specific genes while α2 represses the expression of a-specific genes (Figure 1.3A). In **a** cells the mechanism is even simpler because a-specific genes are expressed by default and α-specific genes are repressed by default. The *MATa* idiomorph encodes a single protein, a1, but it serves no function in haploids. In addition, in both cell types the haploid-specific genes are activated (by *STE12* in many cases; Johnson, 1995). By contrast, in diploids all three sets of haploid genes (a-specific, α-specific and haploid-specific) are actively repressed; α2 represses the expression of a-specific genes; a dimer of α2 and a1 represses the expression of haploid-specific genes; and the same heterodimer represses α1 without which α-specific genes are not expressed. The a1:α2 heterodimer is also an activator of *IME1*, which is the master regulator of meiosis (Kassir *et al.*, 1988). Thus, diploids are asexual (they express neither mating-type) and unlike haploids they may sporulate if appropriately stimulated.

The *MAT* locus alone is sufficient to account for the teleomorphic phenotype of *S. cerevisiae* but cannot explain the fact that it is homothallic. Under the system described above there is no possibility that a single-spore (of either mating-type) could found a new sexual population, since only cells of opposite mating-types may mate and undergo meiosis. However, because *S. cerevisiae* possesses a second genetic system that permits haploids of either mating type to convert to the opposite mating-type, spores can divide, then one can switch mating-type and finally the mother and daughter spores may fuse (Haber, 1998). *S. cerevisiae* can do this because it encodes silent copies of the α and **a** idiomorphs at the left (*HMLα*) and right (*HMRa*) ends of chromosome III respectively and can use these to over-write the information at the *MAT* locus. This over-writing is effectively a gene conversion event that is initiated by the occurrence of a double-strand break at the *MAT* locus. Because in a haploid there is no second *MAT* locus that can be used to direct repair by homologous recombination, one of the *HM* loci is used as the template instead (Haber, 1998). One critical requirement of this system is that the silent cassettes are indeed transcriptionally silent: If they were expressed, haploids would possess the a1:α2 repressor and behave as diploids. This requirement is met by the formation of repressive heterochromatin and the binding of the Silent Information Regulator proteins (*SIR1-4*) at the *HM* loci (Haber, 1998).

Although the system as described is sufficient to give rise to sporadic mating-type switching (and may be similar to that used by *K. lactis*), the actual mechanism employed by *S. cerevisiae* is much more sophisticated. First, *S. cerevisiae* uses the intein-derived endonuclease *HO* to make a cut at the *MAT* locus and initiate the homologous recombination process (Haber and Wolfe, 2005). This increases the rate of mating-type switching by around 10^6 and brings the process under genetic control (Butler *et al.*, 2004). Second, the presence of the recombination enhancer (RE) on the left arm of chromosome III ensures that **a** cells use *HML*α to repair the *MAT* locus rather than *HMRa*, which would result in no net change of mating-type. Similarly, α cells preferentially use *HMRa* to repair the *MAT* locus although the mechanism is different (Haber, 1998). Third, *S. cerevisiae* uses an elaborate cell lineage system to ensure that only half of the cells in a population change at any one time (daughter cell specific repression of HO expression by localizing a repressor, ASH1, to daughter cells; Haber, 1998). These mechanisms ensure that mating-type switching in *S. cerevisiae* is highly efficient and, combined with the axial budding pattern exhibited by haploid cells, provide a means for haploid cells to rapidly become diploid. The low level of outcrossing exhibited by *S. cerevisiae* (noted above) suggests that the evidently strong selective pressure favoring efficient mating-type switching may not be related to the benefits of sex (Keightley and Otto, 2006), but to some advantage conferred by diploidy (Gerstein *et al.*, 2006).

1.1.3.2 Lifecycle of *K. polysporus*

The non-conventional yeast *K. polysporus* that was sequenced during the course of this thesis has a lifecycle that differs in several ways from that of *S. cerevisiae*. In particular, whereas *S. cerevisiae* reproduces primarily by diploid mitoses and produces four or occasionally fewer spores per ascus (Taxis *et al.*, 2005), *K. polysporus* exhibits no appreciable diplophase with zygotes sporulating immediately to produce dozens of spores (van der Walt, 1956). These spores are produced by extra post-meiotic mitoses (Roberts and van der Walt, 1959) and genomic changes that may account for these extra divisions are discussed in Chapter 3. In spite of these differences, the lifecycles of *S. cerevisiae* and *K. polysporus* are broadly similar. They are both haplo-diplontic and both are also homothallic teleomorphs as inferred from the analysis of single-spore cultures (Roberts and van der Walt, 1959). Moreover, the gene content of *K. polysporus* makes it clear that the underlying genetic circuits are similar (a *MAT* locus, *HM* loci, *HO* and cell-type specific genes are present; Appendix X). The presence of *HM* loci is of some importance since *K. polysporus* diverged from *S. cerevisiae* very soon after the WGD (Figure 1.1) and the mechanism we propose for polyploidization requires the presence of silent cassettes for the restoration of fertility after this event (Section 2.3.8).

Early reports suggested that the lifecycle of *K. polysporus* differed from that in *S. cerevisiae* in two additional ways. First, giant multinucleate (and occasionally polyploid) cells were observed. However, these were all derived from a single culture and it is unlikely that this is a feature of the normal *K. polysporus* lifecycle (Roberts and van der Walt, 1959). On the other hand, the conversion of homothallic cells to sterility was observed at a relatively high frequency (Roberts and van der Walt, 1959). In addition, it was shown that although this condition was stable for up to a year, revertants also occurred at a moderate frequency (Roberts and van der Walt, 1959). Although the authors attributed these observations to mutation, the brief description of mating and mating-type switching in *S. cerevisiae* (above) suggests an alternative explanation. Epigenetic silencing (or lack of it) is typically stably inherited but spontaneous changes have been observed in certain genetic backgrounds. It is possible that loss of epigenetic silencing at the *HM* loci may have caused haploid *K. polysporus* cells to behave as a/α diploids and appear sterile for many generations only to subsequently restore silencing. In this respect it is notable that *K. polysporus* possesses no *SIR1* homolog (Appendix X) and that failure to recruit *SIR1* is thought to account for the instability of sub-telomeric silencing relative to *HM* loci in *S. cerevisiae* (Chien *et al.*, 1993).

1.1.3.3 Evolution of mating-type switching and its consequences for polyploidization

Although the evolution of the *MAT* locus and mating-type switching are of interest in their own right (Tsong *et al.*, 2003, Tsong *et al.*, 2006), their main relevance to this thesis is that efficient mating-type switching is required by the model for whole-genome duplication we present in Chapter 2. Since efficient mating-type switching relies on the presence of *HM* loci, *HO* and *SIR* genes (Section 1.1.3.1), it is possible to estimate when homothallism evolved and make inferences about the efficiency of mating-type switching by searching the genomes of sequenced yeasts for homologs of these genes (Butler *et al.*, 2004, Fabre *et al.*, 2005). As can be seen from Figure 1.3B**,** *SIR* homologs are potentially present in all hemiascomycetes suggesting that a mechanism for *HM* silencing existed prior to the *HM* loci themselves. Their ancestral function may be related to their role in sub-telomeric silencing in *S. cerevisiae* (Fabre *et al.*, 2005). By contrast, mating-type switching probably evolved in the common ancestor of *S. cerevisiae* and *K. lactis* (Figure 1.3B), although in the absence of *HO* is likely to have occurred at low frequency. Consistent with this, some but not all strains of *K. lactis* are homothallic (Fabre *et al.*, 2005). Finally, mating-type switching is likely to have become catalyzed by *HO* before the divergence of *S. cerevisiae* and *Z. rouxii* (Figure 1.3B), thus it is likely that efficient mating-type switching emerged just prior to the WGD and has been inherited by all the post-WGD yeasts studied so far (Butler *et al.*, 2004, Haber, 1998).

1.2 Gene Duplication

Gene duplication has been recognized as a potential source of both new genes and new functions since the modern evolutionary synthesis (see references in Long *et al.*, 2003) and before (reviewed by Taylor and Raes, 2004). In this section I briefly review gene duplication in the context of these two phenomena but emphasize that although gene duplication may well be the major contributor of both new genes and new functions to eukaryotic genomes, that they are distinct evolutionary outcomes: Formation of new genes and new functions can occur in the absence of one another and in the absence of gene duplication.

Gene duplication may be considered to consist of three conceptually separable stages; the mutational origin of new gene duplicates; the process of duplicate gene preservation; and the long-term molecular evolution of duplicate gene pairs. In this section I discuss first how redundant genetic material is created, focusing on how the mechanism of gene duplication may affect the subsequent evolution of duplicate gene pairs. The purpose of this is to highlight the distinctive features of whole-genome duplicates. Second, I discuss mechanisms of duplicate gene preservation. Most newly-created duplicate gene pairs will not contribute to long term evolution but a significant minority become preserved in eukaryotic genomes. The circumstances under which this may occur are reviewed. Third, I discuss the little that is known about gene loss after gene duplication. The molecular evolution of gene duplicates is discussed in detail in Chapter 4 and is not repeated here.

1.2.1 The origin of new genes

1.2.1.1 Mechanisms of gene duplication

Ohno famously asserted that, "natural selection merely modified, while redundancy created" (Ohno, 1970) and four mechanisms can be envisaged by which a redundant gene could be obtained: gene duplication (Lynch and Conery, 2000), horizontal gene transfer (Doolittle, 1999, Gogarten and Townsend, 2005), *de novo* creation of a valid gene structure from previously non-functional DNA (Levine *et al.*, 2006), or loss of selection for a gene to perform a previously required function (Duret *et al.*, 2006). In addition, some hybrid mechanisms have been observed, such as the creation of "chimeric" genes from two duplicated genes (Long and Langley, 1993, Long *et al.*, 1999) or from a partially duplicated gene and previously non-coding DNA (Nurminsky *et al.*, 1998, Ranz *et al.*, 2003). The key question therefore is, "what are the relative contributions of these mechanisms?"

In all eukaryotes studied so far it is likely that gene duplication is the primary mechanism of generating novel gene structures (Lynch and Conery, 2000), although it is difficult to estimate the importance of several of the mechanisms outlined above. For instance, most genomes harbor a significant number of "orphan" genes (Dujon *et al.*, 2004, Rubin *et al.*, 2000) for which no convincing homolog can be found. Although many of these are likely to be fast-evolving genes (Cai *et al.*, 2006), it is hard to exclude the possibility that some of them have emerged *de novo* and are functional. Even for genes that have significant homology to other genes, the possibility exists (mainly in the case of metazoan genomes) that they are chimeras of some kind (Ciccarelli *et al.*, 2005). Nevertheless, the presence of thousands of easily detectable duplicate genes in many eukaryotes and the conclusion that the rate of gene duplication is on the same order of magnitude as the per nucleotide substitution rate suggests that gene duplication is by far the most important mechanism for creating redundant genes (Lynch and Conery, 2000). Horizontal gene transfer from bacteria (as opposed to viruses; Bonnaud *et al.*, 2005) has been well studied in both mammals (Salzberg *et al.*, 2001) and yeast (Hall *et al.*, 2005, Dujon *et al.*, 2004) and can account for no more than a handful of cases (Gojkovic *et al.*, 2004).

Duplicate genes arise by a variety of mechanisms: retrotransposition (retrocopies; Schacherer *et al.*, 2004), unequal crossing-over (tandem duplicates; Leh-Louis *et al.*, 2004), replication error (segmental duplications; Schacherer *et al.*, 2005), non-disjunction (aneuploidy; Hughes *et al.*, 2000), and *MAT* locus deletion (polyploidy; See Section 2.3.8; different mechanisms operate in different phyla). The relative contributions of these processes are likely to vary significantly among taxonomic groups. For instance, it is becoming clear that in mammals both retrocopies (Marques *et al.*, 2005) and segmental duplications (Bailey and Eichler, 2006) are important source of genetic redundancy. By contrast, in plants tandem duplicates (Rizzon *et al.*, 2006) and polyploidization (Simillion *et al.*, 2002, Blanc and Wolfe, 2004b) may be of particular relevance and retrogenes may be relatively rare (Benovoy and Drouin, 2006). Nevertheless, following a detailed survey of gene duplication in rice the authors noted that "every conceivable class of duplication that could have happened did in fact happen" (Yu *et al.*, 2005) and any variation in the contribution of different mechanisms should probably be ascribed to quantitative rather than qualitative differences. The composition of multi-gene families (Section 1.1.2.2) and the chromosomal distribution of duplicate genes (Section 1.1.2.3) in *S. cerevisiae* were discussed previously and suggest that when very old duplicates are excluded, unequal crossing-over and polyploidization are the primary sources of redundancy in *S. cerevisiae*.

Duplicate genes created by different mechanisms have very different properties and this can have a significant impact on their subsequent evolution. For instance, retrocopies are often inserted far away from their progenitor locus and do not possess a functional promoter (Cusack and Wolfe, 2006). This may result in either non-expression or mis-expression of newly created genes. Similarly, intra-chromosomal segmental duplications may result in duplicate genes that retain their proximal promoters but no longer have access to distal enhancer elements. Genes created by different mechanism may also differ significantly in their population genetic properties and consequently may differ in their probabilities of preservation. For instance, in moderately sized populations (effective population size ≈ 10^4 - 10^6) the probability of preservation of a pair of duplicate genes by subfunctionalization (Lynch and Force, 2000a) can vary by orders of magnitude depending on whether the duplicates are linked (*e.g.* tandem duplicates) or unlinked (*e.g.* retrocopies). In addition, linked duplicates are less likely to be preserved by neofunctionalization (Lynch *et al.*, 2001).

1.2.1.2 Polyploidization and whole-genome duplicates

Paleopolyploids have been identified in all four eukaryotic kingdoms: plants (Simillion *et al.*, 2002, Blanc and Wolfe, 2004b, Yu *et al.*, 2005, Tuskan *et al.*, 2006), animals (McLysaght *et al.*, 2002, Dehal and Boore, 2005, Amores *et al.*, 1998, Jaillon *et al.*, 2004, Evans *et al.*, 2005), fungi (Wolfe and Shields, 1997b, Dietrich *et al.*, 2004, Kellis *et al.*, 2004), and protists (Laurent Duret, pers. comm.). Nevertheless, polyploidization is likely to have occurred by different mechanisms in different lineages because it relies critically on the mode of reproduction (sexual or asxual) and in the former case on the genetics of the sex determining system. For instance, polyploidy is common amongst plants and parthenogenetic animals since a polyploid may be obtained by meiotic or mitotic non-reduction respectively, without compromising the ability to reproduce again (Otto and Whitton, 2000). By contrast, Müller is credited with realizing that in animals with a sex determination system based on the ratio of autosomes to X chromosomes (*e.g. Drosophila*) polyploids will suffer from aberrant sexual development. Similarly, with an XY/XX (or equivalent) sexual system, dosage compensation may be disturbed (Otto and Whitton, 2000). It has also been suggested that polyploidization in vertebrates is rare simply because it occurs at low frequency and newly-created polyploids have no partner to mate with. In support of this, polyploidization is relatively common in African clawed frogs (Evans *et al.*, 2005), in which the sex of developing young can be determined by temperature (Otto and Whitton, 2000). As in our model for polyploidization in *S. cerevisiae* (Section 2.3.8) this provides a mechanism to restore fertility after polyploidization by permitting two sexes to emerge from a single rare event.

If a polyploid is created by autopolyploidization (the two parental genomes are form the same species) or by allopolyploidization between two moderately diverged genomes then the newly created species will initially be tetraploid. Four alleles will come together at each locus to form quadrivalents at meiosis. As DNA changes accumulate however previously similar chromosome pairs can no longer from quadrivalents and instead form bivalents resulting in a restoration of disomic inheritance (Wolfe, 2001). The relative prevalence of auto- and allopolyploidization are not known and the details of the diplodization process (the reversion from tetrasomic to disomic inheritance) are also far from understood (Wolfe, 2001). It is possible however that the gene loss that follows polyploidization may be the key to both of these processes. For instance, we show in Chapter 3 that the rate of gene loss immediately after the WGD is staggering and it is very possible that this, rather than sequence divergence via point mutation, prevents tetravalents from forming. In addition, analysis of the timecourse of gene loss (Section 2.3.5) may be informative about the nature of the WGD event. Because gene loss is expected to begin immediately after WGD, in the event of an autopolyploidy we expect that 100% duplicate gene retention would coincide with zero percent sequence divergence between surviving duplicate genes. On the other hand, if gene loss began some time after the duplicate sequences begin to diverge, this would suggest an allopolyploidy.

An additional question for which we as yet have no clear answer is, "how many genes should we expect to see returned to single-copy after polyploidization and how many retained in duplicate?" Among the polyploids noted above the percentage of surviving duplicates varies from approximately 10% - 50% but this is largely a function of the amount of the amount of time since polyploidization. Nevertheless, it is notable that similar functional classes of genes have been retained in duplicate after many of these events. For instance, cytosolic ribosomal protein genes have been retained in duplicate in both plants (Blanc and Wolfe, 2004a) and fungi (Seoighe and Wolfe, 1999b). Similarly, transcription factors and/or kinases ("regulatory" genes) were preferentially retained in duplicate after the WGDs in yeast (Seoighe and Wolfe, 1999b), plants (Maere *et al.*, 2005) and animals (Blomme *et al.*, 2006). In addition, it has been shown that duplicates derived from a first WGD event have a significantly increased chance of being re-retained after subsequent WGD events (Seoighe and Gehring, 2004) and that the types of genes that are retained in duplicate after WGD typically do not give rise to duplicates by other mechanisms (Maere *et al.*, 2005). Because the characteristics of genes coding for cytosolic ribosomal protein genes and "regulatory" genes are very different it is likely that more than one explanation will be required to account for these observations. In the former case, it has been proposed that genes coding for ribosomal proteins are retained for increased dosage (Seoighe and Wolfe, 1999b) and that this occurs primarily via WGD because duplication of only a fraction of ribosomal proteins would lead to dosage imbalance and a dominant negative phenotype (Papp *et al.*, 2003).

No plausible explanation has yet been given to explain the preferential retention of kinases and transcription factors in duplicate after WGD, although a number of possibilities can be considered. First, there is some evidence that genes in these functional classes have more complex promoters (Nelson *et al.*, 2004, Iwama and Gojobori, 2004) and thus they may be particularly good candidates for preservation by subfunctionalization (Section 1.2.2.2). In addition, they may not be preserved by smaller scale duplications that fail to duplicate the entire gene and regulatory region (Katju and Lynch, 2003). Second, kinases and transcription factors often have many substrates (Ptacek *et al.*, 2005) or targets (Harbison *et al.*, 2004) respectively. Because target phosphorylation sites or *cis*-regulatory elements are likely to be heterogeneous (*i.e.* all deviating from the consensus in a slightly different way), partial loss-of-function mutations in each member of a pair of duplicates may result in each having high affinity for only a subset of the ancestral targets. This is reminiscent of both coding region subfunctionalization (Dermitzakis and Clark, 2001) and quantitative subfunctionalization (Lynch and Force, 2000a). Third, it is possible that the simultaneous duplication of multiple regulatory genes prevents dysregulation. Indeed, it is notable that kinases and transcription factors are amongst the functional classes most likely to produce a deleterious phenotype when over-expressed in isolation (Sopko *et al.*, 2006) (*contra* the "balance hypothesis" (Papp *et al.*, 2003), genes in multi-protein complexes display no such bias). Finally, both plant (Blanc and Wolfe, 2004a) and yeast (Conant and Wolfe, 2006) researchers have noted that duplicated pathways may become independently expressed following WGD. It is possible that regulatory genes are only recruited when new pathways require regulation.

In addition to biases towards particular molecular functions, duplicate genes created by WGD have two distinctive population genetic properties. First, if we assume that diploidization is rapid then all duplicate gene pairs are effectively unlinked. This will significantly reduce the probability of preservation by subfunctionalization (introduced fully in Section 1.2.2) if the effective population size of the species is large (Lynch *et al.*, 2001). This is because once one of the duplicates has acquired a subfunctionalizing mutation (it loses the ability to perform an essential ancestral subfunction) the second duplicate is absolutely required. For a pair of completely linked duplicates the second duplicate is guaranteed to be present but for unlinked duplicates it may not be, thus resulting in a lethal genotype. Second, most considerations of duplicate gene preservation assume that duplicates are created by single-gene duplications and that the newly created duplicate must then rise from its initial frequency of $1/2N$ (where N is the effective population size) to fixation. This is not the case for duplicates produced by whole genome duplication, which have an initial frequency in the population of 1, because in contrast to all other types of duplication, WGD defines a new population. This can be referred to as fixation-at-birth and is discussed in more detail below.

1.2.2 Mechanisms of duplicate gene preservation

In cases where "mother" and "daughter" members of duplicate gene pairs can be distinguished, one of three fates awaits all newly-created duplicate gene pairs: loss of the "daughter" duplicate, loss of the "mother" duplicate, or retention of both (Lynch *et al.*, 2001). Because the distinction between "mother" and "daughter" duplicates does not apply to whole-genome duplicates and because the principal consequence of loss of the "mother" copy is to contribute to reproductive isolation by relocating a function to the locus at which the "daughter" copy resides, this scenario is discussed in Section 1.3. Here I consider only two outcomes, duplicate gene preservation and return to the single-copy state. Nevertheless, I do not restrict the discussion to whole-genome duplicates but simply highlight how their behavior differs from that of other duplicates as it arises.

1.2.2.1 Models of duplicate gene preservation

A variety of models have been proposed to explain the process by which newly created duplicate gene pairs become preserved, however all are either variants (Gibson and Spring, 1998, Stoltzfus, 1999) or hybrids (Piatigorsky and Wistow, 1991, Hughes, 1994) of three simple ideas; one duplicate evolves a useful novel function while the other performs the ancestral function (neofunctionalization; Ohno, 1970); the duplicates partition ancestral functions between them so that both duplicates are required (subfunctionalization; Force *et al.*, 1999); or duplicates are preserved because unfit genotypes at one locus can be masked by the presence of a functional allele at the other locus (redundancy; Nowak *et al.*, 1997). The three main models and some other variants are shown in Figure 1.4. It is important to note that the aim of each of these models is not to describe the long-term evolution of duplicate gene pairs (He and Zhang, 2005b) but identify why they are preserved initially (Lynch and Katju, 2004). Most progress towards this goal has been made by the subfunctionalization and neofunctionalization models, which are both well supported in the existing literature (see references in Lynch, 2004) and have been studied intensively using population genetic simulations (Lynch and Force, 2000a, Lynch *et al.*, 2001). Neofunctionalization proposes that a wild-type allele present at one of the two duplicate loci performs the ancestral (essential) function, while a neofunctionalized allele at the second duplicate locus confers a selective advantage by performing a novel beneficial function. The allele at the second locus may become neofunctionalized either before or after the locus is founded (discussed below) but in either case it is assumed to occur at the expense of the ancestral function (Figure 1.4). By contrast, subfunctionalization can occur in the complete absence of adaptive evolution (Force *et al.*, 1999). It proposes that following duplication of a locus that performs two (or more) genetically separable essential functions, complementary degenerative mutations result in each of the duplicates being unable to perform a subset of the ancestral functions and thus, both are required for viability (Force *et al.*, 1999). Tissue specific patterns of gene expression under control of distinct enhancer elements are often cited as examples of genetically separable essential functions (Force *et al.*, 1999) but it is likely that subfunctionalization also occurs by reciprocal degenerative coding-region changes (Dermitzakis and Clark, 2001). The primary attraction of subfunctionalization is that unlike neofunctionalization it does not rely on potentially rare gain-of-function mutations and unlike the redundancy model (discussed below) it does not rely on exotic combinations of partial and complete loss-of-function mutation rates. It is also the model that is most obviously consistent with the distribution of fitness effects obtained in routine genetic screens (Jorgensen and Mango, 2002). Subfunctionalization and neofunctionalization are discussed in detail in Section 1.2.2.2.

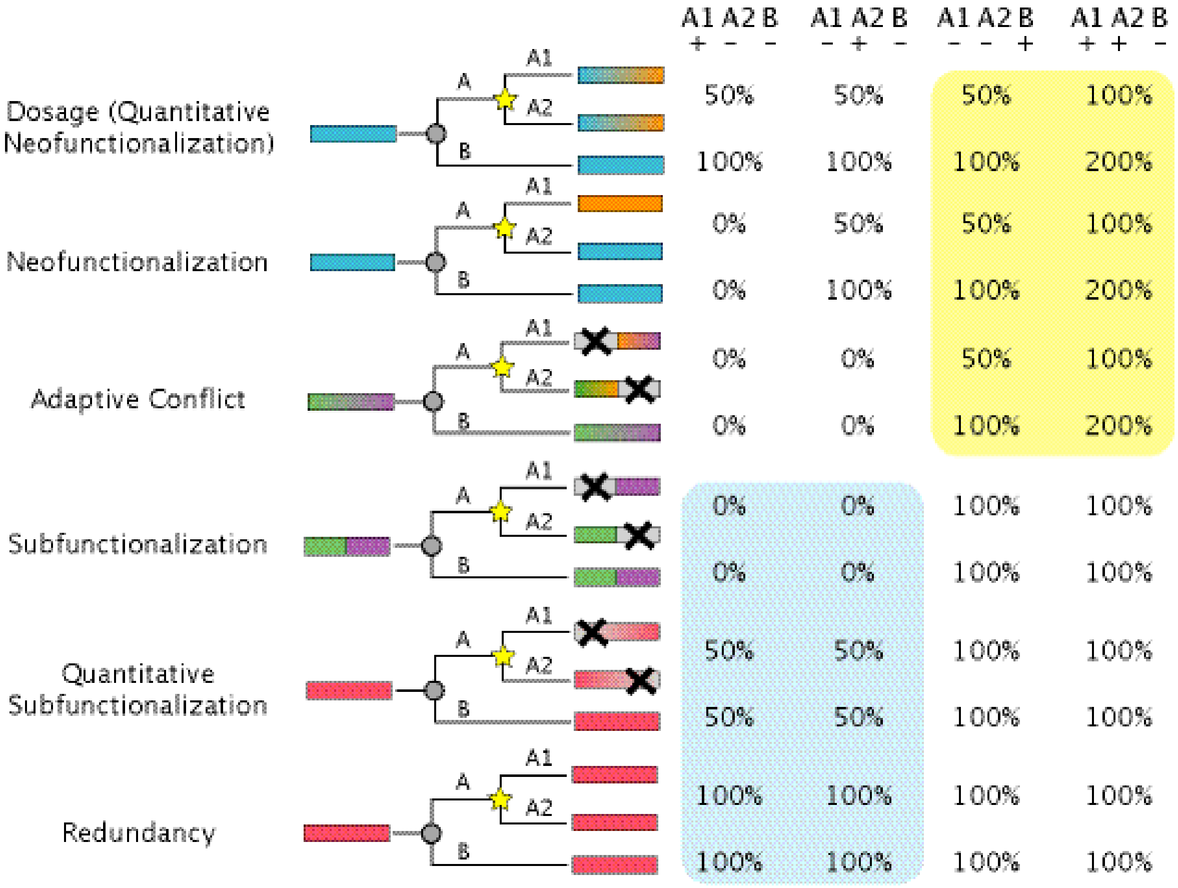


Figure 1.4 Models of duplicate gene pair preservation and the relative fitnesses of different genotypes in inter-specific complementation tests. Boxes represent genes and colors represent functions, except orange, which specifically indicates functions that been gained relative to the ancestral gene. Grey circles indicate speciation events (A and B are orthologous genes) and yellow stars indicate gene duplication events (A1 and A2 are a duplicate gene pair). Black 'X' marks indicate loss-of-function mutations. Blended colors indicate (sub)functions that are not completely genetically separable except in the case of the 'Dosage' model where it indicates that the increased

fitness due to elevated dosage cannot be assigned specifically to either duplicate. Percentages indicate the relative fitness of a particular genotype (indicated at top) compared to the fitness of a wild-type organism of the same species (species A is the top row and species B is the bottom row opposite each model). In all cases, the ancestral function is assumed to be essential, new functions are assumed to double the fitness and partial loss-of-function mutations are assumed to halve the fitness. The pale yellow box highlights complementation tests that can be used to distinguish between models that require neofunctionalization and those that do not (bottom three models). The pale blue box highlights complementation tests that can be used to distinguish among models that do (top three models) or do not require neofunctionalization.

Three versions of the redundancy model exist. In the *näive* version wild-type alleles at both loci in a finite population mutate to null alleles at the same rate and the presence of a duplicate is said to confer an advantage when two null alleles (a lethal genotype in the absence of a duplicate locus) are present at one of the loci. In this model, the selective advantage of the duplicate locus is equal to the mutation rate because the frequency of null homozygotes is expected to be equal to the mutation rate (Lynch, 2004). However, since the mutation rate is the same at the two loci, the selective advantage is effectively cancelled out and one of the two duplicate loci will eventually be lost by drift (Lynch, 2004). In a more sophisticated variant the two loci are not equal. One locus is better at performing the required function but the other experiences a lower mutation rate to null alleles, thus under certain circumstances the system will reach an equilibrium and both loci will be retained indefinitely (Nowak *et al.*, 1997). In the long term this is unlikely to be stable however as movement of the duplicate at the low mutation rate locus to a location with a higher mutation rate or improved performance of the gene at the high mutation rate locus (as could be caused by a gene conversion event between the duplicates) are expected to disrupt the balance and lead to loss of one of the duplicates. Third, a family of models exists that invokes unlikely combinations of mutation rates (null and partial loss-of-function) as a means of duplicate preservation (Nowak *et al.*, 1997, Gibson and Spring, 1998). For instance, Gibson and Spring proposed that a very high rate of mutation to dominant negative missense alleles in duplicate genes and a low rate of mutation to complete loss-of-function alleles would retard loss of duplicate genes and result in large numbers of redundant duplicates (Gibson and Spring, 1998). There is no reason to think that this is correct.

As noted above several additional models of duplicate gene preservation have been proposed, three of which will be considered briefly. First, several authors have suggested that selection for increased dosage may result in duplicate gene preservation (Figure 1.4) and there is evidence that this is the case (Seoighe and Wolfe, 1999a). This may be described as quantitative neofunctionalization since the advantage arises from an increased capacity to perform the ancestral function rather than the ability to perform a novel function *per se*. It is closely related to quantitative subfunctionalization in which both duplicates are required to perform the required function at the ancestral level. Moreover, in the case of both quantitative subfunctionalization and quantitative neofunctionalization there is no reason to believe that the division of labour between the duplicates should be equal (as shown in Figure 1.4): If there is selection for dosage, a genotype in which one duplicate has 80% of the capacity of the ancestral copy and the other has 50% should be favored over the ancestral wildtype genotype (a single copy with 100% capacity).

Second, it has been suggested that prior to duplication genes may perform two (or more) functions that exert a level of pleiotropic constraint on one another, thus preventing one or both functions from being optimized by selection (Piatigorsky and Wistow, 1991, Hughes, 1994). Following duplication each duplicate may accept previously forbidden substitutions that improve their ability to perform one function at the expense of their ability to perform the other ("Adaptive Conflict" in Figure 1.4). Gene duplication may therefore be followed by both subfunctionalization and "reciprocal neofunctionalization". This model is consistent with studies of young gene duplicates such as the *Adh*-derived genes *jingwei*, *Adh-Finnegan* and *Adh-Twain*. The derived genes all appear to have undergone positive selection for fixation of amino acid changes that result in loss-of-function in *Adh* (Jones and Begun, 2005). In the case, of *jingewi* this has resulted in decreased specificity for 1-propanol compared to the ancestral *Adh* gene but an increased specificity for long-chain alcohols (Zhang *et al.*, 2004). It is important to note however, that studies of other young genes, such as those in the *monkey king* family, have found no evidence for positive selection (using either population genetic or molecular evolutionary approaches) but clear evidence of degenerative mutations (Wang *et al.*, 2004).

Third, quantitative subfunctionalization proposes that the ancestral gene performed a single function and that partial loss-of-function mutations in the two duplicates results in a situation where both copies are necessary to perform the required function at the ancestral level (Figure 1.4). This may be due either to a decrease in gene expression or to some kind of coding-region impairment. Although no cases of quantitative subfunctionalization have been reported in the literature, it would surprising if it did not occur in species with small effective population sizes (Lynch and Force, 2000a). In addition, it may provide an unappreciated link between the redundancy model and "classic" subfunctionalization. The version of neofunctionalization proposed by Ohno (mutation during non-functionality; Ohno, 1970) is often criticized on the basis that it assumes that from the moment of duplication selection is able to distinguish between one duplicate which inherits the ancestral function and the second duplicate which is free to evolve a new function. However, a similar criticism can be leveled at critiques of the redundancy model of duplicate gene preservation (the simple version which assumes identical functions and equal mutation rates to nulls; discussed above). These usually assume that one of the duplicates performs the ancestral function and is under purifying selection, while the second copy derives its value purely from its back-up function. It is then shown that this value is negligible and concluded that the back-up duplicate will be lost (Lynch, 2004). If both duplicates are fixed in a moderately sized population however it is more likely that both will be under a reduced level of purifying selection and, under certain conditions (Lynch and Force, 2000a), both duplicates may decline in function and be preserved by quantitative subfunctionalization. As pointed out previously (Section 1.2.1.2) all duplicate gene pairs created by whole-genome duplication are initially fixed in the population and in the case of an autopolyploidization are expected to initially be fully redundant. Quantitative subfunctionalization may therefore be a more common outcome for whole-genome duplicates than is currently appreciated.

1.2.2.2 Factors affecting subfunctionalization and neofunctionalization

In order for a duplicate gene pair to be permanently retained in a genome, both duplicates must first become fixed at their respective loci and then the pair must become preserved by one of the mechanisms described in the previous section. Like many other aspects of genome evolution (Lynch and Conery, 2003), both fixation and preservation depend intimately on the effective population size of a species (N) in the case of both subfunctionalization and neofunctionalization. As well will see however, they are generally oppositely affected, with the net result that subfunctionalization is an important force in smaller populations, while neofunctionalization dominates in larger ones.

It is often remarked that in small populations selection is inefficient because random genetic drift can result in the loss of favorable alleles (Hartl and Clark, 1997, Li, 1997). The dependence of neofunctionalization on large population sizes goes beyond this however because in small populations mutations to favorable alleles may rarely occur (Lynch, 2004). Even if a gene is duplicated and the duplicate becomes fixed by drift, it is expected that a null allele will arise before a neofunctionalized one at one of the two redundant loci and effectively reverse the process of duplicate fixation by itself becoming fixed by drift. By contrast, in a large population both null alleles and favorable alleles are constantly being introduced into the population by mutation and because large populations behave approximately deterministically (Lynch, 2004), they are expected to persist in the population at a frequency close to their selective coefficients, s. Thus, even though neofunctional alleles are assumed to occur at the expense of the original function (Lynch *et al.*, 2001) and are therefore lethal in the homozygte (the same as null alleles), because they confer an advantage to heterozygotes they are expected to segregate in the population at a frequency s. When gene duplication then occurs one of two series of events may occur. Either a neofunctional allele founds the new locus and it will be swept to fixation, or a wild-type allele founds the new locus and the neofunctional allele will be swept to fixation at the original locus. Crucially, this series of events will only occur when $N > 2/s^2$, effectively restricting neofunctionalization to populations with large effective population sizes (Lynch, 2004).

Subfunctionalization also depends critically on N. Because subfunctionalization occurs in the absence of adaptive mutations, the probability that a new duplicate locus founded by a wild-type allele will be fixed is the probability that the allele will drift neutrally to fixation, $1/(2N)$. Thus, only a tiny fraction of duplicates can even begin to be preserved by reciprocal degenerative mutations. However, this is not the only way in which the effective population size impacts the probability of subfunctionalization. If the product of N and the mutation rate to nulls, μ_n, is much greater than 1 (*i.e.* $N\mu_n >> 1$), then null alleles will arise frequently at the duplicated loci and begin to drift to fixation. This will occur on average in $4N$ generations leaving insufficient time for subfunctionalization to occur. If however $N\mu_n << 1$ then the time for a null allele to arise by mutation at one of the two duplicate loci becomes appreciable and subfunctionalization has a reasonable prospect of success. Indeed, if the both duplicates are fixed and $N\mu_n << 1$, then the probability of subfunctionalization is simply the probability that one duplicate will lose one of its two subfunctions, $2\mu_s/(2\mu_{s\,+}\,\mu_n)$, multiplied by the probability that the other duplicate will then lose the other subfunction, $\mu_s/(2\mu_{s} + \mu_n)$. The combined probability of fixation followed by preservation is then $P_{sub} = \alpha^2/4N$, where $\alpha = 2\mu_s/(2\mu_{s} + \mu_n)$. In addition, it can be shown that P_{sub} may not exceed $1/4N$ for the case where the original gene has two subfunctions and P_{sub} may not exceed $1/2N$ for an arbitrary number of subfunctions or for quantitative subfunctionalization. These calculations make it clear that even with the high rate of creation of new gene duplicates in eukaryotes (Lynch and Conery, 2000), subfunctionalization may only be a significant force in small populations.

Although the picture of duplicate gene preservation painted in the previous two paragraphs is largely accurate, two additional factors can have a non-trivial effect on the probability of duplicate gene preservation: linkage and the mechanism of gene duplication. As was pointed out in Section 1.2.1.1, complete linkage increases the probability of duplicate gene preservation by subfunctionalization but decreases the probability of preservation by neofunctionalization (Lynch *et al.*, 2001). More interesting however, is the effect of the mutational process and the "initial conditions" on the subsequent probability of preservation. For instance, neofunctionalization may become important in small populations if the novel function does not occur at the expense of the ancestral function as usually assumed (Lynch *et al.*, 2001). One circumstance in which this occurs is quantitative neofunctionalization ("Dosage" in Figure 1.4). Because both duplicates can perform the ancestral function but the two together confer an advantage a neofunctionalized allele is effectively always present at the ancestral locus and - even in a small population - the system is effectively poised to proceed towards fixation of the pair once gene duplication occurs. Similarly, Francino (2005) has proposed that if a new duplicate gene confers even a small advantage that it may undergo amplification (perhaps by tandem duplication; Section 1.2.1.1) to increase capacity to perform the novel function. Because this increases the size of the mutational target (effectively increasing *N* at this locus) the duplicate now has an increased probability of sustaining additional neofunctionalizing mutations. A duplicate with a weak selective advantage may thus "bootstrap" its way to having a large selective advantage even in a small population. This theory was proposed originally on the basis of observations made in bacteria (Francino, 2005) but it is notable that *sdic* (Nurminsky *et al.*, 1998), a well-studied young chimeric gene in *Drosophila,* which has been swept to fixation, exists as a ten gene tandem array.

The role of mutation in facilitating subfunctionalization is no less important. For example, if a duplicate gene is created without one of its two tissue-specific enhancers (perhaps because the duplication does not span the entire promoter; Katju and Lynch, 2003), then the first subfunctionalization step has already occurred. In this case P_{sub} can exceed the asymptotic limit of $1/4N$ noted above. Similarly, if following a segmental duplication the whole duplicated segment is swept to fixation because one of the duplicated genes confers a dosage advantage, then P_{sub} for all of the other genes in the segment may significantly exceed $1/4N$ (assuming they each have two subfunctions). Indeed, since in this case the probability of fixation is effectively 1 rather than $1/(2N)$ for each of the genes in the duplicated segment, the upper limit on P_{sub} falls from $1/4N$ to 1/2. This is similar to the situation that arises following whole-genome duplication. Because of the fixation-at-birth phenomenon (Section 1.2.1.2), the probability of duplicate gene preservation after WGD is effectively independent of N (provided $N\mu_n << 1$; discussed above) and depends only on the parameter α. This in turn depends only on the ratio of subfunctionalizing to non-functionalizing mutations (μ_s/μ_n) and it can be shown that if $\mu_s/\mu_n = 0.5$ and all genes have two subfunctions, then the frequency of subfunctionalization is expected to be 1/8. Similarly, if $\mu_s/\mu_n = 0.1$ the frequency of subfunctionalization is expected to be 1/72. However, if genes have more than two subfunctions, then the rate of preservation will be even higher after whole-genome duplication. This may partly explain the high rate of duplicate gene preservation after whole-genome duplication (Lynch, 2004) and of course, once genes have been preserved in duplicate (by any mechanism) they become platforms for secondary adaptations. Thus, neutral processes (Lynch *et al.*, 2005) and subfunctionalization in particular (Force *et al.*, 2005) may be key steps in the generation of evolutionary novelty.

1.2.2.3 Duplicate gene preservation in yeast

The theoretical considerations in the previous section suggest two questions. First, what is the effective population size of yeast? Based on levels of silent site diversity in five genes sequenced in 80 strains of *S. cerevisiae* it appears that the effective population size of yeast may be considerably smaller than previously anticipated (Fay and Benavides, 2005b). Indeed, on the basis of larger thirty gene survey it has been suggested that N may be in the range 10^5 - 10^7 (Barry Williams, pers. comm.). Neutral processes such as subfunctionalization are likely to be important towards the lower end of this scale. Second, how many genetically separable subfunctions do yeast genes have on average? This question has been addressed by comparing the growth rates of single-gene deletion strains to that of wild-type strains in multiple environmental conditions (Dudley *et al.*, 2005, Ericson *et al.*, 2006). Studies are limited by the number of growth conditions they consider but Ericson *et al.* (2006) concluded that at least 17% of genes are required in more than two conditions and that 4-5% are hyper-pleiotropic. This latter group is presumably enriched for house-keeping genes, suggesting that no less than 12% (17%-5%) of genes in *S. cerevisiae* have multiple subfunctions. Consistent with the notion that genes with more subfunctions are more likely to be retained in duplicate, it has been noted that longer genes and genes with more protein domains are more likely to have a paralog in *S. cerevisaie* (He and Zhang, 2005a). In addition, it should be noted that even genes that have only single recognizable function may be retained in duplicate by quantitative subfunctionalization.

Because there are relatively few young gene duplicates (defined as dS < 0.02 in Moore and Purugganan, 2003) in yeast, it has not been possible to verify the predictions of theory by studying young duplicate genes as it has been *Drosophila*. Instead, large-scale studies of older gene duplicates, such as those retained in duplicate in yeast since the whole-genome duplication, have been attempted (Kellis *et al.*, 2004). It has been argued that duplicate gene pairs like *SIR3/ORC1* that display highly asymmetric protein sequence evolution must have been preserved by neofunctionalization (the “slow” copy is assumed to perform an ancestral function while the “fast” copy optimizes a novel function) whereas pairs that exhibit equal rates of protein sequence evolution are likely to have been preserved by other mechanisms (Kellis *et al.*, 2004). This is unlikely to be reliable however because neither neofunctionalization nor subfunctionalization make unambiguous predictions about the symmetry of protein sequence evolution after gene duplication. In the latter case, there is no reason why one member of a duplicate should not retain four of five ancestral sub-functions and in the former case, quantitative neofunctionalization may well lead to equal rates of protein sequence evolution.

The hypothesis that duplicate gene pairs that display unequal rates of protein sequence evolution are candidates for neofunctionalization has been directly tested using complementation tests as described in Figure 1.4. Van Hoof (2005) showed that deletions of four pairs of *S. cerevisiae* whole-genome duplicates that had been considered to be likely candidates for neofunctionalization (including *SIR3/ORC1*) could be rescued by expression their single-copy *S. kluyveri* orthologs. This strongly suggests that in the case of these four pairs of duplicates, neither duplicate performs a function not possessed by the ancestral single-copy gene, and that non-adaptive mechanisms may be more important for the preservation of whole-genome duplicates in yeast than commonly recognized. Nevertheless, there are some convincing examples of neofunctionalization in yeast. For instance, Thomson *et al.* (2005) have shown by reconstructing the ancestral sequence of the *ADH1/2* duplicate gene pair and assaying its enzymatic ability *in vitro* that the sequence that existed prior to the duplication was capable of performing only the function currently associated with *ADH1*. This is strong evidence that *ADH2* has acquired a novel function and it seems likely that this is also the reason that the duplication was preserved.

1.2.3 Gene loss

Although the vast majority of new duplicate gene pairs are resolved by loss of one or other gene copy (Lynch, 2004), gene loss has not been well studied except in the context of either reductive genome evolution of endosymbionts (Douglas *et al.*, 2001, Gilson *et al.*, 2006) or birth-and-death models of gene family evolution (Hahn *et al.*, 2005). In contrast to the work presented in Chapter 2 and Chapter 3 of this thesis however, in neither of these cases has the goal been to understand the process of gene loss itself. In this section I review: the small number of well understood examples of gene loss, loss of members of duplicate gene pairs, and evidence suggesting that where gene loss involves a choice between two members of a duplicate gene pair, that the choice is not arbitrary.

1.2.3.1 Circumstances under which gene loss may occur

Assuming that a gene pair is fixed in a population, gene loss may occur in three circumstances; the selection pressure that caused the gene to be maintained no longer exists; another gene is present that can complement the loss of the original gene (Morett *et al.*, 2003); or a new selection pressure emerges that causes the gene to be maladaptive (Olson, 1999). These correspond respectively to the three cases where gene loss is due to a loss of purifying selection, selectively neutral (but without loss of purifying selection), or favored by positive selection. The loss of seven genes in the *GAL* pathway (which function to sense, import and metabolize the sugar galactose) from the genome of *S. kudriavzevii* (Figure 1.1) has been proposed as an example of the first of these (Hittinger *et al.*, 2004). Although it is hard to exclude the possibility that loss of the *GAL* genes was beneficial in some way, the fact that they have been lost independently in several yeast lineages that occupy very different ecological niches argues against the possibility that the *GAL* genes were maladaptive in the specific environment (rotting leaves; Hittinger *et al.*, 2004) preferred by *S. kudriavzevii*.

By contrast, the loss of the *BNA* pathway in *C. glabrata* is likely to have occurred under strong selection because it plays an important role in virulence (Domergue *et al.*, 2005). The *BNA* pathway is responsible for the synthesis of nicotinic acid and allows *S. cerevisiae* and other yeasts to replenish their pool of NAD^+ if it is depleted by the transcriptional repressor *SIR2*. By contrast *C. glabrata* is entirely dependent on external sources of nicotinic acid and when it is unavailable genes, such as the adhesin (*EPA*) genes, which are usually repressed by *SIR2* become expressed. Notably, the human urinary tract is very low in nicotinic acid (Domergue *et al.*, 2005).

Finally, the loss of the a2 gene from the ancestral *MAT* locus (*MTL* in Figure 1.3B) in hemiascomycete yeasts appears to be an example of loss due to redundancy (Tsong *et al.*, 2003, Tsong *et al.*, 2006). In *C. albicans* a2 is required to activate a-specific genes in **a** cells, but in *S. cerevisiae* these genes are expressed by default in **a** cells and are instead repressed by α2 in α cells. By examining how a-specific genes are regulated in yeasts that diverged from the *S. cerevisiae* lineage after it diverged form *C. albicans*, they reconstructed the evolutionary steps that took the a-specific genes from positive control in *C. albicans* to negative control in *S. cerevisiae* and showed that an intermediate stage is likely to have involved redundant control by both systems. Thus, loss of the a2 gene was possible because although there was strong purifying selection for appropriate expression of a-specific genes, compensatory changes arose that could complement the loss. As in the case of the *GAL* pathway, it is hard to exclude the possibility that the change was favored by selection for some unknown reason, but these three examples serve to illustrate the possible conditions under which gene loss may occur.

1.2.3.2 Loss of members of duplicate genes pairs after polyploidization

Although it is known that increased gene dosage can be pathogenic in yeast (Sopko *et al.*, 2006) and in humans (especially neurodegenerative diseases; Lupski and Stankiewicz, 2005, Rovelet-Lecrux *et al.*, 2006, Padiath *et al.*, 2006) it seems likely that most gene loss after gene duplication is neutral and due to the presence of a redundant paralog. Because every gene in the genome is duplicated simultaneously by whole-genome duplication, it is expected that all dosage relationships will be preserved (Veitia, 2005) and thus that the stoichiometry of complexes will not be adversely affected (but see Storchova *et al.*, 2006). Similarly, deleterious duplications created by other mechanisms (Section 1.2.1.1) are unlikely to every be fixed, so there is no need to invoke selection for restoration of the ancestral state. Instead, as is proposed in Chapter 2, it is likely that most gene loss after polyploidization is due to passive inactivation and gene deletion.

Because after whole-genome duplication every chromosome is duplicated, an efficient way to restore diploidy would be to lose whole chromosomes. This is observed in both synthetic plant and synthetic yeast polyploids (references in Comai, 2005). Surprisingly, however this does not appear to have occurred after the yeast whole genome duplication. Gene order comparisons between *S. cerevisiae* and yeast species that diverged prior to the whole-genome duplication show that pairs of sister regions exist in *S. cerevisae* for almost the entire pre-duplication genome (Kellis *et al.*, 2004, Byrne and Wolfe, 2005). This suggests that no large chromosomal segments were lost. In addition, analysis of patterns of gene loss indicates that the median size of deleted segment was likely to have been just one gene long (Kellis *et al.*, 2004, Byrnes *et al.*, 2006). This is consistent with the "interleaving" pattern in Figure 1.2 and with suggestions that gene loss in yeast proceeds by inactivation of the open reading frame and then "erosion" by multiple small deletions (Fischer *et al.*, 2001, Hittinger *et al.*, 2004).

One scenario that can explain the observed pattern of gene loss after whole-genome is as follows. The presence of some genes in duplicate (such as those coding for ribosomal proteins; Section 1.2.1.2) was initially beneficial and thus loss of whole chromosomes was selected against. Gene loss then proceeded by a series of smaller deletions with a small number of genes being lost from each chromosome. If these losses included at least one essential gene from each chromosome however, it would no longer be possible to lose any chromosome in its entirety. Thus, even after any temporary selective advantages conferred by dosage at some loci have subsided all gene losses would have to occur by smaller deletions.

1.2.3.3 Which member of a duplicate gene pair gets lost?

Conant and Wagner have remarked: Much like humans, gene duplicates may be created equal, but they do not stay that way for long. (Conant and Wagner, 2003). Given that this is the case and that one member of a duplicate gene pair will be lost, the question arises, “which member of the pair should be discarded?” It is possible that although the loss of one member of a duplicate pair is effectively neutral because only one gene is required to perform the particular function, that the “choice” of which member of the pair to lose is not. This process can only be studied by comparing how ancestrally duplicated gene pairs have been resolved in different lineages. Specifically, if both members of a duplicate gene pair are equally capable of performing the required function, we should expect that on average 50% of lineages should retain each copy and lose the other. However, if the duplicates have diverged in function then one copy may be favored over the other and the number of lineages retaining a particular copy may deviate from random expectations.

In Chapter 2 and Chapter 3 I describe the only comprehensive studies of duplicate gene loss so far, but anecdotal reports of two lineages independently losing the same (orthologous; Neafsey and Hartl, 2005) or alternative (paralogous; Fischer *et al.*, 2001) copies of ancestrally duplicated genes do exist. For instance, while comparing the genomic locations of homologous genes between *S. cerevisiae* and *S. bayanus* Fischer *et al.* noticed that an apparent single gene transposition event was actually due to an ancient duplication that was resolved differently in the two lineages (Fischer *et al.*, 2001). Conversely, Neafsey and Hartl showed by comparing the genomes of *Tetraodon*, fugu and medaka that *Tetraodon* and fugu had independently lost *RH2-2*, a functionally diverged “green” opsin (RH2-1 detects light of a different frequency; Neafsey and Hartl, 2005). Interestingly, because fugu lost *RH2-2* relatively recently they were able to test - but not support - the hypothesis that the loss was driven by natural selection. Thus, two fish lineages independently dispensed with an apparently unnecessary duplicate gene and retained the functionally useful paralog.

Selection is not the only force that can result in loss of the same (orthologous) gene copy in two independent lineages more often than expected by chance. The same observation may arise by two other processes. First, if a pair of duplicates are fully redundant then at any given time null alleles are expected to be segregating at both loci at a moderate frequency. As pointed out in Section 1.2.2.1 however, at some point a null allele will drift to high frequency at one of the loci and become fixed. If just prior to this event the lineage diverges then both daughter lineages are almost certain to fix a null allele at the same locus and hence convergently lose the same duplicate. Indeed, it can be shown that on average two lineages that inherit identical duplicates genes will lose the same copy 60% of the time (rather than the expected 50%) due to this process alone (D.S. and Mike Lynch, unpublished data). Second, mutation pressure can lead to the preferential loss of one or other duplicate. This occurs in the case of transfers of genes from organelles to nuclear genomes. Because genes are constantly being duplicated from the organellar genome to the nuclear genome but not in the opposite direction, the nuclear copies will eventually be fixed and the organellar copies will be lost in multiple independent lineages even if the nuclear copy confers no advantage. For instance, the mitochondrial ribosomal protein *rps10* has been transferred to the nuclear genome at least 26 separate times (Adams *et al.*, 2000), consistent with the idea that both mitochondrial and nuclear duplicates frequently coexisted, but that eventually the nuclear copy became fixed and the mitochondrial copy was lost by drift.

1.3 Speciation

1.3.1 <u>Species barriers in yeast</u>

Saccharomyces sensu stricto yeast species (Section 1.1.1.2; Figure 1.1) are generally accepted to be distinct on the basis of low viability of spores produced by hybridization. Whereas mating between members of the same *S. cerevisiae* strain produces spores with viabilities of close to 100% (Greig *et al.*, 2002a) and spores produced by mating between *S. cerevisiae* strains often show viabilites of ~80% (Greig *et al.*, 2002a), mating between *S. cerevisiae* and *S. paradoxus* or other *Saccharomyces* species typically result in <1% viable of spores (references in Greig *et al.*, 2002b). The bases of the reproductive barriers among *Saccharomyces* sensu stricto yeasts have been investigated intensely over the last few years. In contrast to animal and plant studies, which have tended to focus either on identifying Dobzhansky-Muller incompatibilities between protein-coding genes or on chromosomal rearrangements respectively (Coyne and Orr, 2004), a variety of mechanisms have been considered and excluded (Liti *et al.*, 2006). Three are reviewed briefly here, chromosomal rearrangements, Dobzhansky-Muller incompatibilities and sequence divergence acted on by the mismatch repair system.

1.3.1.1 Chromosomal Speciation

Chromosomal rearrangements are hypothesized to lead to hybrid inviability by inducing the formation of multivalents at meiosis. Multivalents are prone to mis-segrgation and may result in the production of aneuploid spores with decreased fitness. This may be due either to spores being deficient for essential genes or due to the increased likelihood of mis-segregation in future meioses (a zygote produced by mating involving a +1 aneuploid is expected to be triploid and unstable). Both retrospective and interventionist approaches have been employed to estimate the contribution of chromosomal rearrangements to hybrid viability between *S. cerevisiae* and other sensu stricto yeasts.

Fischer *at al.* used a combination of electrophoresis and PCR to identify karyotype changes in sensu stricto yeasts relative to *S. cerevisiae* (Fischer *et al.*, 2000). They detected no rearrangements in *S. paradoxus* or *S. kudriavzevii* relative to *S. cerevisiae* but four in *S. cariocanus* and *S. bayanus* and two in *S. mikatae*. These observations are inconsistent with the known levels of spore viability among these species. For instance, if each rearrangement reduces spore viability by 50% then the expected viability of viable spores in a cross between *S. cariocanus* and *S. paradoxus* is 6.25% but the observed viability is only one tenth of this. Additional factors must therefore contribute and the authors concluded that chromosomal rearrangements were not a prerequisite for speciation.

Nevertheless, the possibility remained that rearrangements contribute quantitatively to reproductive isolation or that they may reinforce species barriers after they have arisen by another mechanism. To address this question Delneri *et al.* used the *Cre-lox* inducible recombination system to engineer strains of *S. cerevisiae* that are collinear to one of two strains of *S. mikatae* (Delneri *et al.*, 2003). One of these differs from wild-type *S. cerevisiae* (but not the engineered strain Sct1) by a single rearrangement and the other differs from wild-type *S. cerevisiae* (but not the engineered strain Sct1/2) by two rearrangements. In subsequent crosses between these strains and wild-type *S. cerevisiae* spore viabilities of 60% and 25% were obtained with Sct1 and Sct1/2 respectively. These percentages are close to what is expected under the assumption of 50% loss of viability per rearrangement noted above and suggests that mis-segregation contributes to spore death. In addition, inter-specific crosses between Sct1 and the *S. mikatae* strain with which it is collinear, resulted in 20-30% spore viability in 4 of 10 crosses. These data clearly support the view that chromosomal rearrangements at least have the potential to contribute to species barriers in yeast, however the failure to restore full viability indicates that other mechanisms must also be invoked. Indeed, it was noticed that all of the viable spores were aneuploid with some having up to 25 chromosomes. It is therefore possible that these extra chromosomes are masking recessive Dobzhansky-Muller incompatibilities (discussed below) that might otherwise reduce viability.

1.3.1.2 Dominant and recessive Dobzhansky-Muller incompatibilities

An alternative to the chromosomal basis for hybrid infertility is the existence of Dobzhansky-Muller incompatibilities between epistatically interacting genes. This model posits that after an ancestral lineage diverges to create two daughter lineages incompatible changes arise in alternative members of a pair of interacting loci. Thus, in one lineage one of the genes diverges from the ancestral sequence and in the second lineage the other gene diverges from the ancestral sequence. These changes are neutral (or possibly beneficial) provided the ancestral sequence is present at the alternative locus, but if the diverged versions of both genes are brought together in a hybrid they will interact in such a way as to reduce fitness. The mechanism by which fitness is reduced is not specified. It is important to note that the incompatibility can be either dominant or recessive. In the former case, the presence of the two diverged genes will reduce fitness irrespective of what other genes are present. In the latter case however, the existence of an incompatibility can be masked by the presence of an ancestral type sequence at both loci (*e.g.* in an F_1 hybrid) – as in the case of the original daughter lineages, the presence of an ancestral type gene at one locus and a diverged gene at the other is sufficient to supply the required function and the presence of any additional sequences (ancestral or diverged) is irrelevant.

In order to test the possibility that dominant Dobzhansky-Muller incompatibilities might play a role in reproductive isolation between sensu stricto yeast lineages Greig *et al.* (2002a) repeated the test originally performed by Dobzhansky in *Drosophila* (Dobzhansky, 1933). Dobzhansky had observed that in infertile *D. pseudoobscura* hybrids, homologous chromosomes failed to pair at meiosis, thus arresting spermatogenesis. In order to distinguish between the possibility that the chromosomes could not pair because they were too diverged and the possibility that genetic incompatibilities between the two parental species had prevented successful meiosis, Dobzhansky examined the pairing of tetraploid spermatocytes. Because tetraploidy is achieved by duplication of the homologous chromosomes that are present in diploids, failure to pair cannot be due to the lack of an homologous partner. When Dobzhansky performed this test using tetraploid spermatocytes he observed that the hybrids were still infertile and concluded that sterility was due to genetic factors. Strikingly, when repeated using sensu stricto yeast species, precisely the opposite result was obtained (Greig *et al.*, 2002a).

Greig *et al.* first created pseudo-haploids of several yeast species by deleting a single copy of the *MAT* locus from non-hybrid diploids (Greig *et al.*, 2002a). They then performed inter-specific crosses between *S. cerevisiae* pseudo-haploids and pseudo-haploids from the other sensu stricto species. In each case, the spore viability of the hybrid was ~90% compared to <1% for true hybrid diploids. Indeed, the spore viability of the hybrids obtained by crossing pseudo-haploids was not significantly different form that obtained in intra-specific crosses of normal haploids. These data indicate comprehensively that hybrid infertility in yeast is not due to dominant Dobzhansky-Muller incompatibilities between species. If dominant interactions between loci were responsible for infertility, increasing the number of copies of each gene present would not be able to rescue the infertile phenotype.

The evidence regarding recessive Dobzhansky-Muller incompatibilities is not so clear, although it appears they also do not have a role to play in yeast speciation. This conclusion is suggested by the fact that Chambers *et al.* were able to replace *S. cerevisiae* chromosome III with *S. paradoxus* chromosome III without any loss of viability in the haploid (Chambers *et al.*, 1996). This indicates that although they are ~15% diverged at the DNA level and ~10% diverged at the protein sequence level (Cliften *et al.*, 2001) that all the functional elements on chromosome III are conserved between these two species. Moreover, because the *S. paradoxus* chromosome III is present in an otherwise completely *S. cerevisiae* background, no recessive Dobzhansky-Muller incompatibilities can exist between loci on *S. paradoxus* chromosome III and other loci in the genome. In Liti *et al.* (Liti *et al.*, 2006) it is reported (without evidence) that all chromosomes in *S. cerevisiae* can be replaced by their *S. paradoxus* homologs. If this is true it strongly suggests that Dobzhansky-Muller incompatibilities little part to play in yeast speciation. Moreover, because *S. paradoxus* and *S. cerevisiae* are collinear, it suggests that sequence divergence acted on by the mismatch repair system is the primary mechanism of speciation in yeast (Section 1.3.1.3).

Although the evidence cited above suggests that recessive Dobzhansky-Muller incompatibilities do not play a significant role in yeast species barriers, indirect evidence supporting their existence has been reported based on inter-specific crosses. Whereas dominant epistatic interactions can be revealed by crossing haploids from two parental species and examining the fertility of the F_1 generation, recessive incompatibilities can only be revealed by examining F_2 or successive generations in which regions of the genome may be homozygous at the locus of interest. In order to investigate the fertility of an F_2 generation, Greig *et al.* exploited the fact that most hybrid diploids are fertile at a low level (typically <1%) and collected 80 gametes from a large cross (Greig *et al.*, 2002b). They then allowed these to auto-diploidize to obtain a homozygous F_2 generation. Interestingly, the F_2 hybrids fulfilled the main two requirements for a new species: High fertility (~80%) and isolation from the ancestral population (back-cross hybrid fertility ~7%). Nevertheless, the reason for the decrease in fertility relative to the pure parental strain (~20%) is unclear. As the authors point out, chromosomal incompatibilities cannot explain the difference, since the F_2 hybrids were produced by auto-diploidization and must therefore be able to pair at meiosis. Similarly, dominant Dobzhansky-Muller incompatibilities cannot be responsible since there is no evidence that they occur in yeast (Greig *et al.*, 2002a). In addition, the authors argue that aneuploidy is not the explanation although they show - as was previously observed for the hybrids obtained by crossing *S. mikatae* to artificially collinear *S. cerevisiae* strains (Delneri *et al.*, 2003) – that the F_2 hybrids are highly aneuploid. By this process of exclusion the authors conclude that the decreased fertility must be attributable to recessive Dobzhansky-Muller incompatibilities, however given the results of the chromosome complementation experiments cited above (Chambers *et al.*, 1996), direct evidence for a role in reproductive isolation will be required to establish their relevance.

Although evidence for a contribution to reproductive isolation between species is equivocal, it should be noted that abundant epistasis has been detected in genome-wide scans for expression QTLs (Brem *et al.*, 2005) and that negative fitness consequences have been demonstrated for certain pairs of alleles from different *S. cerevisiae* strains (Heck *et al.*, 2006). For instance, haploids with a *MLH1* allele from S288C (*cMLH1*) and a *PMS1* allele from SK1 (*kPMS1*) were shown to accumulate mutations at approximately 100 times the rate of any other combination of alleles (*cMLH1-cPMS1*; *kMLH1-kPMS1*; *kMLH1-cPMS1*). This defect was observed in both genetic backgrounds and shown to result in a significant reduction in the number of complete tetrads over the course of ~100 generations, consistent with a fitness cost (Heck *et al.*, 2006). Thus, although the *cMLH1-kPMS1* interaction results neither in inviability nor sterility of spores produced by crossing

S288C and SK1, it indicates that incompatibilities exist between genotypes of different strains and that other more severe incompatibilities may also be segregating.

1.3.1.3 Sequence divergence acted on by the mismatch repair system

In contrast to both the chromosomal and genic (Dobzhansky-Muller incompatibility) models of speciation there is unambiguous evidence that sequence differences between homologous chromosomes can interfere with recombination and lead to nonproductive meioses between diverged yeast species (Hunter *et al.*, 1996). Moreover, there is evidence that this interference is mediated by the mismatch repair system and that it results in spore inviability by two separate mechanisms, meiosis I non-disjunction (Hunter *et al.*, 1996) and mismatch stimulated chromosome loss (Chambers *et al.*, 1996). Both of these result in potentially lethal aneuploidy. Indeed, the most attractive aspect of this model is that it predicts the existence of the widespread aneuploidy that has arisen during (and confounded) attempts to study other possible mechanisms of speciation.

In order to test the hypothesis that sequence divergence detected by the mismatch repair system can lead to aberrant meioses, Hunter *et al.* crossed strains of *S. cerevisiae* and *S. paradoxus* and then measured the rates of both recombination and aneuploidy in the resulting gametes. This was performed using wild-type, *pms1* null, and *msh2* null strains of *S. cerevisiae* and comparisons between crosses performed using the wild-type and mutant strains showed that recombination, non-disjunction and viability changed in concert. For instance, both the spore viability and the rate of recombination seen when wild-type *S. cerevisae* was crossed to wild-type *S. paradoxus* was approximately 1% of that seen in intra-specific crosses. By contrast, when *msh2* null *S. cerevisae* was crossed to wild-type *S. paradoxus* both recombination and viability rose to ~10%. In addition, non-disjunction was significantly lower when an *msh2* null strain was crossed to *S. paradoxus* than when a wild-type strain was used. These data support the view that when diverged sequences pair at meiosis but fail to recombine (due to the mismatch repair system) that non-disjunction may occur and lead to inviable aneuploid spores. Subsequent work by (Chambers *et al.*, 1996) clarified the mechanism by which this occurs. They showed that ascii that contain two viable spores tend to be disomic, consistent with meiosis I non-disjunction but that ascii with three viable spores typically contain no disomes and one recombinant spore. This authors argue that the unpaired recombinant phenotype arises because although the sequences of *S. cerevisiae* and *S. paradoxus* are similar enough that one successful strand invasion may occur, the probability of the reciprocal strand invasion occurring is negligible. Hence, one recombinant chromosome is formed and the other aborted.

Is sequence divergence acted on by the mismatch repair system sufficient to account for reproductive isolation among sensu stricto yeasts species? Two lines of evidence suggest that it may be. First, Grieg *et al.* used the same assays described above to assess the impact of between strain sequence differences on reproductive isolation in *S. cerevisiae* and *S. paradoxus* (Greig *et al.*, 2003) and found in both cases that it could account for at least 50% of the variation: Spore viability and recombination were both increased in a *msh2* null background. Second, Liti *et al.* have shown that once chromosomal rearrangements are taken into account there is a monotonic relationship between sequence divergence and spore viability (Liti *et al.*, 2006). This is consistent with a causal relationship and in the absence of any significant evidence that genic incompatibilities play a role in sensu stricto yeast species barriers, suggests sequence divergence may be a sufficient explanation.

1.3.2 Dobzhansky-Muller incompatibility

In his 1942 book, *Systematics and the Origin of Species*, Ernst Mayr proposed that species should be defined by the “Biological Species Concept” (BSC): species are groups of actually or potentially interbreeding natural populations that are reproductively isolated from other such groups (Mayr, 1942). Although, as is clear from the preceding section, this may not be a useful definition in all contexts, it has spurred intense research and has led to some significant successes (discussed below; Coyne and Orr, 2004). Most of this research has centered on the Dobzhansky-Muller model described in Section 1.3.1.2 and the search for alleles of pairs of protein-coding genes that interact in such a way as to lower fitness. Nevertheless, in Chapter 2 I argue that neither sequence divergence acted on by the mismatch repair system nor the classic Dobzhansky-Muller model can explain the emergence of reproductive isolation among the yeast lineages that emerged after the yeast WGD (Section 1.1.2.3). Instead, I propose that a modified version of the Dobzhansky-Muller model in which epistatically interacting null alleles lead to reduced fitness is responsible.

In this section, I review work on the classic version of the Dobzhansky-Muller model and highlight some of its successes before describing the modified version proposed by Werth and Windham (1991) and refined by Lynch and Force (2000b).

1.3.2.1 Classic Dobzhansky-Muller incompatibility

The classic Dobzhansky Muller model, as described in Section 1.3.1.2, posits that a pair of genes that act together to perform a required function in an ancestral species diverge in a pair of daughter species. Although the diverged genes are still capable of supplying the required functions in the daughter lineages (the interacting genes have diverged together), the diverged gene from one daughter lineage may not be able to function in conjunction with the diverged gene from the second lineage. Thus, if the daughter species are crossed to create a hybrid its offspring may be inviable ("Classic Dobzhansky-Muller" in Figure 1.5).

In spite of the popularity of the Dobzhansky Muller model only a handful of "speciation genes" have been identified and in only a single case have both members of a pair of epistatically interacting loci been identified (Wu and Ting, 2004). The fish *Xiphophorus maculatus* and *Xiphophorus helleri* both possess a gene called *Xmrk-1*, which is a ubiquitously expressed epidermal growth factor receptor (Wittbrodt *et al.*, 1989). In *X. maculatus Xmrk-1* has been duplicated by non-homologous recombination to produce a second gene, *Xmrk-2,* which has inherited a promoter from a neighboring locus *D* (reviewed in Wu and Ting, 2004). *Xmrk-2* is therefore regulated by the same genes that regulate *D*, amongst which is a repressor called *R*. In addition, *Xmrk-2* has diverged at the protein sequence level from *Xmrk-1*. It possesses two amino acid substitutions, which cause it to function constitutively in the absence of ligand binding. Since, as noted above, *Xmrk-1* and *Xmrk-2* are growth factor receptors, mis-expression of *Xmrk-2* causes it to behave essentially as a dominant oncogene and results in the formation of malignant cancers. Nevertheless, because the repressor *R* is present in *X. maculatus* this does not occur and the potentially lethal mutant does not confer any fitness cost. Similarly, *X. helleri* suffers no fitness penalty because although it does not possess the repressor *R*, neither does it possess *Xmrk-2*. This system breaks down however in F_2 hybrids and in back-crosses because it is possible to obtain genotypes that are homozygous null for the repressor R (*X. helleri* background) and also carrying a copy of *Xmrk-2* from *X. maculatus*. This is a lethal genotype and comprises a reproductive barrier between these species.

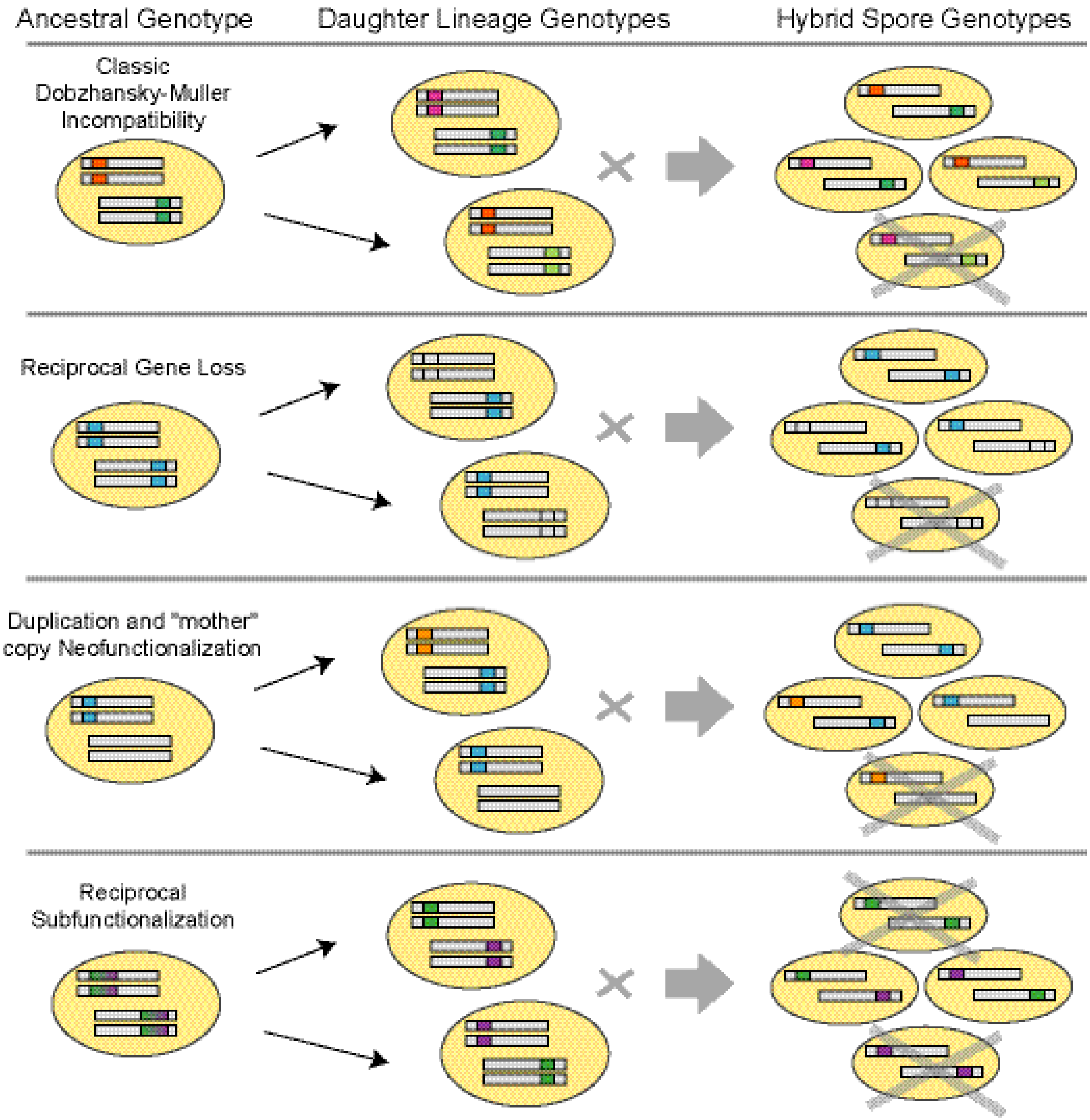

Ancestral Genotype
Daughter Lineage Genotypes
Hybrid Spore Genotypes
Classic Dobzhansky-Muller Incompatibility
Reciprocal Gene Loss
Duplication and "mother" copy Neofunctionalization
Reciprocal Subfunctionalization

Figure 1.5 Models of hybrid incompatibility based on the Dobzhansky-Muller incompatibility model. Ovals containing paired chromosomes represent diploids and ovals containing unpaired chromosomes represent haploids. Boxes represent genes and colors represent functions except grey: grey boxes represent loci where genes formerly existed. All the ancestral functions are required at each stage for viability. Novel functions may increase fitness (not represented) but are not required for viability. In the "Classic Dobzhansky-Muller Incompatibility" model genes may diverge but retain the ancestral function *e.g.* the pink box represents a function derived from the red box. Large grey 'X' marks indicate inviable spores.

Interestingly, this system differs in several ways from the classic Dobzhansky-Muller paradigm. First, it is not clear that the two epistatically interacting loci, *Xmrk-2* and the repressor *R*, were present in the ancestral species. Moreover, it is clear that they need not have been. For instance, the following scenario is compatible with the data provided above. *R* and *Xmrk-1* were both present in the *Xiphophorus* ancestor. Subsequently, *X. maculatus* duplicated *Xmrk-1* to produce *Xmrk-2* and because the repressor *R* was present it drifted neutrally to fixation. It has also sustained two substitutions that would be deleterious were the gene to be expressed. Since it is not however, they have been able to segregate in the population without consequence. In *X. helleri* none of these events occurred, but the *R* gene was lost for some unknown reason, perhaps because the gene that it usually regulates, *D*, was also lost. The second point to take from this therefore is that the Dobzhansky-Muller incompatibility (between *Xmrk-2* and *R*) may have arisen by gene loss rather than by divergence, since the negative interaction is between *Xmrk-2* and the R^- (null) genotype. Finally, in this version of events there is no requirement for positive selection or adaptation to a new environment to drag a “speciation gene” to fixation. Reproductive isolation may therefore arise neutrally under a Dobzhansky-Muller model.

1.3.2.2 The hunt for speciation genes

Much work has been done in the *Drosophila* community to identify genes responsible for post-zygotic reproductive isloation (Noor and Feder, 2006). Most of this has focused on the search for hybrid inviability genes (as opposed to hybrid sterility genes) and the vast majority has done so within a Dobzhansky-Muller framework. Perhaps the most impressive study undertaken so far is a deletion mapping study by Presgraves (2003). In order to identify pairs of genes responsible for recessive Dobzhansky-Muller interactions that cause hybrid inviability between *D. melanogaster* and *D. simulans*, Presgraves created hybrids that were hemizygous for a particular region of the *D. simulans* genome and carried a *D. melanogaster* X chromosome. If a gene in the single-copy region of the *D. simulans* genome was incompatible with a gene some-where on the single *D. melanogaster* X chromosome, then fewer offspring should be observed than when non-hemizygous hybrids were created (*i.e.* they have a *D. melanogaster* chromosome without a deletion which can mask the incompatibility). By scanning the entire *D. simulans* genome Presgraves identified 40 regions that resulted in a lethal phenotype in hybrids bearing a single *D. melanogaster* X chromosome. In total, it was estimated that approximately 200 recessive Dobzhansky-Muller incompatibilities separated the *D. simulans* and *D. melanogaster* (Presgraves, 2003).

In order to verify that these deficiencies represented true Dobzhansky-Muller incompatibilities, Presgraves verified three requirements of the model. First, lethality should only occur in the hybrid. This was confirmed by making pure *D. simulans* flies that were hemizygous in the region of interest. None exhibited any evidence of haploinsufficieny, indicating that the lethality is hybrid-specific. Second, most incompatibilities are thought to be recessive, in line with Haldane's rule. This was verified by examining the viability of hybrid females hemizygous for the same region – less than 25% showed any phenotype and most of these were weak. Finally, Presgraves demonstrated that the lethality was due to true epistatic interactions by replacing the single *D. melanogaster* X chromosome in the hybrid males with a *D. simulans* X chromosome. As predicted, no inviability was observed. These data strongly suggest that many recessive, epistatic, hybrid-specific incompatibilities exist between *D. simulans* and *D. melanogaster* and, consistent with the classic Dobzhansky-Muller model, subsequent fine-mapping and complementation tests in one of these regions showed that the *D. simulans Nup86* gene is incompatible with a locus on the *D. melanogaster* X chromosome (Presgraves *et al.*, 2003). The interacting locus on the X chromosome has yet to be identified.

Does this mean that there are 200 pairs of incompatible genes that can result in lethality in *D. simulans* / *D. melanogaster* hybrids and that there are hundreds of speciation genes waiting to be found? This is still uncertain. For instance, Orr and co-workers have recently reported that a locus on *D. simulans* chromosome three and a locus on *D. melanogaster* chromosome four can also result in inviability if both are homozygous (Masly *et al.*, 2006). In contrast to *Nup86* however, they found that neither locus encodes a gene. Instead, they found that hybrids with this genotype are sterile because neither chromosome possesses a copy of a gene called *JYAlpha* (which is located on *D. simulans* chromosome four and *D. melanogaster* chromosome three). Indeed, the data seem to suggest that the gene was present in duplicate in the common ancestor of *D. simulans* and *D. melanogaster* but subsequently underwent reciprocal loss in the two daughter lineages ("Reciprocal Loss" in Figure 1.5). *D. simulans* lost the copy on chromosome three and *D. melanogaster* lost the copy on chromosome four. This is of interest because this pair of loci fulfill the three criteria used by Presgraves to very the results of his genome-wide scan for Dobzhansky-Muller incompatibilities (Presgraves, 2003). First, the incompatibility occurs only in the hybrid because all *D. simulans* files have a copy of *JYAlpha* on chromosome four and thus intra-specific crosses cannot result in null genotypes. Second, it is recessive because one copy of *JYAlpha* on any chromosome is sufficient for fertility. Third, it is epistatic because the both the *D. melanogaster* null allele on chromosome four and the *D. simulans* null allele on chromosome three must be present together to induce lethality. This raises the possibility that many of the Dobzhansky-Muller incompatibilities identified by Presgraves (2003) are not "Classical" Dobzhansky-Muller incompatibilities but instances of reciprocal gene loss or one of the other mechanisms illustrated in Figure 1.5 (discussed below). In this regard, it is notable that the rate of gene duplication from the X chromosome to autosomes is very high in *D. melanogaster* (Betran *et al.*, 2002)*,* although it remains to be seen how many transfers involve subsequent loss from the ancestral locus on the X chromosome. The recent sequencing of several *Drosophila* genomes should make it possible to investigate this possibility further.

1.3.2.3 Modified Dobzhansky-Muller incompatibility

The mechanism of reproductive isolation suggested above for *JYAlpha* is not novel. It was proposed originally by Werth and Windham in the context of polyploids on largely theoretical grounds (Werth and Windham, 1991). Sufficient data were available (Ferris and Whitt, 1977, Ferris and Whitt, 1979) to indicate that the ultimate fate of most duplicate gene pairs created by whole-genome duplication was silencing and they realized that the loss of alternative copies of duplicated genes in incipient lineages would result in essential genes residing at different map location in different individuals. Subsequent hybridization would result in 1/4 of hybrid gametes receiving no functional copy of each such gene, since the hybrid would be heterozygous at the formerly duplicated loci and the probability of receiving the null allele at both loci is $(1/2)^2$. Werth and Windham showed that even when 70% of the genome is still duplicated and just 500 essential genes exist, that the probability of hybrids producing viable gametes for a pair of lineages that diverged just after the polyploidy event was less than 0.5%. As more genes are returned to single-copy and more realistic numbers of essential genes are considered, the probability of hybrids producing viable gametes, rapidly declines to zero. It is clear that reciprocal gene loss after polyploidization is an extremely powerful mechanism of reproductive isolation. Moreover, it can produce many mutually reproductively isolated lineages, making it perhaps the only mechanism of speciation that can readily explain species radiations.

Lynch and Force subsequently realized that the mechanism proposed by Werth and Windham is a special case of Dobzhansky-Muller incompatibility in which the negative epistatic interaction arises between null alleles fixed at formerly duplicated loci (Lynch and Force, 2000b). In addition, they realized that gene duplication could lead to Dobzhansky-Muller incompatibility in other ways too. For instance, if a single-copy gene were inherited by two daughter lineages and then duplicated in one, a map change may occur depending on how the duplication was resolved. Because the two duplicates initially likely to be identical, either copy can in principle be lost. If that copy is the one at the original "mother" locus and the copy at the "daughter" locus is retained, then the active gene will now be at a different location in the two daughter lineages. More dramatically, if neofunctionalizing mutation is fixed at the "mother" locus (at the expense of the original function) then the ancestral function will be inherited by the "daughter" locus with the result that map location of the ancestral function is again altered ("Duplication and Mother-copy Neofunctionalization" in Figure 1.5). Finally, it is possible that an ancestrally duplicated gene may undergo subfunctionalization independently in the two daughter lineages ("Reciprocal Subfunctionalization" in Figure 1.5). If this occurs there is a 50% chance that the same subfunctions will be retained on homologous chromosomes but equally a 50% chance that reciprocal subfunctionalization will occur and that the two functions will subsequently be found on non-homologous chromosomes. In contrast to the "Classical" Dobzhansky-Muller model none of these mechanisms require any kind of complex interactions between the loci involved, and this alone should suggest that they are likely to be common (Lynch, 2004). Indeed, the only input to the system is new duplicate genes created by mutation. As has been mentioned previously the rate of duplicate gene creation is known to be high in eukaryotes (Lynch and Conery, 2000) suggesting this is unlikely to be a limiting factor and, in the case of polyploids it is clear that the potential for reproductive isolation by reciprocal gene loss (or by either of the other two duplication based mechanisms in Figure 1.5) is enormous.

In Chapter 2 I use the whole-genome duplication that occurred in the ancestor of *S. cerevisiae* and several other yeast species to provide the first evidence that reciprocal gene loss can account for the rapid emergence of multiple new lineages after polyploidization. This establishes that gene duplication maybe responsible not just for the emergence of new genes and new functions, but may also be the basis for the emergence of new species.

Chapter 2. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts

2.1 Preface

This work was published in 2006 in Nature (Scannell *et al.*, 2006a) and is the work of several authors. Kevin Byrne designed and programmed the Yeast Gene Order Browser and performed the statistical tests in Appendix I. Jonathan Gordon worked out the chromosomal rearrangements in Appendix II. Ken Wolfe and I wrote the manuscript and all five authors (those above and Simon Wong) contributed to data curation via the Yeast Gene Order Browser.

2.2 Abstract

A whole-genome duplication (WGD) occurred in a shared ancestor of the yeast species *Saccharomyces cerevisiae*, *Saccharomyces castellii* and *Candida glabrata*. Here we trace the losses of duplicated genes that happened subsequently, and show that the pattern of loss differs among the three species at 20% of all loci. For example, several fundamental transcription factor genes, including *STE12*, *TEC1*, *TUP1*, and *MCM1*, are single-copy in *S. cerevisiae* but were retained in duplicate in *S. castellii* and *C. glabrata*. At many loci, different species lost different members of duplicated gene pairs, so that 4-7% of single-copy genes compared between any two species are not orthologs. This pattern of gene loss provides strong evidence for speciation via a version of the Bateson-Dobzhansky-Muller mechanism, in which the loss of alternative copies of duplicated genes leads to reproductive isolation(Werth and Windham, 1991, Lynch and Force, 2000b). We show that the lineages leading to the three species diverged shortly after the WGD, during a period of precipitous gene loss. The set of loci where single-copy paralogs were retained is biased towards genes involved in ribosome biogenesis and genes that evolve slowly, consistent with the hypothesis that reciprocal gene loss is more likely to occur between duplicated genes that are functionally indistinguishable. We propose a simple unified model in which a single mechanism – passive gene loss – both enabled WGD and led to the rapid emergence of new yeast species.

2.3 Results and Discussion

2.3.1 Using synteny to track the evolution of duplicate gene pairs

We used the Yeast Gene Order Browser (YGOB; ref. (Byrne and Wolfe, 2005)) to compare six yeast species, three of which diverged after their common ancestor experienced a whole-genome duplication (WGD), and three of which diverged from this lineage before the WGD. YGOB compares pairs of genomic regions from post-WGD species (*S. cerevisiae* (Goffeau *et al.*, 1997, Wolfe and Shields, 1997b), *S. castellii* (Cliften *et al.*, 2003) and *C. glabrata* (Dujon *et al.*, 2004)) to single genomic regions in pre-WGD species (*Kluyveromyces waltii* (Kellis *et al.*, 2004), *Kluyveromyces lactis* (Dujon *et al.*, 2004) and *Ashbya gossypii* (Dietrich *et al.*, 2004)) (Figure 2.1). We use the term 'ancestral locus' to describe a locus in a pre-WGD species, or the corresponding duplicated pair of loci in a post-WGD species (i.e., a column in Figure 2.1). Synteny conservation enabled us to determine unambiguously whether each of 2723 ancestral loci was retained in 1 or 2 copies in each post-WGD genome. Where only one copy was retained, the syntenic context allowed orthologs to be distinguished from paralogs (Figure 2.1).

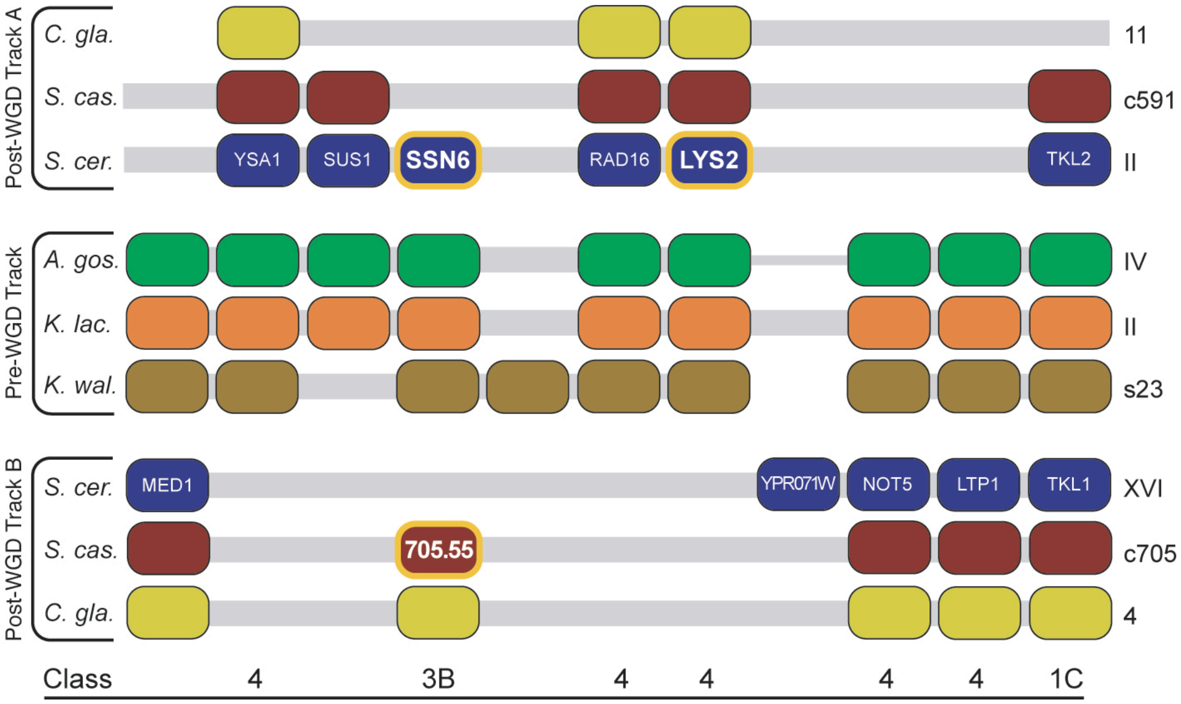


Figure 2.1 Gene order relationships in the region around *S. cerevisiae SSN6* and its homologs, based on YGOB output (http://wolfe.gen.tcd.ie/ygob). Colored boxes represent genes and are not drawn to scale. Chromosomal regions from each pre-WGD species are represented by one horizontal track each. The two corresponding regions in each post-WGD species are represented by two tracks (A and B) at the top and bottom. Homologous genes are arranged in columns. Thick

gray horizontal bars connect genes that are immediate neighbors in the genome. Codes below columns indicate the gene loss class for that ancestral locus, as used in Figure 2.2. Columns without codes did not meet the criteria for scoring.

2.3.2 High rates of differential gene loss among *S. cerevisiae*, *C. glabrata* and *S. castellii*

The fate of an ancestral locus among the three post-WGD species can be classified into one of 14 possible patterns (Figure 2.2). The most common pattern (Class 4, seen at 1957 ancestral loci – 72% of the total) is that all three species have lost the same (orthologous) copy of the gene, such as in the *LYS2* column in Figure 2.1. For clarity we show this as three separate losses in Figure 2.2 but a loss could have occurred in the ancestor of two or three of the species. A further 210 ancestral loci (8%) remain duplicated in all three post-WGD species (Class 0). The other 556 ancestral loci (20%) have had variable fates among the three post-WGD species, which indicates that the consequences of WGD were still being sorted out when these lineages diverged. A striking example is the set of 18 genes that are single-copy in *S. cerevisiae* but two-copy in both *S. castellii* and *C. glabrata* (Class 1B). Transcription factors are disproportionately over-represented in this group (it includes *STE12*, *TUP1*, *GAL11*, *GCR2*, *SFP1*, *YAP3* and *TYE7*; *P* = 0.001 by Fisher test), which suggests that the transcriptional regulatory network in *S. cerevisiae* is simpler than in the other yeasts (Appendix I). *MCM1* and *TEC1* are also in a 1:2:2 relationship among the post-WGD genomes, but these two loci were not counted in Figure 2.2 because the syntenic context around them is not completely conserved.

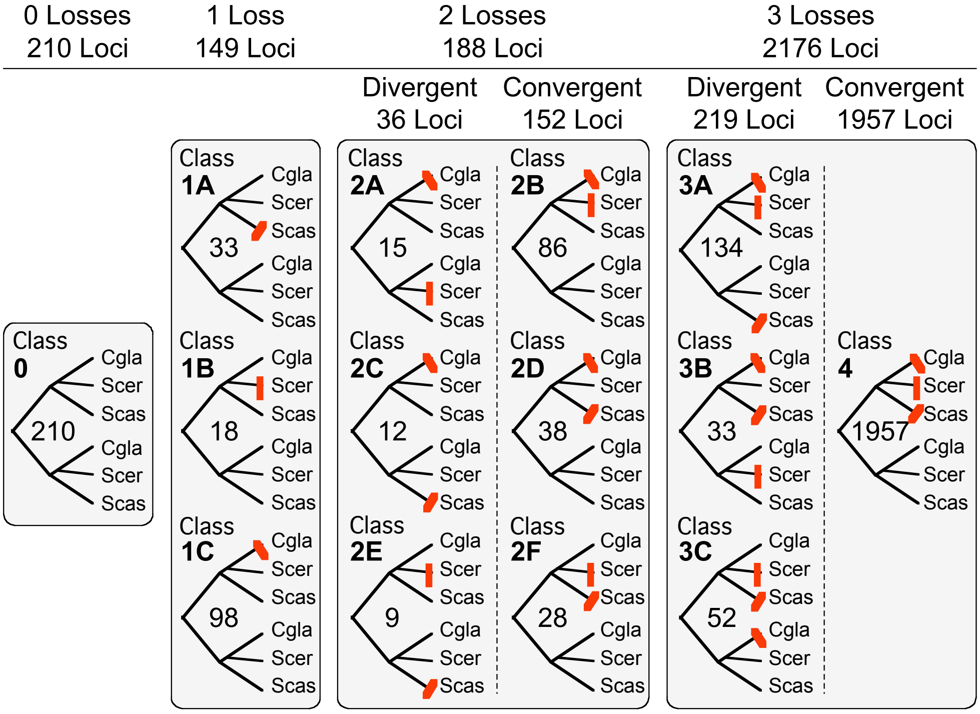


Figure 2.2 Classes of gene loss pattern among 2723 ancestral loci in *S. cerevisiae*, *S. castellii* and *C. glabrata*, and their frequencies. Red marks denote gene absence and are used to group ancestral loci into 14 gene loss classes, described by schematic trees showing the fates of orthologous and paralogous genes. The number of ancestral loci in each gene loss class is shown in the center of its tree. The two sets of species names in each tree denote tracks A and B in arbitrary order. In some cases the absence of a gene copy in two or more species may be due to a single gene loss event on a shared branch, but this does not affect classification. Convergent classes are those where all genes lost are orthologs; divergent classes involve some losses of paralogs in different species.

2.3.3 Reciprocal gene loss is a particular form of differential gene loss that can contribute to reproductive isolation

S. *cerevisiae SSN6* and *S. castellii* gene *705.55* are an example of single-copy paralogs (Figure 2.1). This situation arises when opposite members of a gene pair are lost in two daughter species. Between *S. cerevisiae* and *S. castellii*, 176 of the 2723 loci we surveyed (6.4%; Classes 2E, 3A and 3B in Figure 2.2) show this pattern of reciprocal gene loss (RGL). RGL is a particular form of reciprocal silencing (Werth and Windham, 1991) or divergent resolution (Lynch and Force, 2000b, Taylor *et al.*, 2001) of duplicated genes, and is a property of a pair of genomes. Similarly, there are 198 RGL loci between *C. glabrata* and *S. castellii* (7.3%), and 100 between *S. cerevisiae* and *C. glabrata* (3.7%). Thus, a significant minority of genes that are mutual best BLASTP hits between the post-WGD genomes are not orthologs. More importantly, the process of RGL has the effect of changing the location of the functional copy of a gene (Lynch and Force, 2000b, Werth and Windham, 1991). For instance, *S. castellii* effectively carries a null allele at its locus orthologous to *SSN6*, and *S. cerevisiae* has a null allele orthologous to gene *705.55* (Figure 2.1). If this were the only difference between these two species and they formed a hybrid, the hybrid would be likely to have low fitness because one-quarter of its spores would lack a functional copy of both *SSN6* and gene *705.55* (*S. cerevisiae ssn6* mutants are defective in respiratory growth and sporulation). In fact, 66 of the 176 loci that have undergone RGL between *S. cerevisiae* and *S. castellii* involve essential *S. cerevisiae* genes, so the spore viability of the hypothetical hybrid is reduced to approximately $(0.75)^{66}$ ($= 6 \times 10^{-9}$) due to essential genes alone. Viability will be reduced further by RGL at loci that were not scored in Figure 2.2 due to inadequate synteny conservation (about half the genome), and at loci such as *SSN6* that are not essential but still contribute to fitness. The number of reciprocal losses observed among the post-WGD species is ample to account for their reproductive isolation, notwithstanding the contributions of mechanisms such as interchromosomal rearrangement (Fischer *et al.*, 2001, Delneri *et al.*, 2003) and mismatch repair (Greig *et al.*, 2002b, Hunter *et al.*, 1996).

2.3.4 Reciprocal gene loss is a special case of Bateson-Dobzhansky-Muller incompatibility

The situation described above for *SSN6* and gene *705.55* is a special case of Bateson-Dobzhansky-Muller (BDM) interspecific genomic incompatibility (Coyne and Orr, 2004). The BDM model proposes that negative epistatic interactions between two loci can reduce the fitness of a hybrid. Werth and Windham (Werth and Windham, 1991) and Lynch and Force (Lynch and Force, 2000b) applied the BDM model to duplicated genes, hypothesizing that reciprocal loss (or silencing) of different copies in two species would create a BDM incompatibility, leading to reduced hybrid fitness. RGL at multiple loci could lead to reproductive isolation, and where many duplicated genes exist (as in a polyploid) there is the potential for successive nested speciation events to occur (Werth and Windham, 1991, Lynch and Force, 2000b, Taylor *et al.*, 2001).

2.3.5 Establishing a phylogenetic correlation between reciprocal gene loss and yeast speciation events after whole-genome duplication

To investigate whether RGL was involved in the establishment of reproductive isolation among the post-WGD lineages, we determined the timing of gene losses by estimating the number of duplicated genes surviving at each node on the lineage leading to *S. cerevisiae* (Figure 2.3). To increase the resolution of this analysis we included data from *S. bayanus*

(Kellis *et al.*, 2003), a close relative of *S. cerevisiae* (reproductive isolation between these two species is due to processes other than RGL; ref. (Delneri *et al.*, 2003)). We expressed the ages of the nodes as a proportion of the time (*T*) since the initial divergence of gene pairs created by WGD (see Appendix II and Appendix III). We then estimated the numbers of genes still duplicated in the common ancestors of *S. cerevisiae* and each of *S. bayanus, C. glabrata* and *S. castellii* using two methods: parsimony (which gives the minimum number of genes that must have been retained in duplicate), and a model-based approach (Appendix IV). We consider the latter to be more realistic because it allows for parallel gene losses in different lineages.

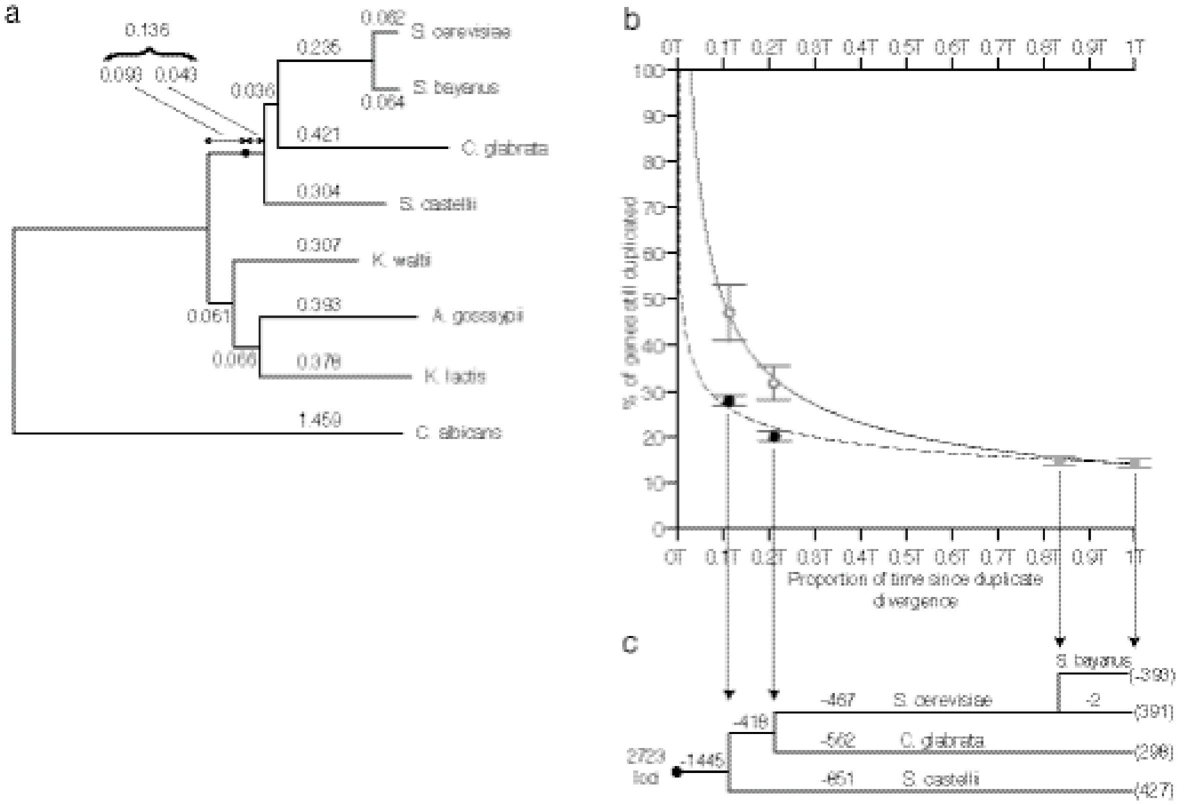


Figure 2.3 Timecourse of duplicated gene loss following WGD. *(a)* Tree reconstructed from 909 protein sequences using a constrained topology (Appendix II) and branch length estimation by maximum likelihood. The black dot indicates the initial divergence of duplicates created by WGD (Appendix III). *(b)* Gene loss curves estimated by the model-based method (open circles and solid curve; Appendix IV) and by parsimony (black circles and dashed curve). Gray circles are common to both methods and show percentages of loci duplicated in *S. cerevisiae* and its common ancestor with *S. bayanus*. The horizontal scale represents the time from the initial divergence of duplicates created by WGD (0*T*) to the present (1*T*) and is derived from the tree in *a* assuming a molecular clock (Appendix III). Power-law curves were fitted to the data(Maere *et al.*, 2005). Standard errors for X (all <2%; omitted for clarity) and Y values were estimated by bootstrapping. *(c)* Numbers of genes lost on each branch leading to post-WGD species, as inferred by the model-based method.

The current numbers of duplicates remaining in each post-WGD genome are shown in parentheses. All numbers refer to the 2723 loci summarized in Figure 2.2.

The parsimony and model-based methods both show a precipitous loss of duplicated genes in the time interval between the WGD and the first speciation event (Figure 2.3b,c). Both methods also show that the fraction of genes retained in duplicate declined appreciably (from 47% to 32% according to the model-based method) in the interval between the first (*S. castellii*) and the second (*C. glabrata*) speciation, even though this corresponds to a very short time period. From this we conclude that gene loss was still occurring rapidly during the emergence of the post-WGD lineages. Moreover, because RGL (by definition) cannot have occurred prior to *S. castellii* diverging from the other post-WGD lineages, and the number of gene losses on the right-hand side of the curve is very few (*S. bayanus* differs from *S. cerevisiae* at only two of the scored ancestral loci), the vast majority of reciprocal losses must have occurred at around the time of the two speciation events. In fact, we estimate that two-thirds of all RGL events occurred between the time of *S. castellii* divergence and time 0.337*T* (Figure 2.3b). The reproductive barriers imposed on these species by RGL are therefore not recent reinforcements but were erected contemporaneously with speciation.

2.3.6 Excess of convergent over divergent gene loss at ancestrally duplicated loci

The fate awaiting most gene pairs formed by WGD was that the duplication was subsequently resolved by deleting one gene copy (Figure 2.2). If the two copies were functionally identical, we would expect that the 'choice' of which copy to delete would be arbitrary. This hypothesis can be tested at ancestral loci that have been resolved independently in more than one post-WGD lineage. We find that in cases of two independent losses, the two retained genes are more often orthologs than paralogs (compare Class 2D to 2C, and 2F to 2E, in Figure 2.2; χ^2 test of homogeneity, $P < 0.05$ for each). A possible explanation for the excess of convergent losses is that at some loci the two copies were not functionally identical, and that the same (better-functioning) copy was retained on both occasions. In contrast, the fact that divergent resolution is seen at some other loci suggests that the choice of survivor at those loci was arbitrary (Classes 2A, 2C, 2E and 3). These observations can be reconciled if some pairs of genes were functionally indistinguishable at the time the duplication was resolved (in which case either copy could be retained), whereas others were functionally distinct (so that a particular copy was preferred by selection).

2.3.7 Slowly evolving loci and those involved in conservative biological processes are more likey to undergo reciprocal gene loss

Differences in the performance of a function can only have been due to sequence differences between the gene copies themselves, or in their cis-regulatory regions. This sequence divergence must have accumulated in the time between WGD and gene loss or, if the WGD was an allopolyploidy, have been inherited from parental species. Therefore, neutral gene loss (which results in divergent resolution half of the time) is expected to be more frequent at ancestral loci that are slowly-evolving or involved in highly conserved biological processes where the potential for functional divergence is low. We tested this prediction and indeed find that loci in Class 3 (all of which underwent RGL between two species) on average evolve 30% slower than Class 4 (where no RGL occurred) (Appendix V; Wilcoxon Rank-Sum test P < 1e-14). Moreover, Gene Ontology terms such as "ribosomal RNA processing", "ribosome biogenesis" and "RNA binding" are disproportionately over-represented among Class 3 loci, as are proteins that are localized to the nucleolus (Huh *et al.*, 2003) and proteins in complexes that bind RNAs (Krogan *et al.*, 2004) (Appendix V). Finally, we also find that genes for snoRNAs, many of which function in rRNA processing, have undergone RGL unusually frequently (Appendix V). Thus, the set of RGL loci appears biased towards those whose functions were most likely to be conserved between duplicates. This functional bias increases the potential contribution of RGL loci to reproductive isolation, because 40% of the Class 3 loci are essential (Guldener *et al.*, 2005) in *S. cerevisiae* as compared to 20% of Class 4 loci (P < 1e-10, χ^2 test).

2.3.8 Passive gene loss as the mechanism for WGD

The passive loss of genes from genomes where there is no selection to retain them is a familiar phenomenon in molecular evolution (Hittinger *et al.*, 2004, Wolfe *et al.*, 1992). We further suggest that passive gene loss is the likely mechanism of the original WGD event in yeast. Our model (Figure 2.4) begins with two haploid cells fusing to form a diploid. If the haploids are from different species, or differ by a chromosomal rearrangement, or carry particular mutations, the resulting diploid may be unable to form viable spores but still able to divide mitotically. If the diploid cell lineage continues to divide mitotically for many generations, it can start to lose one allele from every locus that is not haploinsufficient. During this process there is nothing to prevent an allele at the *MAT* locus being deleted, in which case the cell will behave as a haploid. It can switch mating type, undergo mother-daughter mating, auto-diploidize and so regain fertility (Greig *et al.*, 2002a). Former alleles become separate loci, each of which is homozygous. Continuing loss of redundant gene copies will result in separate lineages that are self-fertile but reproductively isolated from one another by RGL (Figure 2.4).

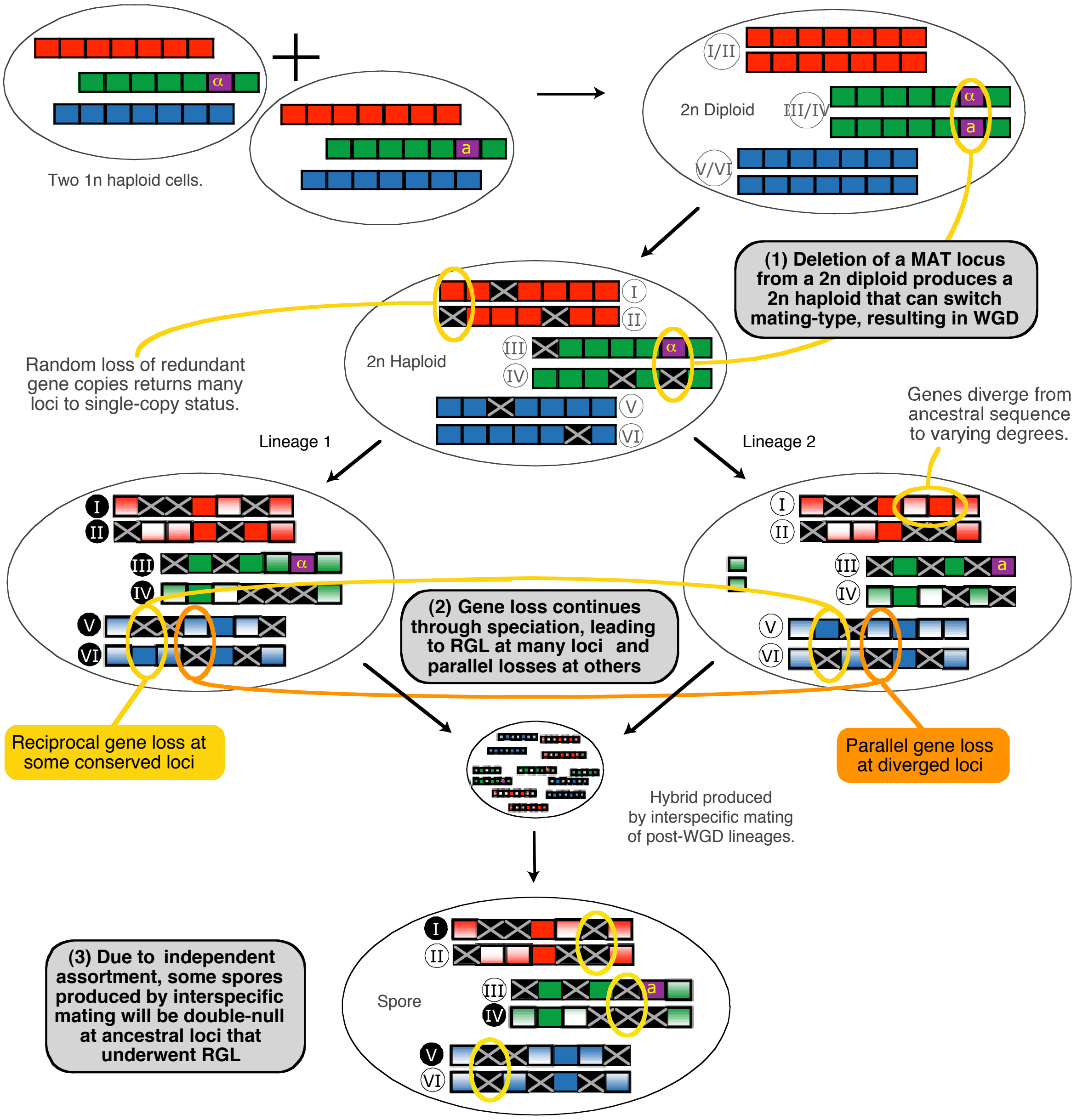

Two 1n haploid cells.
α
a
+
2n Diploid
I/II
III/IV
V/VI
(1) Deletion of a MAT locus from a 2n diploid produces a 2n haploid that can switch mating-type, resulting in WGD
2n Haploid
I
II
III
IV
V
VI
Random loss of redundant gene copies returns many loci to single-copy status.
Lineage 1
Lineage 2
Genes diverge from ancestral sequence to varying degrees.
(2) Gene loss continues through speciation, leading to RGL at many loci and parallel losses at others
Reciprocal gene loss at some conserved loci
Parallel gene loss at diverged loci
Hybrid produced by interspecific mating of post-WGD lineages.
(3) Due to independent assortment, some spores produced by interspecific mating will be double-null at ancestral loci that underwent RGL
Spore

Figure 2.4 Model of passive gene loss as a mechanism of WGD and establishment of reproductively isolated lineages. The steps are discussed in the text. Ovals represent yeast cells. Genes are shown as red, green or blue boxes, except for the *MAT* locus (purple), and are arranged horizontally as chromosomes. Gray X symbols indicate genes that have been deleted. Roman numbering of chromosomes is used to indicate the parent of origin where relevant. Features relevant to each step are ringed in yellow or orange.

2.4 Conclusions

Our results are the first evidence that RGL at multiple ancestrally duplicated genes may lead to speciation, as has previously been hypothesized (but not demonstrated) for polyploid plants (Werth and Windham, 1991, Paterson *et al.*, 2004) and fish (Taylor *et al.*, 2001, Postlethwait *et al.*, 2004). Indeed, because we have shown that RGL is implicated in the emergence of three different lineages, our data support the feature of the modified BDM mechanism (Werth and Windham, 1991, Lynch and Force, 2000b) that most distinguishes it from other theories of reproductive isolation: the ease with which it accounts for multiple speciation events. Finally, by showing that slowly evolving genes and those involved in very fundamental processes are the ones most likely to undergo RGL, our study leads to the remarkable conclusion that these genes, which individually are among the most conservative in the genome, may collectively be responsible for the most radical of evolutionary events.

2.5 Methods

2.5.1 Synteny analysis

We used the YGOB engine (Byrne and Wolfe, 2005) to assess the status and syntenic conservation of loci in *S. cerevisiae*, *S. castellii* and *C. glabrata*. Each ancestral locus (i.e., a column in Figure 2.1, corresponding to two genomic sites in post-WGD species and one site in pre-WGD species) was scored up to 18 times: on tracks A and B in each of the three post-WGD species, and comparing against each of the three pre-WGD genomes. On the basis of homology and syntenic context, the status of each of the six genomic sites in the post-WGD species was designated as one of (1) gene unambiguously present, (2) gene unambiguously absent, (3) gene present but with insufficient syntenic support, (4) gene absent but with insufficient syntenic support. Loci were retained for further analysis if presence or absence could be determined unambiguously on both tracks in all three post-WGD species and if the scoring against all three pre-WGD genomes was not contradictory. This yielded reliable information for 2723 ancestral loci, as summarized in Figure 2.2. The scoring protocol and our implementation are described in ref. (Byrne and Wolfe, 2005). We ignored a small number of ancestral loci where one of the post-WGD species retained neither gene copy. *S. bayanus* was scored relative to the 2723 ancestral loci in *S. cerevisiae* because their genomes are almost completely colinear. 2631 loci in *S. bayanus* had conserved syntenic context (by the criteria above) and manual inspection of candidates generated by the YGOB engine revealed just two differences (*S. bayanus* has retained paralogs as well as orthologs of *HEK2* and *YAT1*).

2.6 Acknowledgements

We are grateful to D. Bradley, G. Conant, J. Conery, B. Cusack, H. Fraser, N. Khaldi, W.-H. Li, A. Lloyd, A. McLysaght, G. Singer, R. Vilgalys and M. Woolfit for discussion. This study was supported by Science Foundation Ireland.

Chapter 3. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication

3.1 Preface

This work has been submitted for publication in the Proceeding of the National Academy of Sciences and is the work of several authors. Carolin Frank assembled the genome (with assistance from Meg Woolfit) and combined the resulting contigs into scaffolds (Appendix VII). Gavin Conant implemented the likelihood model of gene loss after WGD (Appendix XIV). Kevin Byrne modified the Yeast Gene Order Browser code-base to be able to manage the genome data from *K. polysporus* and performed the analyses in Appendix IX and Appendix XI. Ken Wolfe and I wrote the manuscript and all authors contributed to data curation via the Yeast Gene Order Browser.

3.2 Abstract

Among yeasts that underwent whole-genome duplication (WGD), *Kluyveromyces polysporus* represents the lineage most distant from *Saccharomyces cerevisiae.* By sequencing the *K. polysporus* genome and comparing it to the *S. cerevisiae* genome using a likelihood model of gene loss, we show that these species diverged very soon after the WGD, when their common ancestor contained more than 9000 genes. The two genomes subsequently converged onto similar current sizes (5600 protein-coding genes each) and independently retained sets of duplicated genes that are strikingly similar. Almost half of their surviving single-copy genes are not orthologs but paralogs formed by WGD, as would be expected if most gene pairs were resolved independently. This result implicates Dobzhansky-Muller incompatibility after WGD as the likely mechanism of speciation of these yeast lineages. In addition, by comparing the pattern of gene loss among *K. polysporus*, *S. cerevisiae* and three other yeasts that diverged after the WGD, we show that the patterns of gene loss changed over time. Initially, both members of a duplicate pair were equally likely to be lost but loss of the same gene copy in independent lineages was increasingly favored at later timepoints. This trend parallels an increasing restriction of reciprocal gene loss to more slowly evolving gene pairs over time and suggests that as duplicate genes diverged, one gene copy became favored over the other. The apparent low

initial sequence divergence of the gene pairs leads us to propose that the yeast WGD was probably an autopolyploidization.

3.3 Introduction

An ancestor of *S. cerevisiae* underwent whole-genome duplication (WGD) after it had diverged from non-WGD yeast lineages such as *K. lactis*, *K. waltii* and *Ashbya gossypii* (Wolfe and Shields, 1997a, Kellis *et al.*, 2004, Dietrich *et al.*, 2004, Dujon *et al.*, 2004). The WGD had a major impact on the evolution of *S. cerevisiae* and its relatives, most notably by facilitating their adaptation to anaerobic growth (Piskur and Langkjaer, 2004), and contributing to their rapid speciation (Scannell *et al.*, 2006a). In *S. cerevisiae*, about 20% of genes are members of duplicated pairs that were formed in the WGD (Byrne and Wolfe, 2005). The other loci became single-copy again during the sorting-out process (genome reduction) that occurred after the WGD. Similar large-scale loss of copies of duplicated genes from paleopolyploid genomes has occurred during the evolution of plants such as grasses and crucifers (Paterson *et al.*, 2004, Yu *et al.*, 2005, Maere *et al.*, 2005, Schranz and Mitchell-Olds, 2006).

Because the *S. cerevisiae* genome sequence is a single observation of the evolutionary result of the WGD that occurred in a yeast ancestor, it has not been clear whether the set of genes that survived the sorting-out process in *S. cerevisiae* was an inevitable outcome of the WGD, or whether stochastic processes played a major role. Two questions need to be answered: First, are the loci that remain duplicated in *S. cerevisiae* a special subset of the pre-WGD genome, that were somehow more amenable to retention in duplicate after WGD? Second, for loci that are now single-copy in *S. cerevisiae*, was retention of one particular copy preferred over the other? These questions are best addressed by studying the genomes of other yeast species that are descended from the same WGD event. Unfortunately, the post-WGD species whose genomes have been sequenced so far are so closely related to each other that the gene loss process was already nearly complete by the time they diverged (Scannell *et al.*, 2006a). Ideally, we would like to compare genomes that diverged as soon as possible after the WGD, so that relatively little of the sorting-out process occurred on a shared evolutionary branch.

In this study we show that *K. polysporus* is a member of the post-WGD lineage that is most divergent from *S. cerevisiae* and that the vast majority of genes were still duplicated when the lineages leading to these species diverged. We take advantage of the fact that most duplicate gene pairs were resolved twice – once on the *K. polysporus* lineage and once on the *S. cerevisiae* lineage – to study the extent to which the process of gene loss or retention in duplicate was non-random. We find that the two species show similar biases towards retaining duplicated loci with particular biological functions but that, for some functions, the actual genes retained in duplicate are often different. For loci that have become single-copy again, we find that the 'choice' of which copy was discarded became increasingly non-random as time elapsed after the WGD.

3.4 Results and Discussion

3.4.1 *Kluyveromyces polysporus* is a member of the post-WGD clade that is most divergent from *S. cerevisiae*

The phylogeny of hemiascomycete yeasts was recently resolved into 14 clades by Kurtzman and Robnett (Kurtzman and Robnett, 2003) (Appendix VI). The post-WGD species with sequenced (Dujon *et al.*, 2004, Goffeau *et al.*, 1996, Kellis *et al.*, 2003, Cliften *et al.*, 2003, Cliften *et al.*, 2006) or surveyed (Bon *et al.*, 2000, Casaregola *et al.*, 2000, Wong *et al.*, 2003) genomes lie in clades 1-4, while clades 7-14 are outgroups lacking the duplication (Wong *et al.*, 2002). Clades 5 and 6 are monophyletic and sister to clades 1-4, but it was not known if they underwent the WGD or if this event occurred after clades 1-4 split from clades 5-6. We sequenced a few hundred random genomic fragments from *K. polysporus* (in clade 6) and *K. phaffii* (in clade 5). These data suggested that *K. polysporus* and *K. phaffii* both underwent genome duplication, and hence are representative of the WGD lineage most deeply diverged from *S. cerevisiae*. We chose the type strain of *K. polysporus*, originally isolated from soil in South Africa (van der Walt, 1956), for more extensive whole-genome shotgun sequencing.

3.4.2 Genome sequence and gene content of *K. polysporus*

Our *K. polysporus* 7.8x coverage draft genome sequence consists of 290 contigs totaling 14.7 Mb, organized into 41 supercontigs (Appendix VII). We identified 5652 protein-coding genes, 251 tRNAs and at least 39 LTR retrotransposons. The sequence has been submitted to GenBank and can be compared to other yeast genomes using the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe, 2005). In general, the genome is similar in size and gene content to that of *S. cerevisiae*, but some notable differences exist (Appendix X). For instance, several *S. cerevisiae* genes for components of dynein and dynactin (*DYN1*, *DYN3, PAC11*, *ARP1*, *JNM1*, and *NIP100*) have no homologs in *K. polysporus*. It is likely that these gene losses relate to a major phenotypic difference between *K. polysporus* and other yeasts: its asci typically contain 50-100 spores, which are formed by extra mitotic replications after meiosis (van der Walt, 1956, Roberts and van der Walt, 1959). In *S. cerevisiae* dynein and dynactin serve to position the mitotic spindle across the bud neck (Sheeman *et al.*, 2003), but the extra mitoses in *K. polysporus* occur in cells without buds.

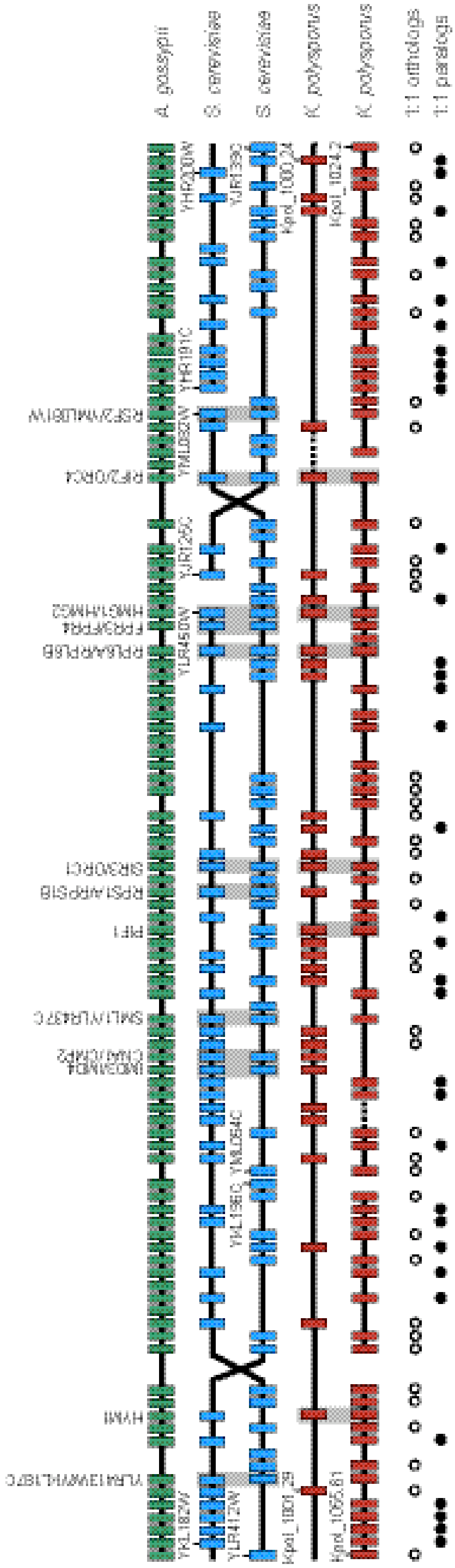


Figure 3.1 Gene order relations in the genomic region around the *SIR3*/*ORC1* gene pair. There are two genomic tracks for each of the post-WGD species *K. polysporus* and *S. cerevisiae*, and a single track for the non-WGD species *A. gossypii*. Colored rectangles represent genes, and genes in the

same column are homologs. Retained duplicated genes in the post-WGD species are highlighted by gray shading and their *S. cerevisiae* names are shown at the top. Solid black lines connect genes that are immediate neighbors on a chromosome or contig. Dashed black lines in *K. polysporus* connect genes that are neighbors on the same supercontig, but between which there is a gap in the genome sequence. The tracks have been drawn to show how YGOB assigns orthology and paralogy between *K. polysporus* and *S. cerevisiae*: the upper tracks in the two species are considered orthologous, as are the two lower tracks. The two X symbols in *S. cerevisiae* show places where YGOB's orthology/paralogy assignments switch between chromosomes. Open and closed circles show how YGOB scored the 74 single-copy loci in this region as 40 orthologs and 34 paralogs, respectively.

3.4.3 The genomes of *S. cerevisiae* and *K. polysporus* are superficially similar but very different in detail

The genome sequence data confirm that *K. polysporus* has undergone WGD. Like *S. cerevisiae*, its genome consists of pairs of sister chromosomal regions that contain some duplicated genes and show a double conserved synteny relationship with single genomic regions in non-WGD species such as *Ashbya gossypii* (Figure 3.1). Among the 3252 ancestral loci that we could reliably compare between the *K. polysporus* and *S. cerevisiae* genomes using the YGOB engine (Byrne and Wolfe, 2005), we identified 450 gene pairs formed by WGD (ohnologs) that have been retained in *K. polysporus* (Table 3.1). Thus, the overall fraction of ancestral loci retained in duplicate in *K. polysporus* is similar to that in *S. cerevisiae* (13.8% and 13.3%, respectively, for the dataset in Table 3.1). However, beneath this superficial similarity, the details of gene loss are so different between the species that it is difficult to tell which of the two sister regions in *K. polysporus* is orthologous to which of the two sister regions in *S. cerevisiae* (Figure 3.1). By contrast, orthologous sister regions are readily identifiable among the other post-WGD species *S. cerevisiae, S. castellii* and *C. glabrata* because they share many gene losses that differentiate them from their paralogous sisters (Scannell *et al.*, 2006a).

Table 3.1 Patterns of differential gene retention between *K. polysporus* and *S. cerevisiae*. Only the 3252 ancestral loci that could be scored reliably (Scannell *et al.*, 2006a, Byrne and Wolfe, 2005) on both sister tracks in both species were counted here. The total numbers of ohnologs are at least 551 in *S. cerevisiae* (Byrne and Wolfe, 2005) and at least 492 in *K. polysporus*, but interspecies rearrangements and gaps in the *K. polysporus* sequence cause some of these loci to be scorable in only one species.

Copy number relationship **(*K. polysporus* : *S. cerevisiae*)**	**Number of ancestral loci**	**Percentage among all loci**	**Percentage among single-copy loci**
2 : 2	212	6.5 %	–
2 : 1	238	7.3 %	–
1 : 2	221	6.8 %	–
1 : 1 (orthologous)	1455	44.7 %	56.4 %
1 : 1 (paralogous)	1126	34.6 %	43.6 %
Total	3252	100.0 %	100.0 %

3.4.4 Approximately equal numbers of single-copy orthologs and paralogs between *K. polysporus* and *S. cerevisiae*

When two closely related genomes are compared, any gene in one species almost invariably has an ortholog in the other species. However, we estimate that only 56% of loci that are single-copy in both *K. polysporus* and *S. cerevisiae* are orthologs (genes that diverged in the speciation event) and the remaining 44% are paralogs (these genes became duplicated in the WGD, and after speciation the two species reciprocally lost different copies) (Table 3.1). The almost equal numbers of orthologs and paralogs around *SIR3*/*ORC1* (Figure 3.1) are typical of the whole genome, as is the loss of approximately equal numbers of genes from both sister regions. Even the apparent small excess of putative orthologs over putative paralogs in Table 3.1 may be an artifact of the algorithm used by YGOB, which assumes that the genomic regions with the greatest shared gene content between species are orthologous (Byrne and Wolfe, 2005). Indeed, the observed 56:44 ratio of orthologs to paralogs among single-copy genes is not significantly different from the 50:50 ratio that would be expected if the two species had gone through completely independent processes of gene loss after WGD (Appendix XI). Importantly, the conclusion that a high proportion of paralogs exists is robust to possible track-assignment errors in YGOB (Appendix XII). The extent of paralogy of single-copy genes observed between *K. polysporus* and *S. cerevisiae* greatly exceeds the levels previously documented in other pairs of species (Scannell *et al.*, 2006a, Town *et al.*, 2006). Our discovery that orthologs do not exist at many loci has negative implications for the prospect of using nuclear gene sequences to resolve the phylogenetic relationships among any group of paleopolyploid species that diverged soon after a WGD.

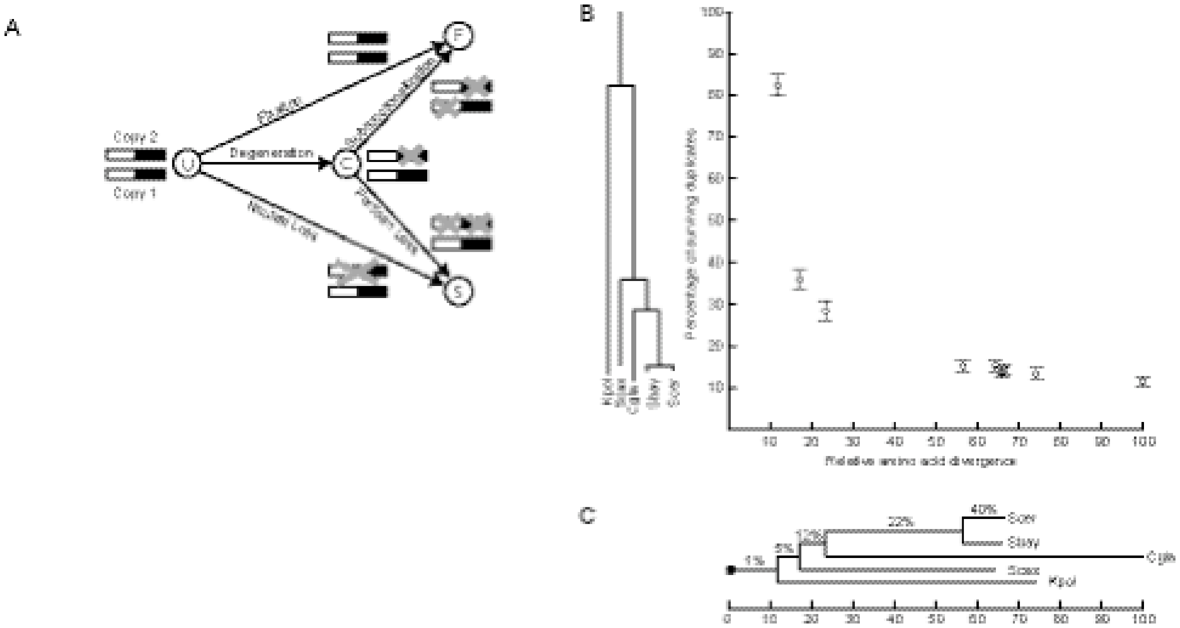
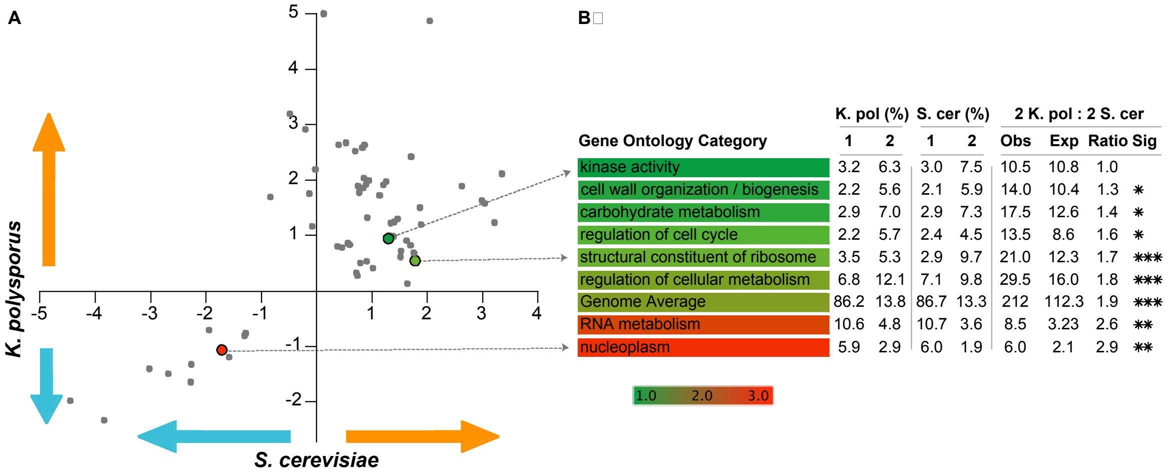


Figure 3.2 Modeling gene pair evolution reveals a changing pattern of gene loss after WGD. **(A)** Our likelihood model of gene pair evolution, showing the four possible states of a pair (**U**, **C**, **S**, **F**; defined in the text), and the permissible transitions between them (arrows). A hypothetical gene pair (copy 1 and copy 2) is shown, containing two domains (white and black boxes). Gray X symbols represent loss-of-function mutations that inactivate either a single domain or a whole gene and cause a pair to move from one state to another. **(B)** Likelihood estimates of the process of gene loss after WGD. Each point on the graph represents the estimated proportion of loci remaining duplicated at a node on the phylogenetic tree. Y-axis values come from the branch lengths of the tree on the left, which was obtained by optimizing the topology and parameters in our likelihood model of gene pair evolution (Appendix XIV). Y-axis values are the total number of loci in states **U** + **C** + **F**, and their error bars were obtained by parametric bootstrapping. X-axis values correspond to amino acid divergence and are taken from the tree in (C); we did not enforce a molecular clock to convert amino acid divergence into time units. **(C)** Tree reconstructed from protein sequences of 11 genes that are duplicated in all five species. Branch-lengths of duplicated branches have been averaged to obtain a species tree. The black dot indicates the time of divergence of duplicated gene pairs. On each branch on the lineage leading to *S. cerevisiae*, the estimated proportion of partisan gene losses (**C** → **S** transitions) is shown as a percentage of all loci returned to single-copy on that branch.

3.4.5 Similar numbers and types of duplicate gene pairs retained in *K. polysporus* and *S. cerevisiae*

The high proportion of paralogs seen between *K. polysporus* and *S. cerevisiae* indicates that these species must have diverged very soon after the WGD and undergone largely independent processes of gene loss. This result was perhaps expected given the phylogenetic position of *K. polysporus*, and is consistent with a Dobzhansky-Muller mechanism of speciation in post-WGD yeasts by reciprocal loss of duplicated genes (Scannell *et al.*, 2006a, Lynch and Force, 2000b, Werth and Windham, 1991). Using a likelihood model of the process of resolution of duplicated gene pairs (described below; Figure 3.2A) we estimate that 82% of loci were still duplicated at the time that *S. cerevisiae* and *K. polysporus* diverged (Figure 3.2B) and the common ancestor of these two species thus had at least 9000 genes (assuming that the pre-WGD yeast had 5000 genes; 5000*1.82 = 9100). Viewed from this perspective it is striking that, after speciation, the numbers of retained duplicates in the two species subsequently dropped independently to the same level (13-14% of the original gene set). Despite this independent history, 47% of the ohnolog pairs in *K. polysporus* have also been retained in duplicate in *S. cerevisiae* (212 of 450; Table 3.1). The number of shared ohnologs is 1.9-fold higher than expected by chance, even allowing for some shared ancestry, and must indicate convergent evolution of genome content ($P < 5 \times 10^{-33}$ by hypergeometric distribution; Appendix XIII). More generally, we find that Gene Ontology (GO) terms that are significantly over- or under-represented among the ohnologs of one yeast species, relative to its singletons, tend to be similarly biased in the other species (Figure 3.3A). Both species show significant under-representation of genes involved in RNA metabolism, mRNA processing, and rRNA processing among duplicates relative to singletons, and significant over-representation of duplicated genes for cytosolic ribosomal proteins, protein kinases, and carbohydrate metabolism.



	K. pol (%)		S. cer (%)		2 K. pol : 2 S. cer			
Gene Ontology Category	1	2	1	2	Obs	Exp	Ratio	Sig
kinase activity	3.2	6.3	3.0	7.5	10.5	10.8	1.0	
cell wall organization / biogenesis	2.2	5.6	2.1	5.9	14.0	10.4	1.3	*
carbohydrate metabolism	2.9	7.0	2.9	7.3	17.5	12.6	1.4	*
regulation of cell cycle	2.2	5.7	2.4	4.5	13.5	8.6	1.6	*
structural constituent of ribosome	3.5	5.3	2.9	9.7	21.0	12.3	1.7	***
regulation of cellular metabolism	6.8	12.1	7.1	9.8	29.5	16.0	1.8	***
Genome Average	86.2	13.8	86.7	13.3	212	112.3	1.9	***
RNA metabolism	10.6	4.8	10.7	3.6	8.5	3.23	2.6	**
nucleoplasm	5.9	2.9	6.0	1.9	6.0	2.1	2.9	**

Figure 3.3 Duplicate gene retention in different Gene Ontology (GO) categories in *K. polysporus* and *S. cerevisiae*. **(A)** Ratios of occurrence of particular GO terms among duplicates, relative to single-copy genes, in the two species. Each point represents a GO term; only terms that are significantly over-represented (direction of orange arrows) or under-represented (direction of blue arrows) in at least one of the two species ($\alpha < 0.001$ by Fisher's exact test) are shown. Colored data-points and dashed arrows show GO terms that also appear in (B). Ratios are presented on a $\log_2$ scale, so 0 indicates a term that is equally frequent among ohnologs and singletons; 3 indicates eightfold over-representation of a GO term among ohnologs, and –3 indicates eightfold under-representation. Note that GO terms are not mutually exclusive so it is not appropriate to calculate a correlation. Details are given in Appendix IX. **(B)** Variation in the extent of overlap between species, within GO categories, of the genes retained in duplicate. The color scale indicates the ratio (Ratio) of the observed number of loci with a GO term retained in duplicate in both species (Obs) to the expected number (Exp). Observed values were obtained from YGOB. Expected values were calculated from the product of the duplicate preservation rates in each species after correcting for the shared evolutionary branch (Appendix XIII). Asterisks show Obs/Exp ratios significantly greater than one (hypergeometric probability: *, $P \leq 0.05$; **, $P \leq 10^{-3}$; ***, $P \leq 10^{-5}$). The other columns show the frequency of the GO term in each species among singletons and among ohnologs (columns labeled "1" and "2" respectively).

3.4.6 The pattern of duplicate gene preservation varies among functional categories

Surprisingly, however, the similarities of GO category biases among duplicates and singletons in the two species do not necessarily mean that the same loci have been retained in duplicate in both. We find that in GO categories that are under-represented among ohnologs relative to singletons, such as 'RNA metabolism' and 'nucleoplasm', the degree to which ohnologs are shared by the two species is greater than in the genome at large (Figure 3.3B). In these categories relatively few loci were retained in duplicate but both species tended to retain the same genes. Conversely, in GO categories that are over-represented among ohnologs relative to singletons, such as 'kinase activity', the level of ohnolog sharing between species is less than the genome average and no more than expected by chance (Figure 3.3B; Appendix XIII). Detailed analysis of a curated set of 75 ancestral protein kinase loci (a subset of the GO term 'kinase activity') shows that *S. cerevisiae* retains 25 duplicated pairs and *K. polysporus* retains 18 pairs, but only six of these pairs are the same; the others are in 2:1 or 1:2 relationships (Appendix VIII). These data suggest that the GO categories that are over-represented among ohnologs are over-represented because certain types of gene (as opposed to particular genes) are favored for preservation in duplicate (Maere *et al.*, 2005, Schranz and Mitchell-Olds, 2006, Seoighe and Gehring, 2004, He and Zhang, 2005a, Hughes and Friedman, 2003). Thus, in answer to the first question we posed in the *Introduction*, there is evidence that *K. polysporus* and *S. cerevisiae* independently converged towards similar categories of retained duplicate genes after WGD. The outcome of the WGD was therefore surprisingly predictable in terms of the functions of retained genes and the eventual overall level of gene retention, but generally unpredictable at the level of the fate of individual genes.

3.4.7 Convergent loss of gene duplicates

To explore the second question – whether the two copies of a gene are equally prone to loss – we included several modes of duplicate gene loss in our likelihood model, and fitted its parameters to YGOB data for five post-WGD species (Appendix XIV). In our previous study of *S. castellii*, *C. glabrata* and *S. cerevisiae* (Scannell *et al.*, 2006a) we found that, at loci where two of the species had each lost one member of an ohnolog pair through independent loss events, convergent losses of orthologous copies were seen about three times more frequently than reciprocal losses of paralogous copies, instead of the 50:50 ratio expected for independent events (Classes 2C/2D and 2E/2F in Figure 2.2). This result suggested that there were selective differences between copies (a particular copy was preferentially retained), but it did not indicate whether these selective differences were present at the time of the WGD or emerged gradually afterwards. By including data from *K. polysporus* it now becomes possible to study how the patterns of gene loss changed over time.

3.4.8 A likelihood model of gene loss after WGD that incorporates partisan gene loss

Our model of gene pair evolution (Figure 3.2A) proposes that after WGD, all gene pairs are initially in a state **U** ('undecided') where the two copies are functionally equivalent and either of them could be lost. Over time, the pair can transition into one of three other possible states: **F** ('fixed') where the duplication has been fixed; **S** ('single-copy'), where one member of the pair has been lost; or **C** ('converging'), a state where both gene copies remain in the genome but there are selective differences such that the loss of one copy (copy 1, for instance) would be deleterious whereas loss of the other (copy 2) would be neutral. We included state **C** in our model to account for the aforementioned excess of convergent losses over reciprocal losses at loci where two independent losses had occurred (Scannell *et al.*, 2006a). Note that loci cannot remain in states **C** or **U** indefinitely. As a hypothetical example, state **C** could include a pair of genes coding for a two-domain protein, but where one of the domains has been inactivated in gene copy 2, with the result that copy 1 is essential but copy 2 is not (Figure 3.2A). This situation can be resolved either by inactivation of the other domain in copy 1 (subfunctionalization and transition to state **F**), or by complete loss of gene copy 2 (transition to state **S**). We refer to the latter as partisan gene loss (as distinct from neutral gene loss) because the identity of the lost gene copy is not arbitrary. If a speciation occurs while the **C**-state pair is still duplicated, any subsequent losses in the descendant species must be of gene copy 2 and so will be convergent. Inclusion of state **C** in the likelihood model significantly improves the fit to the data (Appendix XIV). Moreover, when we compare the likelihoods of the model across all possible branching orders of the post-WGD species, the tree with the highest likelihood (Figure 3.2B, Y-axis) has the expected topology (Scannell *et al.*, 2006a) and places a significant number of gene losses on the shared branch between the WGD and first speciation (of *K. polysporus* from the other post-WGD species), which is evidence against the unparsimonious possibility that *K. polysporus* and *S. cerevisiae* might be descended from two independent WGD events.

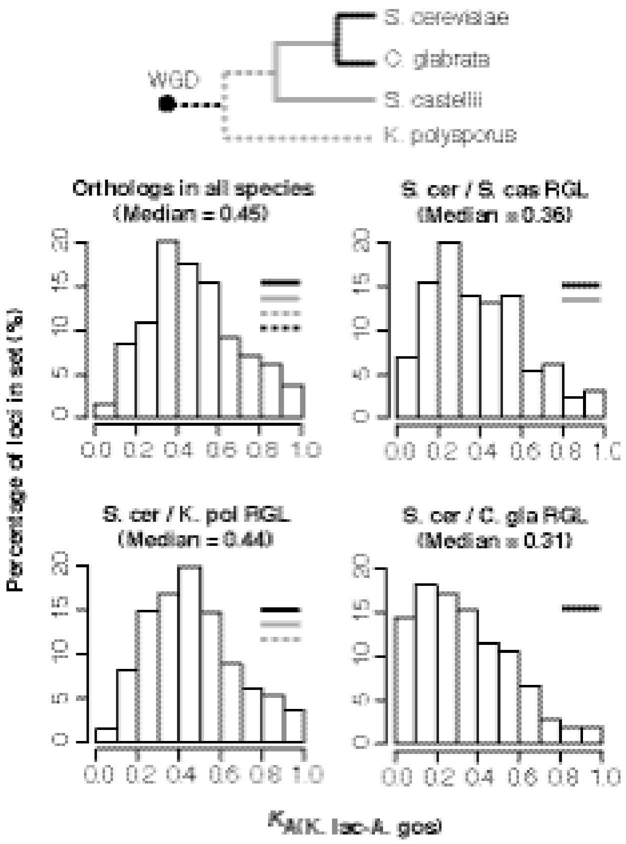


Figure 3.4 Reciprocal gene loss (RGL) is restricted to slower-evolving loci at later timepoints. Histograms show the distribution of levels of nonsynonymous substitution (K_A) between *K. lactis* and *A. gossypii* (a proxy for rate of sequence evolution) for orthologs and sets of loci that have undergone RGL during different time intervals. The patterned lines beside each histogram show the branches of the phylogenetic tree (top) on which RGL could have occurred. RGL loci were always assigned to the most recent category possible. All datasets contain at least 100 loci, and all K_A distributions, except the two on the left, differ significantly from one another ($0.0001 < P < 0.05$ by Wilcoxon rank-sum test).

3.4.9 The pattern of gene loss from duplicated loci changes with time

In our model, gene pairs gradually move out of state **U** and into other states (Figure 3.2A). Because state **U** is the only one that can give rise to neutral gene losses, it is the only state that can lead to reciprocal gene loss (RGL, where two species lose alternative copies of the gene). Therefore we expect that the proportion of duplicated loci that are amenable to RGL will decrease as time elapses after WGD. Furthermore, because the accumulation of sequence divergence presumably tends to make gene pairs leave state **U**, we expect that the set of loci that remain in state **U** will gradually become enriched in slower-evolving loci. The model therefore predicts that loci that underwent RGL soon after WGD will tend to be a random subset of the genome, whereas more recent instances of RGL will tend to have been at more slowly-evolving loci. We tested this hypothesis by partitioning RGL events into different time periods during the evolution of the post-WGD species, and indeed find that RGL events have become increasingly restricted to the slowest-evolving loci (Figure 3.4). The loci that underwent RGL in the most recent interval, after *C. glabrata* and *S. cerevisiae* diverged, have a median rate of amino acid substitution that is only 70% of the median for loci that underwent RGL between *K. polysporus* and *S. cerevisiae*. A separate direct comparison between loci that underwent RGL and those that underwent convergent loss indicates that the former evolve significantly more slowly than the latter, thus excluding the possibility that there is a general trend towards resolving slower evolving loci at later timepoints (P = 0.006 by Wilcoxon rank-sum test; Appendix XV). Furthermore, the loci that underwent RGL between *K. polysporus* and *S. cerevisiae* do not show any significant differences in GO categories compared to single-copy orthologs, contrary to what is seen for later RGL events (Scannell *et al.*, 2006a).

We estimate that the proportion of gene losses that were partisan (*i.e.*, losses from state **C** as opposed to state **U**) rose from 1% immediately after WGD to 40% for losses that occurred after the *S. bayanus*-*S. cerevisiae* speciation (Figure 3.2C and Appendix XVI). This increase can be explained by the accumulation of sequence divergence between the two gene copies, which will inevitably introduce selective differences between them and may cause them to have different deletion phenotypes (state **C**). The answer to our second question is therefore that initially there was little or no selective difference between the two gene copies, but that differences emerged quite quickly as the sequences diverged, which then caused particular gene copies to be favored for retention at single-copy loci. We note also that the fact that only low levels of partisan gene loss are estimated for the earliest timepoints after WGD indicates that the gene pairs were initially very similar in sequence. This inference in turn shows that the WGD event must have been an autopolyploidization or an allopolyploidization between two parental lineages with only minimal sequence divergence between them.

3.5 Conclusion

Our results show that the most recent common ancestor of *K. polysporus* and *S. cerevisiae* must have had more than 9000 protein-coding genes. The two species show markedly convergent subsequent evolution, with both genomes shrinking to about 5600 protein-coding genes, and both retaining similar functional categories of genes in duplicate. That such similarities exist despite the fact that almost half of their single-copy genes are paralogs is remarkable and suggests that WGD provides unique evolutionary opportunities that can be capitalized upon in relatively predictable ways.

3.6 Materials and Methods

3.6.1 Genome survey sequencing of Kluyveromyces polysporus and Kluyveromyces phaffii

The type strains of *Kluyveromyces polysporus* (DSMZ 70294) and *Kluyveromyces phaffii* (MUCL 31247) were obtained from the culture collections of the DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen) and MUCL (Mycothèque de l'Université catholique de Louvain). DNA cloning and sequencing was done by GATC-Biotech (Konstanz, Germany). Genomic DNA was sheared by nebulization and random fragments of 1-2 kb were cloned into plasmids. Both ends of the inserts in 384 plasmids from each species were sequenced. Genes were identified by BLASTX and the gene order in fragments containing >1 gene was compared to other hemiascomycetes. In both *K. polysporus* and *K. phaffii* we found examples of neighboring genes that were close, but not immediate neighbors, in non-WGD species. This suggested that *K. polysporus* and *K. phaffii* are post-WGD species.

3.6.2 Draft genome sequence of *K. polysporus* DSMZ 70294

The type strain of *Kluyveromyces polysporus* (DSMZ 70294) was obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen and used to create genomic DNA libraries. A total of 101,838 sequence reads (79,976 reads from a plasmid library and 21,862 reads from a fosmid library) were assembled into 546 initial contigs using the Phred (Ewing *et al.*, 1998) and Phrap (www.phrap.org) software. Sequence coverage in the Phrap assembly is 7.8x. We manually ordered and oriented 90% of the contigs into 41 supercontigs (Appendix VII), using a combination of physical scaffolds constructed by the program Bambus (Pop *et al.*, 2004) based on fosmid read-pair information, and gene order information from comparisons to other yeast genomes. Within the supercontigs, adjacent contigs with overlapping or consecutive genes at their ends (as inferred by comparison with the non-WGD species *A. gossypii*, *K. waltii* and *K. lactis*) were physically joined by a stretch of 100 N’s into longer contigs, reducing the total number of contigs from 546 to 424. The set of 290 contigs that are larger than 2 kb was retained for subsequent annotation and analysis. The total size of these contigs is 14,703,743 bp, and their N_{50} value is

125,449 bp (that is, half of the bases are in contigs of this size or larger). N_{50} for the supercontigs is 421,604 bp.

3.6.3 Annotation

We wrote a suite of Perl modules to automate identification of conserved features in the genome of *K. polysporus*. The modules provide data-structures to represent genomes at various levels of resolution from exons to scaffolds and wrappers to run external applications. We performed a three-step annotation. First, tRNAscan-SE (Lowe and Eddy, 1997) was used to identify tRNA genes and HMMER v1.8.4 (Eddy *et al.*, 1995) was used to identify putative telomeres and introns. Next, open reading frames (ORFs) above a context-dependent minimum length were identified and all possible gene structures were constructed by merging ORFs across introns, possible sequencing errors and scaffold gaps. Finally, a single gene structure was selected at each locus and all gene structures were evaluated with respect to conservation of sequence in other sequenced yeast genomes, synteny, learned codon-usage patterns and other heuristics. In total, 5927 possible protein-coding genes were identified and 5652 were retained as likely real genes. Perl modules are available on request from scannedr@tcd.ie (D.R.S). Genes were initially named using the scheme *Kpol_{contig_number}.{gene_number}* where the gene numbers were consecutive within the contig. Subsequent manual curation resulted in the elimination of some numbered genes, and the discovery of some extra genes that were given names with lettered suffixes. Sequences have been deposited in GenBank with accession numbers XXXXX-XXXXX and the data can be browsed in the Yeast Gene Order Browser (YGOB).

3.6.4 Yeast Gene Order Browser (YGOB) and Gene Ontology (GO) analysis

We imported the *K. polysporus* genome annotation into our YGOB database, which also includes genome data from the post-WGD species *S. cerevisiae*, *S. bayanus, S. castellii*, and *C. glabrata*, and the non-WGD species *A. gossypii*, *K. lactis* and *K. waltii* (Byrne and Wolfe, 2005). The YGOB engine was then used to classify ancestral loci into different categories of gene loss or retention status, similar to ref. (Scannell *et al.*, 2006a). In this study we worked with two datasets: 3252 ancestral loci that can be reliably scored as either present or absent in both *K. polysporus* and *S. cerevisiae*, and 2299 ancestral loci that can be reliably scored among *K. polysporus*, *S. cerevisiae*, *S. castellii*, *C. glabrata* and *S. bayanus.*

Gene Ontology terms associated with *S. cerevisiae* genes were downloaded from the *Saccharomyces* Genome Database (www.yeastgenome.org) in March 2006 and mapped to the 3252 ancestral loci that satisfy YGOB's quality criteria (Byrne and Wolfe, 2005). Among these, in *S. cerevisiae* 2819 ancestral loci have been returned to single-copy (singletons) and 433 ancestral loci have retained both gene copies (ohnologs), while in *K. polysporus* there are 2802 singletons and 450 ohnolog pairs.

In the analysis shown in Appendix IX we counted the number of singletons in *S. cerevisiae* annotated with each GO term and the number of ohnolog loci at which both gene copies had been annotated with the term. For ohnolog loci at which a GO term had been assigned to only one of an ohnolog pair, the ohnolog count was incremented by one half. We identified GO terms that are either under- or over-represented among ohnolog loci relative to singleton loci using a two-sided Fisher's exact test and report all terms for which the P-value is less than or equal to 0.05, after applying the Benjamini and Hochberg correction for multiple-testing. We transferred all GO annotations mapped to *S. cerevisiae* genes present at an ancestral locus (either a singleton or an ohnolog pair) to the *K. polysporus* genes at that locus and identified GO terms that are either under- or over-represented among ohnolog loci relative to singleton loci as described above.

In Appendix XIII we describe two methods to calculate the expected number of shared duplicate pairs between *S. cerevisiae* and *K. polysporus* and the significance of the observed deviation from these values. In Figure 3.3 we calculated the expected number of shared duplicate pairs for individual GO categories using Method 2 (which accounts for the presence of a shared evolutionary branch) with the additional assumption that the proportion of loci preserved in duplicate on the shared evolutionary branch is the same as the genome average (1.93% / 7.35% = 0.26) and does not vary among GO categories.

3.6.5 Phylogenetics

We used YGOB to select loci that have been retained in duplicate since the WGD by *S. cerevisiae, S. bayanus, C. glabrata, S. castellii* and *K. polysporus* and for which single-copy orthologs were also available in four additional yeast species (*K. lactis, K. waltii, A. gossypii* and *C. albicans*). Ignoring the *K. polysporus* genes, we first used YGOB to determine which of the two gene copies in *S. bayanus, C. glabrata* and *S. castellii* are orthologous to each of the two gene copies in *S. cerevisiae*. We were able to partition these duplicates into two clades (DC1, DC2), each consisting of four syntenic orthologs, for 92 loci.

Because of the high level of reciprocal gene loss between *K. polysporus* and *S. cerevisiae* we used phylogenetic methods rather than YGOB (which relies on conservation of synteny) to determine which of the two gene copies in *K. polysporus* is orthologous to each of the two gene copies in *S. cerevisiae.* For each locus we used ClustalW (Thompson *et al.*, 1994) and Gblocks (Castresana, 2000) to generate an alignment from all 14 sequences and used Shimodaira-Hasegawa tests (Shimodaira and Hasegawa, 2001) (implemented in Tree-Puzzle (Schmidt *et al.*, 2002)) to determine whether one of the two possible topologies was preferred: either *K. polysporus* copy 1 clusters with DC1 and *K. polysporus* copy 2 clusters with DC2 or *vice versa*. Loci at which there was significant (α = 0.05 level) support for one topology over the other were retained.

We also sought to exclude loci that may have undergone gene conversion (Sugino and Innan, 2005). We used Phyml (Guindon and Gascuel, 2003) to draw unconstrained trees for each locus with all five pairs of duplicates and the corresponding single ortholog in *K. lactis*. Any loci for which either DC1 or DC2 (including the appropriate *K. polysporus* ortholog) were not reconstructed were discarded. Eleven loci were retained for further analysis (*S. cerevisiae* gene names: *YBP2*/*YBP1*, *SWH1*/*OSH2*, *HST1*/*SIR2*, *FAR10*/*VPS64*, *SBE2*/*SBE22*, *GEA1*/*GEA2*, *SDT1*/*PHM8*, *SIR3*/*ORC1*, *FSH2*/*FSH3*, *CDC50*/*YNR048W* and *TRF4*/*TRF5*), and super-alignments of these loci were used for phylogenetic analysis.

At any given locus all the gene copies in DC1 (or DC2) are orthologous to one another and are paralogous to the gene copies in DC2 (or DC1). There is however no relationship between the gene copies in DC1 at one locus and the gene copies in DC1 at other loci. It is therefore possible to concatenate gene copies from DC1 at one locus with gene copies from DC2 at other loci (provided all gene copies in DC1 are treated consistently) when constructing a super-alignment. We used this fact to exclude the possibility that generating a single super-alignment might result in concatenation of the faster-evolving clades (DC1 and DC2 can evolve at very different rates) at several loci. Instead, we generated 100 super-alignments (4045 amino acid sites each) in which the DC1/DC2 designation was randomly reversed with probability 0.5 for each locus. Finally, for each of the 100 super-alignments a single bootstrap-replicate was generated using ‘seqboot’ in the Phylip package and these – rather than the original super-alignments – were retained for phylogenetic reconstruction.

Because the phylogenetic relationships between the yeasts used in this study are known (Scannell *et al.*, 2006a, Kurtzman and Robnett, 2003) we optimized branch-lengths but not the topology (modified to include *K. polysporus*) for each of 100 bootstrap-replicates using a WAG + I + G(8) + F model. Finally, branch-lengths were averaged between duplicate clades and across all 100 bootstrap-replicates to obtain the tree in Figure 3.2C. We did not correct the tree in Figure 3.2C for the effect of accelerated protein sequence evolution after WGD because we found that the method used in (Scannell *et al.*, 2006a) yielded a small negative length for the branch between the WGD and the *K. polysporus* divergence.

3.7 Acknowledgements

We thank D. Lundin and G. Manning for comments, and the SFI/HEA Irish Centre for High-End Computing (ICHEC) for computational facilities and support. A.C.F. is supported by a Swedish Research Council (Vetenskapsrådet) fellowship. This project was supported by Science Foundation Ireland.

Chapter 4. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast

4.1 Preface

This work has been submitted to Genome Research and is the work of two authors. I designed and carried out the research with supervision from Ken Wolfe. Ken Wolfe and I wrote the manuscript together.

4.2 Abstract

It is widely accepted that newly arisen duplicate gene pairs experience an altered selective regime that is often manifested as an increase in the rate of protein sequence evolution. Many details about the nature of the rate acceleration remain unknown, however, including its typical magnitude and duration, and whether it applies to both gene copies or just one. We provide initial answers to these questions by comparing the rate of protein sequence evolution, among eight yeast species, between a large set of duplicate gene pairs that were created by a whole-genome duplication (WGD) and a set of genes that were returned to single-copy after this event. Importantly, we employ a new method that takes account of the tendency for slowly-evolving genes to be retained preferentially in duplicate. We show that on average proteins encoded by duplicate gene pairs evolved at least three times faster immediately after the WGD than equivalent single-copy genes. Although this rate subsequently declined rapidly, it has not yet returned to the typical rate for single-copy genes. In addition, we show that although duplicate gene pairs often have highly asymmetric rates of evolution, even the slower members of pairs show evidence of a burst of protein sequence evolution immediately after duplication.

4.3 Introduction

Theory indicates that one of three fates awaits all newly-created duplicate gene pairs (Force *et al.*, 1999, Lynch *et al.*, 2001): nonfunctionalization (one copy is disabled and eventually lost, restoring the ancestral genotype and phenotype), subfunctionalization (the ancestral gene functions are partitioned between the two duplicates, thus restoring the ancestral phenotype but altering the genotype) or neofunctionalization (retention of both gene copies confers an advantage, so both genotype and phenotype are altered). In the event that one member of a pair becomes nonfunctionalized, the selective constraints that operated on the single ancestral gene copy are presumed to be inherited by the remaining functional duplicate. Indeed, unless the retained gene copy resides at a different genomic location than the ancestral gene copy (in which case reproductive isolation may emerge between lineages; Lynch and Force, 2000b, Scannell *et al.*, 2006a), the net effect of nonfunctionalization is likely to be the restoration of the pre-duplication *status quo*. By contrast, if a gene pair is either subfunctionalized or neofunctionalized then both members will be maintained by selection but the presence of (partial) redundancy may result in one or both genes experiencing an altered selective regime relative to the ancestral single-copy state. Thus, gene duplication may initiate a period of altered molecular evolution and duplicate preservation may result in this being prolonged. However, because the vast majority of new genes originate by gene duplication, the distinction between duplicated and single-copy genes is essentially a semantic one, and it is apparent that the evolutionary dynamics of a genes formed by duplication must eventually change into the dynamics of single-copy genes.

Several authors have reported that duplicate genes exhibit an elevated rate of protein sequence evolution (Lynch and Conery, 2000, Nembaware *et al.*, 2002, Jordan *et al.*, 2004) and this has been interpreted to mean that both members of a pair are subject to weaker purifying selection than single-copy genes (Kondrashov *et al.*, 2002). However, it has also been observed that duplicated genes may exhibit asymmetric protein sequence evolution (i.e. the pair consists of a "slow" gene copy and a "fast" gene copy; Van de Peer *et al.*, 2001, Conant and Wagner, 2003, Zhang *et al.*, 2003, Brunet *et al.*, 2006) and this has been taken as support for the Ohno model of evolution after gene duplication (Kellis *et al.*, 2004), which hypothesizes that one member of a pair (the "slow" copy) maintains the ancestral rate of evolution (and the ancestral role) while the "fast" copy may evolve to optimize a novel beneficial function (Ohno, 1970). It is worth pointing out however, that the observations themselves are not mutually exclusive. For instance, it is possible that young and old duplicated pairs are subject to different selection pressures and that age differences between datasets have contributed to different conclusions. Moreover, in either case substitutions are presumed to be accepted for the same underlying reason: the presence of a redundant gene copy complements any loss of the ancestral function (either due to a loss-of-function mutation or due to the gain of an alternative function) in its paralogous partner. An important corollary of this is that as duplicates accumulate substitutions they become progressively less able to complement one another (Gu *et al.*, 2003) and at some point must fail to do so completely. Surprisingly, few authors have tried to estimate how long after gene duplication this loss of complementation occurs (Lynch and Conery, 2000). In this study we address this question and attempt to clarify previous observations by simultaneously examining three aspects of duplicate gene pair evolution: the magnitude of the increase in the rate of protein sequence evolution exhibited by duplicate genes; the symmetry of this effect (whether it is exhibited equally by both copies); and the duration of the effect (how soon after gene duplication the rate of protein sequence evolution returns to the pre-duplication level). We do this by comparing rates of protein sequence evolution in 85 loci that were retained in duplicate and 808 loci that were returned to single-copy after the yeast whole-genome duplication (WGD; Wolfe and Shields, 1997b, Dietrich *et al.*, 2004, Kellis *et al.*, 2004). In addition, our approach differs in a number of ways from those taken by previous authors.

First, we have chosen to study only with genes for which either single-copy orthologs or double-copy co-orthologs are available in eight yeast species, four of which diverged after a WGD in their common ancestor (post-WGD yeasts; *S. cerevisiae, S. bayanus, C. glabrata* and *S. castellii*) and four of which diverged from this lineage prior to the WGD (*K. waltii, K. lactis, A. gossypii* and *C. albicans*; we refer the first three as non-WGD yeasts and use *C. albicans* as an outgroup). More specifically, our set of single-copy loci consists of genes that are single-copy in the three non-WGD yeasts and that are also currently single-copy in all four post-WGD yeasts. By contrast, although the genes in our double-copy dataset also possess only a single ortholog in each of the non-WGD yeasts, they have been retained in duplicate since the WGD in the other four yeasts. Our motivation for requiring that all genes in our datasets have single-copy orthologs in multiple non-WGD species is discussed below, but the motivation for studying gene pairs that are retained in duplicate in multiple post-WGD yeasts is simple: it allows us to study the same gene pairs at successive time intervals after gene duplication.

The second major difference between our approach and previous studies is that we use concatenated alignments to study the group properties of duplicates and single-copy genes. We estimate the average increase, after the WGD, in the rate of protein sequence evolution in double-copy sequences on different branches of the phylogenetic tree. Although concatenating alignments in this manner prevents us identifying individual gene pairs that exhibit particularly asymmetric protein sequence evolution or that are evolving very rapidly, it increases our power to identify general evolutionary trends associated with gene duplication. In this regard our experimental design is similar to the study by Lynch and Connery (2000), in which data from a large number of pairs were fit to an evolutionary model in order to make inferences about the evolution of the "average" or "ideal" gene pair.

Finally, we use a method we have developed recently (Scannell *et al.*, 2006a) to correct for the fact that genes that are retained in duplicate do not comprise a random sample of the genome but are, on average, more slowly evolving (prior to duplication) than genes that are not retained in duplicate (Davis and Petrov, 2004). This bias can lead to a scenario where an inter-species comparison of the rates of protein sequence evolution between sets of orthologous genes that either have paralogs or do not have paralogs can fail to detect a true increase in the rate of protein sequence evolution in the former set. It is likely that this effect has been a significant source of error in previous studies (Davis and Petrov, 2004) and, by correcting for it, we show that although the rate of protein sequence evolution in duplicated genes in modern *S. cerevisiae* has declined significantly from its high immediately after the WGD, that it has still not returned to the pre-duplication rate for at least one member of most gene pairs.

4.4 Results

4.4.1 Assessing the affect of gene duplication on protein sequence evolution

Our method for assessing the impact of gene duplication on the rate of protein sequence evolution consists of two steps. First we assembled two super-alignments, called A1 and A2. A1 is a concatenation of the aligned protein sequences of genes in our single-copy dataset (324,540 columns from 808 loci that are single-copy in all seven species), and A2 is a concatenation of the aligned protein sequences in our double-copy dataset (33,720 columns from 85 loci that are double-copy in the four post-WGD species and single-copy in the three non-WGD species). We then used the procedure described in Scannell *et al.* (2006a) to mitigate any rate biases between sequences in the super-alignments A1 and A2 due to the preferential retention of slowly evolving genes in duplicate (Davis and Petrov, 2004). Briefly, this procedure matches each column in A2 with a "control" column in A1 that contains exactly the same amino acid residues in some non-WGD species, and so can be considered to be following a similar evolutionary trajectory in the non-WGD species.

We then use only these matched columns to assemble two new super-alignments, A1’ and A2’. Because there are about ten times more columns in A1 than A2, it is possible to find a matching column in A1 for almost every column in A2.

The second step in our procedure is to perform maximum likelihood branch-length evaluation on A1’ and A2’ using the established phylogenetic relationships among the yeast species represented in these super-alignments (see *Methods* and Scannell *et al.*, 2006a). Tree T1 is derived from the single-copy sequences in super-alignment A1', and tree T2 is derived from the double-copy sequences in super-alignment A2' (Figure 4.1A, left). We then modify the topologies of T1 and T2 to produce a final pair of trees, T1’ and T2’, with a single topology (Figure 4.1A, right). In the case of T2 we simply average the lengths of all the duplicated branches between the clades labeled ‘Copy 1’ and ‘Copy 2’ (Figure 4.1A, bottom) and collapse one of the redundant clades. In the case of T1, we partition the branch on which the WGD occurred into pre- and post-duplication branches as described (Figure 4.1A, top; Chapter 2). Because T1’ and T2’ have identical topologies (Figure 4.1A, right) we can estimate the rate of protein sequence evolution on T2’ relative to T1’ by comparing branch lengths between them. For convenience, we report the length of each branch on T2’ as a percentage of the length of the corresponding branch on T1’ in all subsequent analyses. In addition, because we are only interested in this scaled value (*i.e.* the rate of protein sequence evolution of double-copy sequences relative to appropriate single-copy control sequences) and not the actual length of the branches on either T1’ or T2’ we will refer to this percentage simply as the rate of protein sequence evolution.

In Appendix XVII we show that our column-matching procedure can substantially reduce the effect of the bias noted by Davis and Petrov (2004), that slowly evolving genes are more likely to be retained as duplicates. We first confirm their result for our dataset. In the non-WGD species *K. waltii*, *K. lactis* and *A. gossypii* the average rate of protein sequence evolution of genes that were retained in duplicate in post-WGD species is only 78-80% of the average rate of those that were not retained in duplicate (Appendix XVII, Panel A). That is, the median evolutionary rate in the non-WGD species for genes in set A2 is about 20% lower than for those in set A1, even though the distinction between sets A2 and A1 concerns whether or not they are duplicated in a different group of species. We then demonstrate that the column-matching can reduce this rate bias. We performed column-matching in three ways, by matching columns in A1 to those in A2 on the basis of having identical amino acid residues in two, three or four non-WGD species. As the number of matched species increases, the median rate of protein sequence evolution in non-WGD species in A2' relative to A1' increases from 92% to 94% to 97% (Appendix XVII, Panels B-D), indicating that Davis and Petrov's bias is being eliminated. We note that this is not a trivial consequence of the column-matching procedure (which causes the non-WGD sequences to be identical in A1’ and A2’) because the branch lengths in the post-WGD clade change by a similar amount (compare Appendix XVII, Panel A and Appendix XVII, Panel D; the median changes in the rate of protein sequence evolution in the non-WGD and post-WGD clades are 18% and 19% respectively). Column-matching with four non-WGD species is the most effective method, but to achieve this we had to use data from the non-WGD species *S. kluyveri*, which has not been completely sequenced, and the missing data has the consequence that we cannot find matches for 17% of the columns in A2 (Appendix XVII, Panel D). For the remainder of this study we therefore chose to use super-alignments made by column-matching for three non-WGD species (*K. lactis, K. waltii* and *A. gossypii*), because this criterion allows matching of almost all columns in A2 (99.7%) and the amelioration of the rate bias is only slightly less than when four non-WGD yeasts are used (Appendix XVII, Panel C).

4.4.2 Elevated rate of protein sequence evolution in double-copy sequences relative to single-copy control sequences

After controlling for the Davis and Petrov effect as described above, we find that the relative rate of sequence evolution of proteins in the A2 set is greater than the expected 100% in all branches descended from the WGD (median 128%; range 111–342%) but very close to this value for all others (median 95%; range 93–107%) (Figure 4.1B). The observation that all of the branches in the post-WGD clade are significantly longer than expected indicates that double-copy sequences experience a considerable increase in the rate of protein sequence evolution relative to equivalent single-copy sequences. As we discuss in more detail below, this appears to be true for duplicates derived from the WGD even in modern *S. cerevisiae* (the rate of protein sequence evolution on the terminal *S. cerevisiae* branch is 111 ± 3%) and appears to be especially true on the earliest branch after duplication (342 ± 54%). We also note that the change in the rate of protein sequence evolution on successive branches after the WGD in Figure 4.1B declines monotonically on successive branches from the WGD to modern *S. cerevisiae* (342% > 128% > 124% > 111%), which is consistent with a progressive restoration of purifying selection after gene duplication and conforms precisely to the expectation under the model outlined in the *Introduction*.

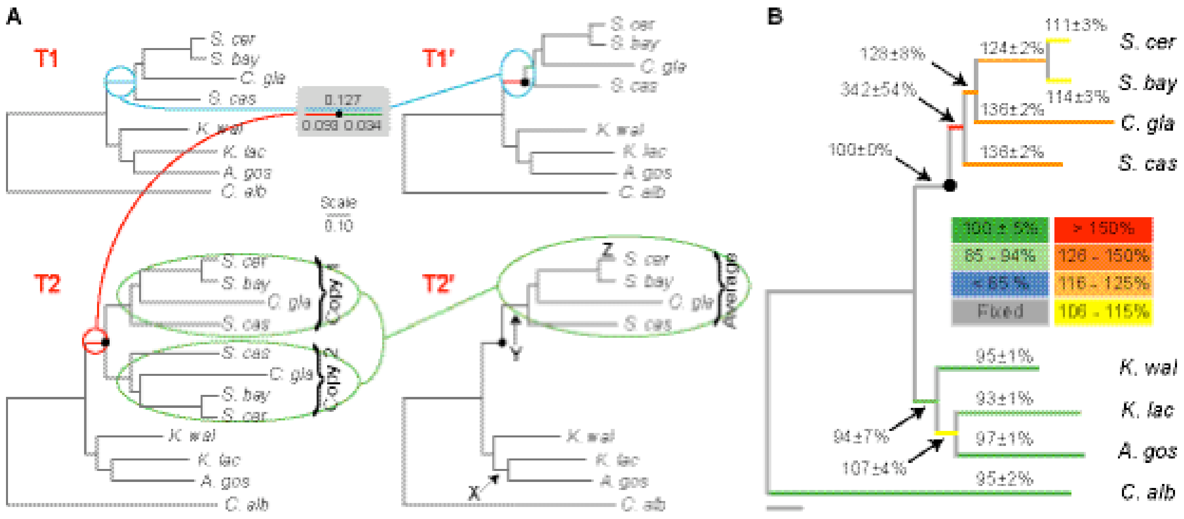


Figure 4.1 Measuring the increase in the rate of protein sequence evolution after gene duplication. (A) Construction of a pair of topologically identical trees, T1' and T2' (right), from a tree derived from single-copy sequences only (T1; obtained from super-alignment A1') and a tree derived from single- and double-copy sequences (T2; obtained from super-alignment A2'). The tree T1' was derived from the tree T1 by partitioning the branch between the divergence of the non-WGD yeasts and the divergence of *S. castellii* from the *S. cerevisiae* lineage into pre- and post-duplication segments (light-blue line and grey box). As in Scannell et al. (2006a) we assumed that the length of the pre-duplication branch on T1 is the same as that on T2 (red line). The tree T2' was derived from the tree T2 by averaging the lengths of all duplicated branches between the post-WGD clades labeled 'Copy 1' and 'Copy 2' (light-green ovals). A black circle (●) indicates the inferred point of duplicate gene divergence. The branches labeled X, Y and Z on T2' are referred to in the text. (B) Tree showing the length of branches on T2' as a percentage of the length of the corresponding branches on T1', which is a measure of the rate of evolution of double-copy sequences relative to single-copy sequences. Percentages (± one standard deviation) are averages from 100 bootstrap replicates (see *Methods*). The branch lengths drawn are the averages on T1' from the same 100 bootstrap replicates. Branches are colored according to the arbitrary scale shown.

We performed a variety of control experiments to confirm our observations. First, we considered the possibility that the column-matching procedure we employed might artificially inflate the estimated rate of protein sequence evolution among double-copy sequences (although Appendix XVII, Panel A strongly suggests that this is not the case). To test this we replaced A2 with an equal number of randomly sampled columns from A1 and carried out all other steps as previously. As expected for a negative control we detected no acceleration on any branch (0, Panel A). We also considered that spurious matches between columns in A1 and A2 based on rare combinations of amino acids in *K. lactis, K. waltii* and *A. gossypii* might cause us to overestimate rate of protein sequence evolution in double-copy sequences. We therefore excluded all columns from A1' and A2' that possessed an amino acid combination in *K. lactis, K. waltii* and *A. gossypii* that was observed less than five times in either A1 or A2. Excluding these columns causes our estimates of the rate of protein sequence evolution in double-copy sequences to be slightly increased for the post-WGD clade, and probably slightly improved for the non-WGD clade (0, Panel B), but ultimately supports the same conclusions as Figure 4.1B. Finally, to exclude the possibility that the differing numbers of sequences in A1 and A2 made a comparison between trees derived from these super-alignments inappropriate or that the tree processing steps introduced an error of some kind, we removed all the sequences from one of the duplicate clades (*e.g.,* the sequences corresponding to 'Copy 2' in Figure 4.1A, bottom left) from A2 and repeated all other steps as previously. Again, the results were not significantly affected (0, Panel C) and we conclude that sequences of retained duplicate gene pairs evolve faster at the protein sequence level than equivalent single-copy sequences.

4.4.3 Double-copy sequences experience a burst of protein sequence evolution immediately after duplication

As expected the greatest increase in the rate of protein sequence evolution among double-copy sequences is observed immediately after the WGD. On the branch between the WGD and the divergence of *S. castellii* from the *S. cerevisiae* lineage we estimate that double-copy sequences evolved on average at 342±54% the rate of equivalent single-copy sequences (Figure 4.1B), and this is probably a lower bound estimate for several reasons. First, we averaged the rate of protein sequence evolution between the two duplicate clades and if (as we show below) the increase in the rate of sequence evolution is usually experienced primarily by one member of each duplicate pair, the increase in some gene copies could be up to twice that shown in Figure 4.1B. This is similar to the tenfold average increase in the nonsynonymous substitution rate detected by (Lynch and Conery, 2000). Second, we did not attempt to remove duplicated pairs that are undergoing gene conversion from our dataset, except for those encoding cytosolic ribosomal proteins (see *Methods*). Since gene conversion will cause us to underestimate the lengths of branches on T2 only (Figure 4.1A, bottom left), it is possible that it has depressed our estimates of the rate increase in double-copy sequences. Finally, we note that all the sequences in A1 must have been duplicated for at least a short period of time after the WGD (Scannell *et al.*, 2006a) so it is possible that they also experienced a brief increase in the rate of protein sequence evolution. If this is the case then comparing branch lengths between T1' and T2' will tend to underestimate the increase in the rate of protein sequence evolution attributable to gene duplication.

The branch from the WGD to the first speciation event accounts for approximately 10% of the time from the WGD to the present so the increase in the rate of protein sequence evolution we observe is the average value over a reasonably long period of time. This suggests that the increase may have been more modest towards the end of this branch and potentially much greater immediately after the WGD. We used the genome sequence of *K. polysporus* (Scannell *et al.*, 2006b) to investigate this possibility further. Because *K. polysporus* diverged from the *S. cerevisiae* lineage on the branch between the WGD and the divergence of *S. castellii*, it should allow us to partition the branch immediately after the WGD into two segments. On the branch immediately after the WGD we expect the estimated rate of protein sequence evolution to be greater than 342±54% and on the other we expect it to be less. Surprisingly however, when we applied our method to super-alignments that included *K. polysporus* sequences, $A1_{Kpol}$ and $A2_{Kpol}$ (similar to A1 and A2 above but with sequences from *K. polysporus*; see *Methods*), we were unable to estimate reliably the length of the branch between the WGD and the divergence of *K. polysporus* on tree T1' (this is done by comparison to T2'; see Figure 4.1A). In 34 of 100 pseudo-replicates we obtained a very short branch length (on the order of 0.01 amino acid substitutions per site) and consequently estimated the rate of protein sequence evolution in double-copy sequences immediately after the WGD to be >1000% of the single-copy rate in many cases. However, the remaining 66 pseudo-replicates indicated a short negative branch, and the average of all one hundred pseudo-replicates was not distinguishable from zero (-0.003 ± 0.01 amino acid substitutions per site). Although this is nominally consistent with our previous conclusion that *K. polysporus* and *S. cerevisiae* diverged very soon after the WGD (Scannell *et al.*, 2006b) additional data (not shown) indicate that two sources of error may be contributing to underestimation of the length of the branch between the WGD and this divergence event. First, it is possible that gene conversion that occurred between duplicate pairs prior to the divergence of the *K. polysporus* and *S. cerevisiae* lineages cause the WGD to appear to occur at a later time on tree T2 than was actually the case. Second, we have previously shown that it is very difficult to determine whether genes in *K. polysporus* are orthologs or paralogs (created by the WGD) of their closest homologs in the other post-WGD yeast species (Scannell *et al.*, 2006b). If some of the single-copy *K. polysporus* sequences in $A1_{Kpol}$ are paralogs rather than orthologs of the sequences from the other post-WGD species in $A1_{Kpol}$, then we will infer that *K. polysporus* diverged from these species earlier than was actually the case. The combination of these two sources of error (gene conversion in T2 and cryptic paralogs in T1) will cause us to underestimate the length of the branch between the WGD and the divergence of *K. polysporus* on T1' (Figure 4.1A, top). We are therefore currently unable to confirm that the rate of protein sequence evolution on the branch between the WGD and the divergence of *K. polysporus* is greater than 342±54%. However, we were able to estimate that the rate of protein sequence evolution on the branch between the divergence of *K. polysporus* and *S. castellii* is 252±37%, which is consistent with the pattern of a sudden rate increase after WGD followed by a gradual slowdown.

4.4.4 An elevated rate of protein sequence evolution persists in double-copy sequences for an extremely long period of time after duplication

The rate of protein sequence evolution in double-copy sequences on the terminal *S. cerevisiae* branch is higher than for equivalent single-copy sequences (111±3%), suggesting that duplicate pairs still experience a more permissive selective regime due to the presence of a partially redundant gene-copy. Because it is surprising that this effect is still observed so long after the WGD (100 - 300 Myr; Wolfe and Shields, 1997b, Friedman and Hughes, 2001), we verified this result by performing a codon-based analysis of selective constraint between orthologous sequences from *S. cerevisiae* and *S. bayanus* that are either derived from single-copy or double-copy sequences (see *Methods*). The divergence time between this pair of species is approximately 15% of the age of the WGD (Scannell *et al.*, 2006a). The non-synonymous substitution rate is significantly (19.83%) higher between orthologs that are members of duplicate pairs than between orthologs that are single-copy genes (Table 4.1). The former are also ~10% less constrained (as inferred from the *dN*/*dS* ratio) and we note that this effect is only observed if the biased retention of slowly-evolving sequences in duplicate identified by Davis and Petrov is corrected for (compare the '% Difference' in *dN*/*dS* values between columns labeled 'Column-matched' and 'Random sample'). Table 4.1 also confirms that the column-matching procedure operates by selecting a subset of columns from the super-alignment A1 that are more evolving slowly than average, but does not otherwise affect the data (compare the *dN* and *dS* values between single-copy loci for the columns labeled 'Column-matched' and 'Random sample'). Most importantly however, it is clear that the altered molecular evolution of duplicated gene pairs can persist for a very long period of time after the initial duplication event.

Table 4.1 Sequences derived from duplicate gene pairs ('Double-copy') have experienced an elevated nonsynonymous substitution rate and decreased selective constraint relative to single-copy sequences ('Single-copy') since the divergence of *S. cerevisiae* and *S. bayanus*. Codon super-alignments were obtained by back-translating the protein super-alignments used to create Figure 1B. The values shown are the averages of 100 pseudo-replicates. The site-matching procedure corrects for the Davis and Petrov effect and is described in the text.

	Random sample of columns from A1 and A2			**Site-matching procedure between A1 and A2**		
	Single-copy (A1)	**Double-copy (A2)**	**% Difference**	**Single-copy (A1')**	**Double-copy (A2')**	**% Difference**
dN	0.069	0.076	9.61 %	0.063	0.076	19.83 %
dS	1.119	1.250	11.76 %	1.131	1.250	10.46 %
dN / *dS*	0.062	0.061	-1.92 %	0.056	0.061	8.48 %

4.4.5 Double-copy sequences evolve asymmetrically at the protein sequence level

To examine the possibility of asymmetric rates of protein sequence evolution between members of duplicated pairs we performed maximum-likelihood branch-length evaluation individually on each of the 85 double-copy loci we collected. We designated the duplicate clades (*e.g.* the clades labeled 'Copy 1' and 'Copy 2' in Figure 4.1A, bottom left) as either "fast" or "slow" evolving based on the relative lengths of only the first branches after the WGD. We discarded ten loci where the difference between the lengths of these branches was negligible (see *Methods*) and then assembled a new super-alignment, A2$_{\text{asym}}$, from the remaining loci but being careful to concatenate all the "fast" clades together. Using A2$_{\text{asym}}$ and A1 we repeated all the steps performed to create Figure 4.1B except the final averaging step between the two duplicate clades (Figure 4.1A, bottom). Instead, we compared the lengths of branches in the "fast" and "slow" clades on the tree reconstructed from A2$_{\text{asym}}$ separately to the equivalent branches on the tree reconstructed from single-copy loci. Branches in the "fast" and "slow" clades exhibit radically different rates of protein sequence evolution (Figure 4.2). Indeed, on average, the "fast" *S. cerevisiae* copy has evolved at 150% of the rate of the "slow" copy. Importantly, although we treat these data as a measure of the asymmetry of protein sequence evolution, we note that the method by which we constructed A2$_{\text{asym}}$ (on the basis of the first branch after the WGD only) represents an implicit test of the hypothesis that duplicated sequences evolve asymmetrically. If this hypothesis is false, then no difference in the rate of protein sequence evolution between "fast" and "slow" clades should be observed on any branch other than the first branch after the WGD (which we forced to be asymmetric). In fact, the distinction is apparent on every branch (Figure 4.2) and the sum of the difference in rates between duplicate branches (excluding the first branches after the WGD) is much greater in this case than in any of 100 randomized datasets, suggesting a minimum significance of $P < 0.01$. A comparison of the number of substitutions observed on the terminal *S. cerevisiae* (or *S. bayanus*) branches to that expected assuming equal rates of protein sequence evolution supports this conclusion (χ^2 goodness-of-fit test, $P <$ 1e-10).

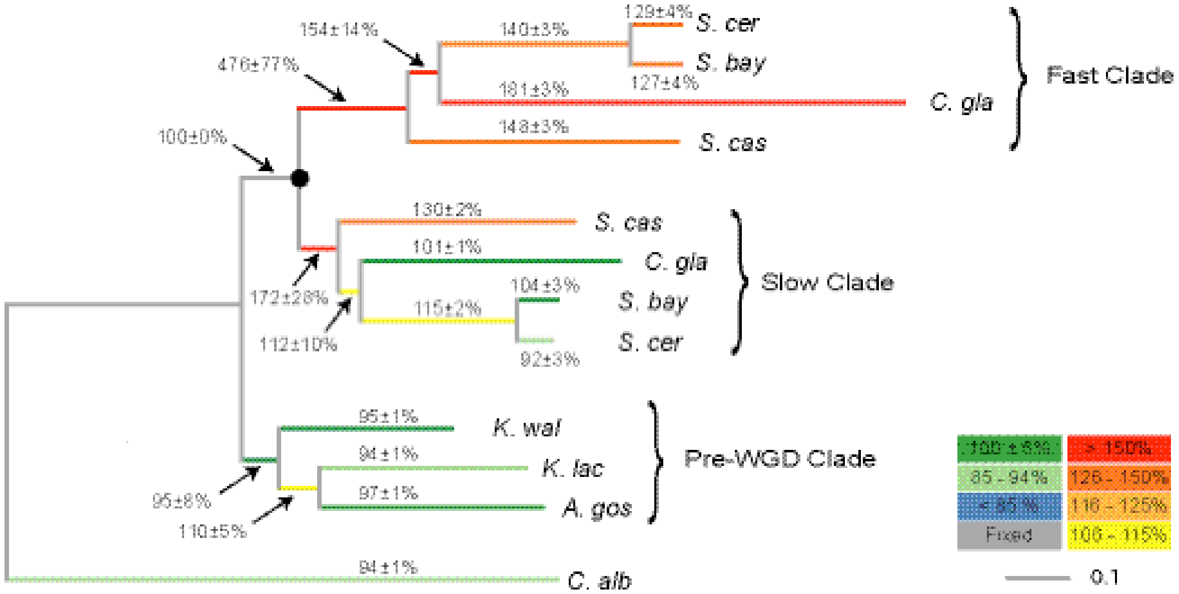


Figure 4.2 Asymmetric protein sequence evolution is initiated very soon after gene duplication and persists in modern duplicates. The tree was reconstructed from A2$_{asym}$ and shows branch lengths expressed as percentages of the length of the corresponding branches on a tree reconstructed from equivalent single-copy sequences (see text for details). Branch lengths are the averages of 100 pseudo-replicates and the coloring scheme is the same as in Figure 4.1.

Three features of Figure 4.2 are notable. First, the rate of protein sequence evolution on the first branch after the WGD is significantly greater than 100% in both the "fast" and "slow" clades. The rate on this branch in the "fast" clade is close to five times the expected (single-copy) rate. More surprisingly, its rate in the "slow" clade is 172±28%, almost twice the expected rate. This strongly suggests that both members of duplicated pairs experience a burst of protein sequence evolution after gene duplication. This result is unlikely to be an artifact of the method we used to estimate the rate on this branch (Figure 4.1A, top) because even if we assume that the WGD occurred immediately after the divergence of the non-WGD yeasts (*i.e.*, reducing the red branch in Figure 4.1A to zero, which minimizes the estimated increase in the rate of sequence evolution on the first branch after the WGD) we find that the first branch after the WGD in the slow clade is 120% of the length of the equivalent branch on T1'. In addition, the terminal *S. castellii* branch in the "slow" clade also shows significant acceleration (Figure 4.2; 130±2%).

Second, although both the "fast" and "slow" duplicate clades experience a rapid decline in the rate of protein sequence evolution (Figure 4.2), the levels to which they fall are very different. The terminal branches in the "fast" clade are still evolving much faster than expected (127–181%), but in the "slow" clade the rate increase attributable to the presence of a paralog has virtually disappeared on all branches after the divergence of *S. castellii* and it is possible that the slower-evolving members of gene pairs created in the WGD no longer experience an altered selective regime due to the presence of a duplicate sequence (see *Discussion*). Finally, we infer that the rapid emergence of "fast" and "slow" members of gene pairs represents a decisive and largely irreversible evolutionary change, because our partitioning of genes into "fast" and "slow" copies based on the rate on the first branch after WGD is a remarkably accurate predictor of the rates of evolution on all subsequent branches (Figure 4.2). In independent work, our laboratory has described this evolutionary pattern as "consistent asymmetry" and attributed it to early neofunctionalization of the faster copy (Byrne and Wolfe, manuscript submitted).

4.4.6 The pattern of amino acid substitution does not differ between double-copy and single-copy sequences

Because gene duplication is often associated with evolutionary innovation we considered the possibility that the mode as well as the tempo of protein sequence evolution may be affected by gene duplication. We therefore compared the pattern of amino acid substitutions occurring in T2' and T1' on three different branches (labeled X, Y and Z in Figure 4.1A, bottom right). We chose these branches because they are short (minimizing the number of sites that have sustained multiple substitutions), they have similar lengths (so results can be compared between branches), and because branches Y (immediately after the WGD) and X (the branch from the divergence of *S. bayanus* to modern *S. cerevisiae*) are of particular interest. We used maximum-likelihood to reconstruct internal nodes in the trees and inferred substitutions by parsimony. We classified substitutions on a spectrum from 'Conservative' to 'Radical' using the Universal Evolutionary Index (Tang *et al.*, 2004), an empirically derived index specifying the relative frequencies of amino acid changing single nucleotide substitutions (see *Methods*). We did not detect any difference in the proportions of substitutions of different types between equivalent branches on T2' and T1' (Table 4.2). We obtained similar results (data not shown) when substitutions were classified using the "Grantham Matrix" (Li *et al.*, 1985), which is based on physico-chemical properties of amino acids (Grantham, 1974). Because we have sufficient statistical power to detect even a small departure from expected values we conclude that gene duplication does not lead to a disproportionate increase in certain types of amino acid substitutions but results in a general increase in the rate of protein sequence evolution. This is consistent with recent results suggesting that neither positive selection (which is likely to have contributed to asymmetric evolution of duplicate pairs after the WGD (Fares *et al.*, 2006)) nor gene duplication *per se* are associated with altered patterns of amino acid substitution (Hanada *et al.*, 2006, Conant *et al.*, 2006).

Table 4.2 The pattern of amino acid substitution does not differ between sequences derived from duplicate gene pairs (from A2') and single-copy sequences (from A1') either prior to the WGD (branch X), immediately after the WGD (branch Y), or in modern sequences (branch Z). P-values were calculated using a χ^2 test of homogeneity.

		Number of amino acid substitutions of type				
		Conservative	**Moderately Conservative**	**Moderately Radical**	**Radical**	**P-value**
Non-WGD (branch X)	Single-copy	275	102	42	14	0.312
	Double-copy	314	121	69	22	
Post-WGD (branch Y)	Single-copy	419	203	153	48	0.240
	Double-copy	811	474	285	93	
Modern (branch Z)	Single-copy	578	302	201	56	0.337
	Double-copy	1255	699	435	156	

4.5 Discussion

Gene duplication is a hugely important process in genome evolution and is of interest for at least three reasons: the requirement for a redundant functional gene structure as a possible prerequisite for the evolution of novel functions (Ohno, 1970, Thomson *et al.*, 2005, but see Piatigorsky and Wistow, 1991, Hughes, 1994) the unexplained "excess" of duplicate genes in vertebrate genomes (Force *et al.*, 1999); and the contribution of gene duplication and subsequent reciprocal gene loss to the creation of new species (Scannell *et al.*, 2006a, Lynch and Force, 2000b). Although the relationships between the accumulation of sequence changes in duplicate genes and these processes are not well understood (Lynch and Katju, 2004), the altered molecular evolution of paralogs is a possible factor in all of them. In addition, preservation of gene pairs by either subfunctionalization (van Hoof, 2005) or neofunctionalization (Thomson *et al.*, 2005) ultimately ensures that the altered selective regime experienced by gene pairs will continue and opens up the possibility that it will contribute to secondary evolutionary changes over a much longer period than was required for the initial preservation of the pair (He and Zhang, 2005b). Thus, understanding how the molecular evolution of duplicate gene pairs differs from that of single-copy genes is critical for understanding genome evolution.

In order to estimate the rate of protein sequence evolution in different time intervals after WGD we compared the lengths of equivalent branches between trees drawn from double-copy and single-copy sequences (Figure 4.1). Because the time between speciation events is fixed, any differences in branch-lengths compared in this way must be due to differences in substitution rates. Moreover, provided no other systematic differences exist, any observed rate differences can be attributed to gene duplication. This approach is conceptually similar to that taken by (Halligan and Keightley, 2006), who compared the rate of nucleotide substitution between putatively neutrally evolving intronic sites and promoter regions and concluded that the rate of evolution in promoters is constrained by purifying selection. In our case we needed to identify a sample of single-copy sites that were under a level of constraint similar to that experienced by the double-copy sequences prior to the WGD. We showed empirically that matching columns between A1 and A2 that were following similar evolutionary trajectories in non-WGD yeasts is an effective way of doing this. Nevertheless, we note that the column-matching procedure could be improved in at least two ways. First, in this study we used sequences from the same three non-WGD yeasts to both pair columns between A1 and A2 and to subsequently evaluate the efficacy of the procedure (Appendix XVII). It would be desirable to be able to use different sets of taxa for these two purposes. More generally, the site-matching procedure as implemented in this work is an *ad hoc* approximation of a principled method. By matching columns with identical combinations of amino acids we sought to match columns that – on average – were evolving at similar rates in species that have not undergone the WGD. If genome sequences from more non-WGD species were available however, it should be possible to assign sites in A1 and A2 to rate classes by maximum-likelihood and then derive A1’ and A2’ by simply sampling the appropriate number of sites from each rate class. As more genomes become available, this strategy may be useful in other contexts too.

By splitting duplicate pairs into "fast" and "slow" evolving copies we have shown that the two members of duplicate pairs are on average under very different levels of constraint. Indeed, on the first branch after the WGD we estimate that the "slow" clade is evolving at almost twice the expected rate (172±28%) while the "fast" clade is evolving at more than twice this rate again (476±77%). An important question is how this rate asymmetry arises, but we currently favour the view that no specific explanation is required. In the event that a particular gene pair evolves asymmetrically the identities of the "fast" and "slow" copies may be determined stochastically. One member of the duplicate pair must eventually sustain a mutation that sets it on a new evolutionary course and the other duplicate by default becomes the "slow" copy. This does not contradict our observation that both members of duplicate pairs tend to experience an increase in the rate of protein sequence evolution because in this model the decisive substitution need not be the first one. However, this model does predict that prior to this event both duplicates should experience a similar rate of protein sequence evolution.

The observation that both the "fast" and "slow" duplicate clades experienced a burst of protein sequence evolution after the WGD (Figure 4.2) is the most striking result of this study. Previous work has typically focused either on identifying cases of asymmetric protein sequence evolution (Van de Peer *et al.*, 2001, Conant and Wagner, 2003, Zhang *et al.*, 2003, Brunet *et al.*, 2006) or on testing whether gene duplication leads to an increase in the rate of protein sequence evolution (Lynch and Conery, 2000, Nembaware *et al.*, 2002, Jordan *et al.*, 2004) and has not attempted to quantify the relative contributions of the two processes. In addition, it was frequently unclear whether observation of the latter effect was a consequence of failure to control for the former. As far as we are aware our results represent the first simultaneous demonstration that both an increase in the rate of protein sequence evolution (in both copies) and asymmetric protein sequence evolution are consequences of gene duplication. These data suggest that even if Ohno's model of evolution after gene duplication is substantially correct (a "slow" copy performs the ancestral function while a "fast" copy optimizes a novel function) it cannot explain the evolution of duplicate pairs in the immediate aftermath of duplication.

We note that there appears to be a discrepancy between our results and those of (Kellis *et al.*, 2004), who observed that just 17% of duplicate pairs created by the WGD evolved significantly faster than their *K. waltii* ortholog but that 95% of these exhibited asymmetric protein sequence evolution (using *K. waltii* as the sole outgroup species). It seems likely however that there is no true contradiction but that the limited number of available genome sequences (3) afforded insufficient resolution and prevented Kellis et al. from recognizing that both members of duplicate pairs may experience an elevated rate of protein sequence evolution. For instance, because the only post-WGD yeasts they considered were *S. cerevisiae* and the closely related yeast *S. bayanus* they were forced to consider the average rate of protein sequence evolution over a very long post-WGD branch (compare to Figure 4.2). Similarly, because *K. waltii* was the only available outgroup, they will have overestimated the length of the *K. waltii* branch and suffered from reduced power to detect an increase in the rate of protein sequence evolution on post-WGD branches.

When does the altered selective regime experienced by gene pairs end? Our comparison between *S. cerevisiae* and *S. bayanus* suggests that on average members of duplicate pairs created by the WGD are still ~10% less constrained than equivalent single-copy sequences (Table 4.1). In addition, although the separate analysis of rates of protein sequence evolution in the “fast” and “slow” duplicate clades indicates that the two members of duplicate pairs may have very different histories (Figure 4.2), both clades show a progressive decline in the estimated rates of protein sequence evolution and provide no compelling indication that this process has reached equilibrium (as would be indicated by successive branches showing similar rates of protein sequence evolution). Nevertheless, it appears possible that the sequences in the “slow” clade may have returned to the rate of evolution that prevailed prior to the WGD. This is suggested by the fact that the rate of protein sequence evolution on the terminal *S. cerevisiae* branch is similar to that in the non-WGD clade (Figure 4.2; 92±3% compared to a median of 95%), but the much higher rate of protein sequence evolution on the terminal *S. bayanus* branch indicates that this conclusion should be treated with some suspicion. Had our site-matching procedure fully corrected for the rate bias between single-copy and double-copy sequences, then a rate of protein sequence evolution of 100% (rather than 95% as applied above) would indicate the complete restoration of the ancestral level of constraint. Even by this more lenient measure however, it is clear that sequences in the “fast” duplicate clade are still evolving rapidly relative to their single-copy progenitors (Figure 4.2). However, based on the continuing decline in the rate of protein sequence evolution on the lineage from the WGD to *S. cerevisiae* it seems possible that sequences in the “fast” clade will eventually be restored to the ancestral rate of protein sequence evolution.

How do we account for the prolonged period of asymmetric protein sequence evolution after gene duplication? It seems unlikely that 100 Myr after the WGD these sequences are still optimizing novel functions as predicted by the Ohno theory of evolution after gene duplication (Ohno, 1970). In addition, the elevated rate of protein sequence evolution in these genes cannot be governed exclusively by factors such as expression level (Drummond *et al.*, 2006, Kim and Yi, 2006) because immediately after the WGD both duplicates should be expressed at the same level as the gene that existed just prior to the WGD but this is the period during which their rates of protein sequence evolution differ most form the expected rate. One model that can account for these observations is quantitative sub-functionalization (Lynch and Force, 2000a). Under this model a single ancestral function is partitioned between a pair of duplicate genes, such that both are necessary to supply the required function at a level sufficient to prevent loss of fitness. As discussed in Section 1.2.2.2 this is likely to be much more common after WGD than after other types of duplication event. Crucially, quantitative sub-functionalization is compatible with three of the main features of Figure 4.2. First, it predicts that the rate of evolution should be highest just after the WGD because either copy can accept many substitutions that would be forbidden to a single-copy gene. Second, it predicts that the rate of protein sequence evolution should decline as the ability of duplicates to perform the ancestral function is eroded by slightly deleterious mutations: every time one copy fixes a partial loss-of-fitness substitution the other copy is committed to supplying more of the required function and will consequently be able to accept fewer substitutions. Third, quantitative sub-functionalization is highly likely to result in an unequal division of labour between duplicates and the existence of “fast” and “slow” clades in Figure 4.2 may be a reflection of the existence of “major” and “minor” gene duplicates. Moreover, because in this model neither duplicate performs the entire ancestral we expect that even a long time after the WGD both duplicates will still be evolving slightly faster than control single-copy sequences. The fact that the “slow” copy has not declined completely to the expected rate of protein sequence evolution reflects the fact that the “minor” duplicate still performs some of the required function (Figure 4.2).

Finally, our comparison of substitution patterns between double-copy and single-copy loci in different time intervals both before and after the WGD (Table 4.2) indicates that although the rate of protein sequence evolution changes dramatically after gene duplication and may have long-lasting effects on the molecular evolution of duplicate genes, the relative proportions of different amino acid substitutions are not altered. This is consistent with previous work showing that highly conserved sites in proteins are more likely to differ between gene pairs (produced by duplication) than between orthologs (produced by speciation) but that there is no difference in the nature of the observed changes (Seoighe *et al.*, 2003). We conclude that an increase in the rate of protein sequence evolution due to the presence of a redundant gene copy is sufficient to explain the altered molecular evolution of duplicate pairs relative to single-copy sequences. In addition, the observation that highly conserved and presumably functionally important sites are substituted after gene duplication suggests that loss-of-function mutations may be important for the preservation of duplicates after WGD and supports the view that quantitative sub-functionalization may be involved. In either case, we propose that on average both members of gene pairs exhibit an initial burst of protein sequence evolution but that this gives way to a period of highly asymmetric evolution during which one copy evolves at almost the ancestral rate while the other continues to evolve rapidly for a very long period of time.

4.6 Methods

4.6.1 Generation of super-alignments

We used the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe, 2005) to identify loci at which both duplicates derived from the WGD have been retained in the four post-WGD species *S. cerevisiae, S. bayanus, C. glabrata* and *S. castellii* and for which the orthology/paralogy relationships between duplicates in different species are known (double-copy loci; Scannell *et al.*, 2006a). We also assembled a set of loci at which single-copy syntenic orthologs only have been retained in the same four post-WGD species (single-copy loci). We discarded any loci for which syntenic orthologs in the non-WGD species *K. waltii, K. lactis* and *A. gossypii* were unavailable in YGOB as well as any loci for which we could not identify an ortholog in *C. albicans* using the reciprocal-best-hit BLAST methodology between *C. albicans* and *K. lactis*. We also discarded any loci that code for ribosomal proteins. Coding sequences for all genes were obtained from the website of the consortium that sequenced the relevant genome (Kellis *et al.*, 2003, Dujon *et al.*, 2004, Dietrich *et al.*, 2004, Kellis *et al.*, 2004) except for *S. castellii* (Cliften *et al.*, 2003), which we have previously reannotated (Scannell *et al.*, 2006b) (wolfe.gen.tcd.ie/ygob/scas/), *S. cerevisiae* (www.yeastgenome.org) and *C. albicans* (www.candidagenome.org). We translated and aligned the sequences at each locus, removed gapped sites and discarded alignments shorter than 50 amino acid sites in length.

Finally, we generated super-alignments by concatenating all of the alignments derived from the remaining single-copy (808) or double-copy (85) loci as appropriate. The resulting super-alignments, which we refer to as A1 and A2, consist of 324,540 and 33,720 amino acid sites respectively.

4.6.2 Generation of pseudo-replicates and confidence estimates

In the main text we review a sampling procedure (which we have previously described; Scannell *et al.*, 2006a) to select sub-alignments from A1 and A2 (called A1' and A2') that we use for phylogenetic reconstruction (See *Phylogenetics* below). Unless otherwise stated however we always performed the sampling procedure 100 times and generated 100 pairs of pseudo-replicate super-alignments ([A1'$_1$, A2'$_1$]…[A1'$_{100}$ A2'$_{100}$]). All subsequent steps were then performed separately on each pseudo-replicate pair ([A1'$_n$, A2'$_n$]) and we report the average results with the associated standard deviations. Prior to generating each pseudo-replicate pair we also randomized the relationships between duplicate clades from different loci in A2. Consider two alignments each of which consists of two duplicate clades (DC1 and DC2), which in turn consist of four orthologs each. Because all the sequences in DC1 (or DC2) at each locus are orthologs, but there is no relationship between sequences in DC1 (or DC2) at different loci, it is possible to concatenate the sequences in DC1 from locus one with the sequences from either DC1 or DC2 from locus two.

4.6.3 Phylogenetics

We determined the topology of the species tree for the yeasts used in this study by removing the *C. glabrata* sequence from the super-alignment A1 (See *Generation of super-alignments* above) and generating 100 pseudo-replicates (30,000 sites each) from the remaining sequences. *C. glabrata* was omitted because although its phylogenetic relationship to the other species is known with certainty from gene loss and other data, phylogenetic inferences based on *C. glabrata* sequence data have been shown to be unreliable (Scannell *et al.*, 2006a). We then used the WAG+G(8)+I+F model (as implemented in Tree-Puzzle; Schmidt *et al.*, 2002) to determine the maximum-likelihood topology for each bootstrap replicate and obtained a consensus topology, which is supported by all 100 pseudo-replicates. Since this topology (modified to include *C. glabrata*; Fig. 1A) recapitulates the putative relationships between these yeast species (Scannell *et al.*, 2006a) it was imposed for all subsequent analyses: all parameters (branch-lengths, gamma rate classes, *etc.*) other than the topology were optimized for all trees derived from the super-alignments A1’ and A2’. The imposed topology was modified as necessary to accommodate the existence of duplicate gene pairs in A2’ and the WAG+G(8)+I+F model as implemented in Tree-Puzzle was used for all figures in the main text. The model WAG+G(8)+F (no invariant sites) was used for all Supplementary Figures because it significantly reduces the required computation time.

4.6.4 Saccharomyces kluyveri and Kluyveromyces polysporus analyses

We generated super-alignments including the non-WGD species *S. kluyveri* exactly as described above in *Generation of super-alignments* but with the additional requirement that orthologous *S. kluyveri* genes could be identified at each locus using the reciprocal-best-hit BLAST methodology between *S. kluyveri* and *K. lactis* proteins. We obtained 793 single-copy and 81 double-copy loci and the resulting super-alignments, $A1_{Sklu}$ and $A2_{Sklu}$, consist of 307,374 and 29,918 aligned sites respectively. The phylogenetic relationship between *S. kluyveri* and the other species was determined using the super-alignment $A1_{Sklu}$ and the procedure described above in *Phylogenetics*. For analyses involving *K. polysporus* we used 11 alignments of double-copy loci and 59 alignments of single-copy loci from (Scannell *et al.*, 2006b). These were concatenated (as described above in *Generation of super-alignments*) to produce, $A1_{Kpol}$ and $A2_{Kpol}$, which consist of 23,157 and 4,904 aligned sites respectively.

4.6.5 Calculation of synonymous and non-synonymous substitution rates

We calculated the average *dN* and *dS* between orthologous single-copy *S. cerevisiae* and *S. bayanus* sequences by removing all sequences other than those from *S. cerevisiae* and *S. bayanus* from A1 either before or after performing the site-matching procedure to correct for the over-representation of slow-evolving genes in double-copy (See *Main Text*). We then replaced each amino acid with the codon that encodes it and used yn00 in the PAML package to estimate synonymous and non-synonymous distances between the two nucleotide sequences. The procedure to estimate the average *dN* and *dS* between orthologous double-copy *S. cerevisiae* and *S. bayanus* sequences was identical except that duplicated sequences from each species were concatenated to produce a single pairwise nucleotide super-alignment (201,708 nucleotides in length; the nucleotide super-alignment derived from single-copy sequences is 100,854 nucleotides in length) prior to using yn00 to estimate synonymous and non-synonymous distances.

4.6.6 Partitioning duplicates into fast-evolving and slow-evolving copies

We performed maximum-likelihood branch-length estimation individually for each of the 85 double-copy loci in our dataset as described in the *Phylogenetics* section but with four, rather than eight, gamma rate classes. We then compared the lengths of the branches between the nodes corresponding to the gene duplication event and the divergence of *S. castellii* (i.e. the first branches after the WGD) and considered the longer branch to be at the base of the fast-evolving clade and the shorter branch to be at the base of the slow-evolving clade. If the difference between the lengths of these branches was less than 5% of the sum of their lengths the locus was discarded. We assembled a super-alignment, $A2_{asym}$, from the alignments of the remaining loci by ensuring that the fast-evolving clades were always concatenated together.

4.6.7 Comparison of substitution patterns between single-copy and double-copy sequences

We performed a joint reconstruction of the sequences at internal nodes of a randomly chosen pair of super-alignments A1' and A2' (See *Generation of pseudo-replicates and confidence estimates*) using Fastml and the model WAG+G(8) (Pupko *et al.*, 2002). We then used parsimony to infer the substitutions between nodes at which the marginal probability of the most likely amino acid was at least twice the probability of the next most likely one and not less than 0.25. Finally, we classified all substitutions as 'Conservative', 'Moderately Conservative', 'Moderately Radical' and 'Radical' using either the "Universal Evolutionary Index" (Tang *et al.*, 2004) or the "Grantham Matrix" (Grantham, 1974, Li *et al.*, 1985) and compared the proportions of substitutions of each type between equivalent branches on T1' and T2' using a χ^2 test of homogeneity.

4.7 Acknowledgements

We are grateful to Mark Johnston for allowing us to use unpublished *S. kluyveri* sequences and Aoife McLysaght, Cathal Seoighe and Laurent Duret for comments at an early point in this work. We thank the members of the Wolfe lab for helpful discussions. This project was supported by Science Foundation Ireland.

Chapter 5. Conclusions

The genomes of almost twenty different hemiascomycete yeasts have been sequenced (Wolfe, 2004) since the genome of *S. cerevisiae* was published a decade ago (Goffeau *et al.*, 1996). The number of available genomes and the remarkable conservation of gene order means that yeast comparative genomics can be used to address hypotheses about genome evolution that would be impossible in any other system. Indeed, the two aspects of gene duplication studied in this thesis – the altered molecular evolution of duplicate genes and the loss of duplicate genes – both involved comparisons between at least seven genomes and it seems appropriate to speak of the "awesome power of yeast comparative genomics". The combined results of these analyses also highlight the power of two other forces in molecular evolution: neutral processes and whole-genome duplication. The potential for passive gene loss to drive the creation of new species (Chapter 2 and Chapter 3) in the absence of adaptive evolution is a departure from conventional wisdom and suggests that a fresh look at the study of reproductive isolation may be necessary. In addition, in Chapter 4 I have argued that quantitative sub-functionalization may be important for the initial preservation of duplicate genes. Although additional data will be required to resolve this issue, it highlights the potential for neutral processes to result in genome expansion. In addition, and as discussed in the *Introduction*, the preservation of duplicate genes by any mechanism may be a platform for the subsequent evolution of novel functions. Indeed, because of the large numbers of duplicate genes created by whole-genome duplication the potential for significant evolutionary innovation cannot be ignored. Moreover, the demonstration in this thesis that whole-genome duplication may also result in the rapid emergence of multiple new species highlights the potential for whole-genome duplication to contribute simultaneously to the emergence of both new species and new gene functions. Indeed, given recent evidence that at least one and possibly two whole-genome duplications occurred at the base of vertebrates (McLysaght *et al.*, 2002, Dehal and Boore, 2005), Ohno may have been correct to argue that whole-genome duplication is a key factor in the emergence of eukaryotic complexity (Ohno, 1970).

Appendix I Biased representation of Gene Ontology terms in gene loss classes among *S. cerevisiae*, *C. glabrata* and *S. castellii*

Yeast Gene Ontology data was downloaded from Incyte's Yeast Proteome Database (www.incyte.com, April 2003) website. Tests were carried out for all possible pairs of gene loss class (Figure 2.2) and Biological Process and Molecular Function Gene Ontology terms using two-tailed Fisher's exact tests. Halves were counted for ancestral loci that are still duplicated in S. cerevisiae but where a GO term had been assigned to only one of the pair. Totals involving halves were rounded up and down and the less significant P-value used. P-values were corrected with the Benjamini and Hochberg False Discovery Rate Correction for Multiple Testing. Results are listed in descending order of P-value and colored by significance; orange for P<0.001, peach for P<0.01, yellow for P<0.05 and white for P<0.1.

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
0	(24 / 210)	11.43%	Kinase activity	(122.5 / 2723)	4.50%	0.0003019228
0	(18 / 210)	8.57%	Protein kinase activity	(88.5 / 2723)	3.25%	0.0009975473
0	(13 / 210)	6.19%	Protein serine/threonine kinase activity	(56 / 2723)	2.06%	0.001353231
0	(15 / 210)	7.14%	Response to drug	(71.5 / 2723)	2.63%	0.001570295
0	(25 / 210)	11.90%	Carbohydrate metabolism	(149.5 / 2723)	5.49%	0.001657241
0	(16 / 210)	7.62%	Transcription factor activity	(77.5 / 2723)	2.85%	0.001906076
0	(19.5 / 210)	9.29%	Phosphate metabolism	(119 / 2723)	4.37%	0.00734047
0	(9.5 / 210)	4.52%	Carbohydrate biosynthesis	(39.5 / 2723)	1.45%	0.008184801
0	(16.5 / 210)	7.86%	Phosphorylation	(93.5 / 2723)	3.43%	0.0081987
0	(5 / 210)	2.38%	Cyclin-dependent protein kinase, regulator activity	(14 / 2723)	0.51%	0.009870484
0	(22.5 / 210)	10.71%	Cellular morphogenesis	(149 / 2723)	5.47%	0.01024425
0	(15 / 210)	7.14%	Protein amino acid phosphorylation	(87 / 2723)	3.20%	0.01052995
0	(29 / 210)	13.81%	Transcription regulator activity	(216.5 / 2723)	7.95%	0.01166270
0	(7.5 / 210)	3.57%	Metal ion transport	(28 / 2723)	1.03%	0.01180234
0	(8.5 / 210)	4.05%	Protein localization	(36 / 2723)	1.32%	0.01306002
0	(18.5 / 210)	8.81%	Cell wall organization and biogenesis	(120 / 2723)	4.41%	0.01848046
0	(18 / 210)	8.57%	Signal transduction	(122.5 / 2723)	4.50%	0.02026004
0	(6.5 / 210)	3.10%	Glucose metabolism	(25.5 / 2723)	0.94%	0.02602050
0	(3 / 210)	1.43%	Structural constituent of cell wall	(7 / 2723)	0.26%	0.03140796
0	(5.5 / 210)	2.62%	Enzyme inhibitor activity	(19.5 / 2723)	0.72%	0.03178749
0	(5 / 210)	2.38%	Carbohydrate catabolism	(21 / 2723)	0.77%	0.03713372
0	(6.5 / 210)	3.10%	Kinase regulator activity	(28.5 / 2723)	1.05%	0.03892537
0	(6.5 / 210)	3.10%	Invasive growth	(29.5 / 2723)	1.08%	0.04403117

0	(38 / 210)	18.10%	Response to stress	(337.5 / 2723)	12.39%	0.04924421
0	(9.5 / 210)	4.52%	Transcriptional activator activity	(58 / 2723)	2.13%	0.06010529
0	(2 / 210)	0.95%	Glucosidase activity	(4 / 2723)	0.15%	0.06662392
0	(32.5 / 210)	15.48%	Regulation of transcription, DNA-dependent	(284 / 2723)	10.43%	0.07308453
0	(36 / 210)	17.14%	Transferase activity	(326.5 / 2723)	11.99%	0.07417447
0	(4 / 210)	1.90%	Amino acid transport	(19 / 2723)	0.70%	0.08377406
0	(3 / 210)	1.43%	DNA repair	(105 / 2723)	3.86%	0.08854010
0	(7.5 / 210)	3.57%	Transcriptional repressor activity	(42 / 2723)	1.54%	0.0899391
0	(2 / 210)	0.95%	Aerobic respiration	(88 / 2723)	3.23%	0.09435876
0	(9 / 210)	4.29%	Actin cytoskeleton organization and biogenesis	(60 / 2723)	2.20%	0.09918827

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
1A	(3 / 33)	9.09%	Lyase activity	(43.5 / 2723)	1.60%	0.02181516
1A	(2 / 33)	6.06%	Metal ion homeostasis	(32.5 / 2723)	1.19%	0.0733048
1A	(4 / 33)	12.12%	Cell wall organization and biogenesis	(120 / 2723)	4.41%	0.07559531
1A	(4.5 / 33)	13.64%	Signal transduction	(122.5 / 2723)	4.50%	0.08105871
1A	(2 / 33)	6.06%	Exocytosis	(37 / 2723)	1.36%	0.08905365
1A	(1 / 33)	3.03%	Channel/pore class transporter activity	(7 / 2723)	0.26%	0.0998860
1A	(1 / 33)	3.03%	Structural constituent of cell wall	(7 / 2723)	0.26%	0.0999378

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
1B	(4 / 18)	22.22%	Transcriptional activator activity	(58 / 2723)	2.13%	0.001180972
1B	(8 / 18)	44.44%	DNA binding activity	(256.5 / 2723)	9.42%	0.001258077
1B	(7 / 18)	38.89%	Transcription regulator activity	(216.5 / 2723)	7.95%	0.001899603
1B	(4 / 18)	22.22%	Transcription factor activity	(77.5 / 2723)	2.85%	0.003264378
1B	(2 / 18)	11.11%	Mating-type determination	(18 / 2723)	0.66%	0.008841317
1B	(3 / 18)	16.67%	RNA polymerase II transcription factor activity	(70.5 / 2723)	2.59%	0.01673971
1B	(4 / 18)	22.22%	Budding	(131.5 / 2723)	4.83%	0.01825018
1B	(5 / 18)	27.78%	Regulation of transcription from Pol II promoter	(204 / 2723)	7.49%	0.02031332
1B	(6 / 18)	33.33%	Transcription from Pol II promoter	(277.5 / 2723)	10.19%	0.02077900
1B	(8 / 18)	44.44%	Transcription, DNA-dependent	(435 / 2723)	15.98%	0.02115941
1B	(6 / 18)	33.33%	Regulation of transcription, DNA-dependent	(284 / 2723)	10.43%	0.02267404
1B	(3 / 18)	16.67%	Conjugation with cellular fusion	(92.5 / 2723)	3.40%	0.03289662
1B	(3 / 18)	16.67%	DNA recombination	(94.5 / 2723)	3.47%	0.034643
1B	(2 / 18)	11.11%	Mitotic recombination	(41.5 / 2723)	1.52%	0.03978037
1B	(2 / 18)	11.11%	Transcriptional repressor activity	(42 / 2723)	1.54%	0.03979037
1B	(6 / 18)	33.33%	Response to stress	(337.5 / 2723)	12.39%	0.04482505
1B	(12 / 18)	66.67%	Nucleobase, nucleoside, nucleotide and nucleic acid metabolism	(809 / 2723)	29.71%	0.04793574
1B	(2 / 18)	11.11%	RNA localization	(46.5 / 2723)	1.71%	0.0485210
1B	(1 / 18)	5.56%	Channel/pore class transporter activity	(7 / 2723)	0.26%	0.05599967

1B	(3 / 18)	16.67%	Establishment and/or maintenance of cell polarity (sensu Saccharomyces)	(120.5 / 2723)	4.43%	0.0618639
1B	(4 / 18)	22.22%	Nuclear organization and biogenesis	(198.5 / 2723)	7.29%	0.06261600
1B	(2 / 18)	11.11%	Response to DNA damage	(54.5 / 2723)	2.00%	0.06371482
1B	(2 / 18)	11.11%	Response to pheromone	(58 / 2723)	2.13%	0.06981687
1B	(1 / 18)	5.56%	DNA helicase activity	(10 / 2723)	0.37%	0.07687171
1B	(2 / 18)	11.11%	Intracellular signaling cascade	(61.5 / 2723)	2.26%	0.07831512
1B	(1 / 18)	5.56%	Flocculation	(10.5 / 2723)	0.39%	0.08388065
1B	(3 / 18)	16.67%	Establishment and/or maintenance of chromatin architecture	(137.5 / 2723)	5.05%	0.08402174
1B	(4 / 18)	22.22%	Cytoskeleton organization and biogenesis	(225.5 / 2723)	8.28%	0.08938588
1B	(1 / 18)	5.56%	TCA intermediate metabolism	(12 / 2723)	0.44%	0.09089485
1B	(1 / 18)	5.56%	Auxiliary transport protein activity	(13 / 2723)	0.48%	0.09783880

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
1C	(33 / 98)	33.67%	Structural constituent of ribosome	(99 / 2723)	3.64%	0.0000000000
1C	(34 / 98)	34.69%	Structural molecule activity	(171 / 2723)	6.28%	0.0000000000
1C	(35 / 98)	35.71%	Protein biosynthesis	(270 / 2723)	9.92%	0.0000000155
1C	(33.5 / 98)	34.18%	RNA binding activity	(298.5 / 2723)	10.96%	0.0000011609
1C	(0.5 / 98)	0.51%	Nucleotide binding activity	(195.5 / 2723)	7.18%	0.02042767
1C	(3 / 98)	3.06%	Protein serine/threonine phosphatase activity	(13.5 / 2723)	0.50%	0.02104194
1C	(0 / 98)	0.00%	ATPase activity	(130 / 2723)	4.77%	0.02156450
1C	(0 / 98)	0.00%	ATP binding activity	(136.5 / 2723)	5.01%	0.02289903
1C	(23 / 98)	23.47%	Molecular_function unknown	(1015 / 2723)	37.28%	0.04949949
1C	(3.5 / 98)	3.57%	Vitamin metabolism	(21 / 2723)	0.77%	0.05298168
1C	(2 / 98)	2.04%	Nucleobase metabolism	(10 / 2723)	0.37%	0.06735552
1C	(5 / 98)	5.10%	Protein serine/threonine kinase activity	(56 / 2723)	2.06%	0.06747097
1C	(7 / 98)	7.14%	Protein kinase activity	(88.5 / 2723)	3.25%	0.08749110
1C	(7 / 98)	7.14%	Phosphorylation	(93.5 / 2723)	3.43%	0.09701540

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
2A	(4 / 15)	26.67%	Amino acid biosynthesis	(74 / 2723)	2.72%	0.001544299
2A	(4 / 15)	26.67%	Amino acid metabolism	(117.5 / 2723)	4.32%	0.007481854
2A	(3 / 15)	20.00%	Cell ion homeostasis	(63.5 / 2723)	2.33%	0.008296315
2A	(3 / 15)	20.00%	Aerobic respiration	(88 / 2723)	3.23%	0.01879276
2A	(2 / 15)	13.33%	Metal ion homeostasis	(32.5 / 2723)	1.19%	0.01898122
2A	(3 / 15)	20.00%	Structural constituent of ribosome	(99 / 2723)	3.64%	0.02530976
2A	(4 / 15)	26.67%	Organic acid metabolism	(179.5 / 2723)	6.59%	0.02859496
2A	(3 / 15)	20.00%	Kinase activity	(122.5 / 2723)	4.50%	0.04307419
2A	(5 / 15)	33.33%	Protein modification	(301 / 2723)	11.05%	0.04481100
2A	(1 / 15)	6.67%	Amino acid transporter activity	(8 / 2723)	0.29%	0.05309427
2A	(1 / 15)	6.67%	Carbohydrate kinase activity	(8.5 / 2723)	0.31%	0.05891693
2A	(5 / 15)	33.33%	Transferase activity	(326.5 / 2723)	11.99%	0.058951
2A	(6 / 15)	40.00%	Transcription, DNA-dependent	(435 / 2723)	15.98%	0.06221322
2A	(2 / 15)	13.33%	Ion transport	(70.5 / 2723)	2.59%	0.07346630

2A	(2 / 15)	13.33%	Transcription from Pol I promoter	(76.5 / 2723)	2.81%	0.08454179
2A	(1 / 15)	6.67%	Hydro-lyase activity	(14 / 2723)	0.51%	0.08828624
2A	(3 / 15)	20.00%	Regulation of cell cycle	(167.5 / 2723)	6.15%	0.08951903
2A	(1 / 15)	6.67%	Molecular_function unknown	(1015 / 2723)	37.28%	0.09110131
2A	(4 / 15)	26.67%	Protein biosynthesis	(270 / 2723)	9.92%	0.09209951
2A	(3 / 15)	20.00%	Structural molecule activity	(171 / 2723)	6.28%	0.09319893
2A	(2 / 15)	13.33%	Ribosome biogenesis	(82 / 2723)	3.01%	0.09414585
2A	(0 / 15)	0.00%	Biological_process unknown	(654 / 2723)	24.02%	0.09694582

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
2B	(8 / 86)	9.30%	Transcriptional activator activity	(58 / 2723)	2.13%	0.001169329
2B	(8 / 86)	9.30%	Intracellular signaling cascade	(61.5 / 2723)	2.26%	0.001714328
2B	(10 / 86)	11.63%	Signal transduction	(122.5 / 2723)	4.50%	0.01043576
2B	(5 / 86)	5.81%	Membrane fusion	(38 / 2723)	1.40%	0.01106607
2B	(3 / 86)	3.49%	Small monomeric GTPase activity	(14 / 2723)	0.51%	0.01510777
2B	(9 / 86)	10.47%	Biological_process unknown	(654 / 2723)	24.02%	0.01655370
2B	(3 / 86)	3.49%	Tubulin binding activity	(17 / 2723)	0.62%	0.02376991
2B	(10 / 86)	11.63%	Protein catabolism	(142 / 2723)	5.21%	0.03014308
2B	(2 / 86)	2.33%	Carbohydrate transport	(7 / 2723)	0.26%	0.03081065
2B	(10 / 86)	11.63%	Carbohydrate metabolism	(149.5 / 2723)	5.49%	0.03729369
2B	(7 / 86)	8.14%	Conjugation with cellular fusion	(92.5 / 2723)	3.40%	0.03991940
2B	(12 / 86)	13.95%	Vesicle-mediated transport	(193 / 2723)	7.09%	0.04039246
2B	(6 / 86)	6.98%	Lipid biosynthesis	(75.5 / 2723)	2.77%	0.04579421
2B	(9 / 86)	10.47%	Budding	(131.5 / 2723)	4.83%	0.04586231
2B	(5 / 86)	5.81%	Response to pheromone	(58 / 2723)	2.13%	0.04904696
2B	(4 / 86)	4.65%	Transcriptional repressor activity	(42 / 2723)	1.54%	0.05708462
2B	(13 / 86)	15.12%	Cytoskeleton organization and biogenesis	(225.5 / 2723)	8.28%	0.05782238
2B	(8 / 86)	9.30%	Cell wall organization and biogenesis	(120 / 2723)	4.41%	0.06683538
2B	(22 / 86)	25.58%	Cell proliferation	(431.5 / 2723)	15.85%	0.0668969
2B	(8 / 86)	9.30%	Establishment and/or maintenance of cell polarity (sensu Saccharomyces)	(120.5 / 2723)	4.43%	0.06838692
2B	(4 / 86)	4.65%	Monosaccharide metabolism	(45.5 / 2723)	1.67%	0.07358698
2B	(4 / 86)	4.65%	Cytokinesis	(46 / 2723)	1.69%	0.07360578
2B	(0 / 86)	0.00%	DNA repair	(105 / 2723)	3.86%	0.07755396
2B	(10 / 86)	11.63%	Regulation of cell cycle	(167.5 / 2723)	6.15%	0.07872504
2B	(3 / 86)	3.49%	Heavy metal binding activity	(28.5 / 2723)	1.05%	0.08038587
2B	(4 / 86)	4.65%	Endocytosis	(49 / 2723)	1.80%	0.08758125
2B	(5 / 86)	5.81%	RNA polymerase II transcription factor activity	(70.5 / 2723)	2.59%	0.094573

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
2C	(2 / 12)	16.67%	Guanyl nucleotide binding activity	(57.5 / 2723)	2.11%	0.03584873
2C	(1 / 12)	8.33%	Small monomeric GTPase activity	(14 / 2723)	0.51%	0.07170893
2C	(1 / 12)	8.33%	Guanyl-nucleotide exchange factor activity	(13.5 / 2723)	0.50%	0.07172722
2C	(3 / 12)	25.00%	Nucleotide binding activity	(195.5 / 2723)	7.18%	0.08001242

2C	(2 / 12)	16.67%	Structural constituent of ribosome	(99 / 2723)	3.64%	0.0912453

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
2D	(2 / 38)	5.26%	Small monomeric GTPase activity	(14 / 2723)	0.51%	0.02175669
2D	(5.5 / 38)	14.47%	Oxidoreductase activity	(116 / 2723)	4.26%	0.03305999
2D	(5.5 / 38)	14.47%	Carbohydrate metabolism	(149.5 / 2723)	5.49%	0.07867121
2D	(2 / 38)	5.26%	Nucleotide metabolism	(33.5 / 2723)	1.23%	0.09796489

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
2E	(1 / 9)	11.11%	Amine/polyamine transporter activity	(6 / 2723)	0.22%	0.02584979
2E	(1 / 9)	11.11%	Flocculation	(10.5 / 2723)	0.39%	0.04424913
2E	(1 / 9)	11.11%	Cell adhesion	(15 / 2723)	0.55%	0.05893188
2E	(1 / 9)	11.11%	Mannosyltransferase activity	(17 / 2723)	0.62%	0.06622047

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
2F	(8 / 28)	28.57%	Nuclear organization and biogenesis	(198.5 / 2723)	7.29%	0.002764243
2F	(6 / 28)	21.43%	Chromatin modification	(118.5 / 2723)	4.35%	0.00296578
2F	(3 / 28)	10.71%	Chromatin binding activity	(23.5 / 2723)	0.86%	0.003087855
2F	(3 / 28)	10.71%	Glucose metabolism	(25.5 / 2723)	0.94%	0.003793825
2F	(2 / 28)	7.14%	Glycolysis	(7 / 2723)	0.26%	0.003946533
2F	(5 / 28)	17.86%	Chromatin modeling	(93 / 2723)	3.42%	0.00483044
2F	(6 / 28)	21.43%	Establishment and/or maintenance of chromatin architecture	(137.5 / 2723)	5.05%	0.005818135
2F	(2 / 28)	7.14%	Carbohydrate kinase activity	(8.5 / 2723)	0.31%	0.005947410
2F	(8 / 28)	28.57%	DNA binding activity	(256.5 / 2723)	9.42%	0.01110253
2F	(3 / 28)	10.71%	Monosaccharide metabolism	(45.5 / 2723)	1.67%	0.01627033
2F	(1 / 28)	3.57%	Permease activity	(1 / 2723)	0.04%	0.02132592
2F	(2 / 28)	7.14%	Carbohydrate catabolism	(21 / 2723)	0.77%	0.02537763
2F	(6 / 28)	21.43%	Regulation of transcription from Pol II promoter	(204 / 2723)	7.49%	0.03053449
2F	(7 / 28)	25.00%	Transcription from Pol II promoter	(277.5 / 2723)	10.19%	0.04164072
2F	(2 / 28)	7.14%	Heavy metal binding activity	(28.5 / 2723)	1.05%	0.04439343
2F	(2 / 28)	7.14%	Isomerase activity	(30 / 2723)	1.10%	0.04710518
2F	(2 / 28)	7.14%	Double-strand break repair	(31.5 / 2723)	1.16%	0.05261737
2F	(4 / 28)	14.29%	Mitosis	(124.5 / 2723)	4.57%	0.05417652
2F	(3 / 28)	10.71%	Chromatin silencing	(76 / 2723)	2.79%	0.05567092
2F	(7 / 28)	25.00%	Regulation of transcription, DNA-dependent	(284 / 2723)	10.43%	0.07651873
2F	(2 / 28)	7.14%	Carbohydrate biosynthesis	(39.5 / 2723)	1.45%	0.07706991
2F	(2 / 28)	7.14%	Mitotic recombination	(41.5 / 2723)	1.52%	0.08373153
2F	(1 / 28)	3.57%	Phosphatase regulator activity	(6.5 / 2723)	0.24%	0.08526738
2F	(4 / 28)	14.29%	Carbohydrate metabolism	(149.5 / 2723)	5.49%	0.09079350
2F	(1 / 28)	3.57%	Amino acid transporter activity	(8 / 2723)	0.29%	0.09590364

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
3A	(12 / 134)	8.96%	Transcription from Pol I	(76.5 / 2723)	2.81%	0.001217642

			promoter			
3A	(6 / 134)	4.48%	Ribonuclease activity	(26.5 / 2723)	0.97%	0.004375919
3A	(9 / 134)	6.72%	rRNA processing	(59.5 / 2723)	2.19%	0.005728802
3A	(6 / 134)	4.48%	tRNA binding activity	(31 / 2723)	1.14%	0.007794253
3A	(5 / 134)	3.73%	Structural constituent of cytoskeleton	(23 / 2723)	0.84%	0.00977319
3A	(4 / 134)	2.99%	Actin binding activity	(15.5 / 2723)	0.57%	0.01364448
3A	(26 / 134)	19.40%	RNA binding activity	(298.5 / 2723)	10.96%	0.01525617
3A	(4 / 134)	2.99%	RNA ligase activity	(21 / 2723)	0.77%	0.02973713
3A	(4 / 134)	2.99%	tRNA ligase activity	(21 / 2723)	0.77%	0.02974457
3A	(20 / 134)	14.93%	RNA processing	(228 / 2723)	8.37%	0.03165878
3A	(23 / 134)	17.16%	RNA metabolism	(278 / 2723)	10.21%	0.03598592
3A	(6 / 134)	4.48%	Regulation of transcription, DNA-dependent	(284 / 2723)	10.43%	0.03700997
3A	(27 / 134)	20.15%	Hydrolase activity	(339.5 / 2723)	12.47%	0.04151970
3A	(9 / 134)	6.72%	Ribosome biogenesis	(82 / 2723)	3.01%	0.04160543
3A	(6 / 134)	4.48%	Nuclease activity	(48.5 / 2723)	1.78%	0.04779065
3A	(5 / 134)	3.73%	Pre-mRNA splicing factor activity	(38 / 2723)	1.40%	0.05428013
3A	(3 / 134)	2.24%	snRNA binding activity	(16 / 2723)	0.59%	0.06082118
3A	(0 / 134)	0.00%	Response to drug	(71.5 / 2723)	2.63%	0.08091311
3A	(7 / 134)	5.22%	Peptidase activity	(62.5 / 2723)	2.30%	0.08333249
3A	(7 / 134)	5.22%	Ligase activity	(64 / 2723)	2.35%	0.08530303
3A	(3 / 134)	2.24%	DNA-directed RNA polymerase activity	(19 / 2723)	0.70%	0.08817042
3A	(5 / 134)	3.73%	Primary active transporter activity	(46 / 2723)	1.69%	0.0993295

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
3B	(2 / 33)	6.06%	Transcription co-repressor activity	(13.5 / 2723)	0.50%	0.01687550
3B	(3 / 33)	9.09%	Regulation of translation	(44.5 / 2723)	1.63%	0.02306279
3B	(3 / 33)	9.09%	Transcription cofactor activity	(44.5 / 2723)	1.63%	0.02306854
3B	(2 / 33)	6.06%	Hydrogen-transporting two-sector ATPase activity	(17 / 2723)	0.62%	0.02355504
3B	(2 / 33)	6.06%	Monovalent inorganic cation transporter activity	(17 / 2723)	0.62%	0.02356091
3B	(2 / 33)	6.06%	Hydrogen-/sodium-translocating ATPase activity	(17 / 2723)	0.62%	0.02356679
3B	(2 / 33)	6.06%	Hydrogen ion transporter activity	(17 / 2723)	0.62%	0.02357267
3B	(4 / 33)	12.12%	Ribosome biogenesis	(82 / 2723)	3.01%	0.02462681
3B	(2 / 33)	6.06%	Hydrogen transport	(23.5 / 2723)	0.86%	0.0425977
3B	(3 / 33)	9.09%	rRNA processing	(59.5 / 2723)	2.19%	0.04644672
3B	(2 / 33)	6.06%	Monovalent inorganic cation transport	(29.5 / 2723)	1.08%	0.06237397
3B	(1 / 33)	3.03%	Isocitrate dehydrogenase activity	(4 / 2723)	0.15%	0.0624698
3B	(2 / 33)	6.06%	Metal ion homeostasis	(32.5 / 2723)	1.19%	0.0732861
3B	(3 / 33)	9.09%	Transcription from Pol I promoter	(76.5 / 2723)	2.81%	0.08318391
3B	(1 / 33)	3.03%	Lipid transporter activity	(7 / 2723)	0.26%	0.0999119

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
3C	(8 / 52)	15.38%	rRNA processing	(59.5 / 2723)	2.19%	0.0000671770

3C	(8 / 52)	15.38%	Transcription from Pol I promoter	(76.5 / 2723)	2.81%	0.0003173696
3C	(8 / 52)	15.38%	Ribosome biogenesis	(82 / 2723)	3.01%	0.0004655319
3C	(3 / 52)	5.77%	tRNA ligase activity	(21 / 2723)	0.77%	0.01096918
3C	(3 / 52)	5.77%	RNA ligase activity	(21 / 2723)	0.77%	0.01097190
3C	(4 / 52)	7.69%	Lyase activity	(43.5 / 2723)	1.60%	0.01432652
3C	(5 / 52)	9.62%	Amino acid biosynthesis	(74 / 2723)	2.72%	0.01976364
3C	(10 / 52)	19.23%	RNA processing	(228 / 2723)	8.37%	0.02798703
3C	(3 / 52)	5.77%	tRNA binding activity	(31 / 2723)	1.14%	0.02830996
3C	(2 / 52)	3.85%	Guanyl-nucleotide exchange factor activity	(13.5 / 2723)	0.50%	0.03808268
3C	(4 / 52)	7.69%	Ligase activity	(64 / 2723)	2.35%	0.04464350
3C	(11 / 52)	21.15%	RNA metabolism	(278 / 2723)	10.21%	0.04657850
3C	(2 / 52)	3.85%	Exonuclease activity	(17 / 2723)	0.62%	0.05257846
3C	(3 / 52)	5.77%	Acyltransferase activity	(44 / 2723)	1.62%	0.06465933
3C	(2 / 52)	3.85%	Carbohydrate catabolism	(21 / 2723)	0.77%	0.07466374
3C	(2 / 52)	3.85%	Mitochondrial transport	(22 / 2723)	0.81%	0.08064727
3C	(5 / 52)	9.62%	Amino acid metabolism	(117.5 / 2723)	4.32%	0.09710884

Class	Ratio in Class	% Class	GO Term	Ratio in Genome	% Genome	Corrected P-value
4	(44 / 1957)	2.25%	Structural constituent of ribosome	(99 / 2723)	3.64%	0.009607444
4	(93 / 1957)	4.75%	Structural molecule activity	(171 / 2723)	6.28%	0.03517732
4	(82 / 1957)	4.19%	Carbohydrate metabolism	(149.5 / 2723)	5.49%	0.06726693

Appendix II Phylogenetic relationship among *S. castellii*, *S. cerevisiae* and *C. glabrata*

Phylogenetic trees reconstructed by a variety of methods from either single-copy orthologous loci (Class 4 in Figure 2.2) or double-copy loci (Class 0) consistently show *C. glabrata* branching off outside a clade containing *S. cerevisiae* and S. *castellii*. However, the large numbers of ancestral loci in Class 2B (relative to 2D and 2F) and in Class 3A (relative to 3B and 3C), suggest an alternative topology where *S. castellii* is an outgroup to *C. glabrata* and *S. cerevisiae.* The number of ancestral loci is not homogeneously distributed among Classes 2B, 2D and 2F, nor among Classes 3A, 3B and 3C (χ^2 tests; $P < 0.05$).

In order to determine which of these topologies is correct, we searched for genomic rearrangements that could be phylogenetically informative about the relationship among *S. castellii*, *S. cerevisiae* and *C. glabrata*. The YGOB engine was used to search for all instances of a chromosomal inversion that is present in one track in any two post-WGD species, but absent from the same track in the third post-WGD species and absent from the pre-WGD species. The resulting list of about 200 candidate sites was examined manually. We searched specifically for chromosomal inversions, but in examining the candidate regions we also noticed some other rearrangements (interchromosomal translocations) that are phylogenetically informative. In total, five rearrangements shared by *S. cerevisiae* and *C. glabrata* to the exclusion of *S. castellii* were found, as described in Figure 5.1 and Figure 5.2. No rearrangements supporting the alternative topology were identified, strongly suggesting that the novel phylogeny we propose is the correct one.

In addition, for loci in Class 3 (Figure 2.2) the topology of the gene tree does not depend on the species phylogeny and can be inferred directly from synteny information. We exploited this to examine the ability of various tree reconstruction methods to recover correct topologies and show that all the methods we employed tend to return a topology where *C. glabrata* diverges from the *S. cerevisiae* lineage before *S. castellii*, even when synteny information clearly shows an alternative topology to be correct. This suggests that a systematic bias (Phillips *et al.*, 2004) may be affecting tree reconstruction and that the trees placing *C. glabrata* as an outgroup to *S. cerevisiae* and *S. castellii* are unreliable.

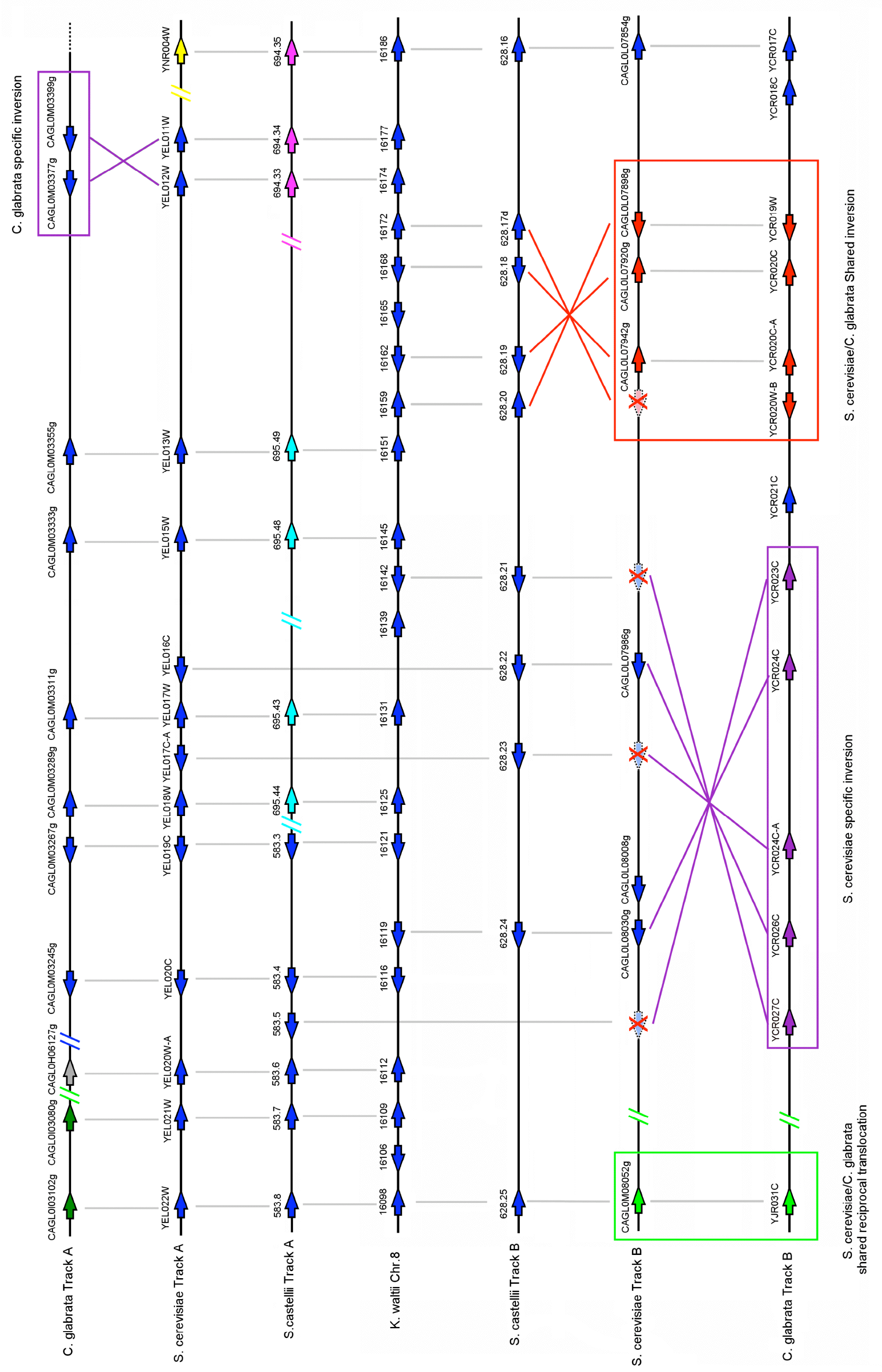


Figure 5.1 An inversion and a translocation suggesting *S. cerevisiae* and *C. glabrata* are more closely related to one another than either is to *S. castellii.* The inversion (boxed in red) is shared by Track B of *S. cerevisiae* and *C. glabrata* but not *S. castellii*. The gene order in *S. castellii* Track B is the same as in *K. waltii*, a representative pre-WGD species. The interchromosomal translocation (boxed in green) is shared by Track B of *S. cerevisiae* and *C. glabrata*, while Track B of *S. castellii* is colinear with the ancestral gene order.

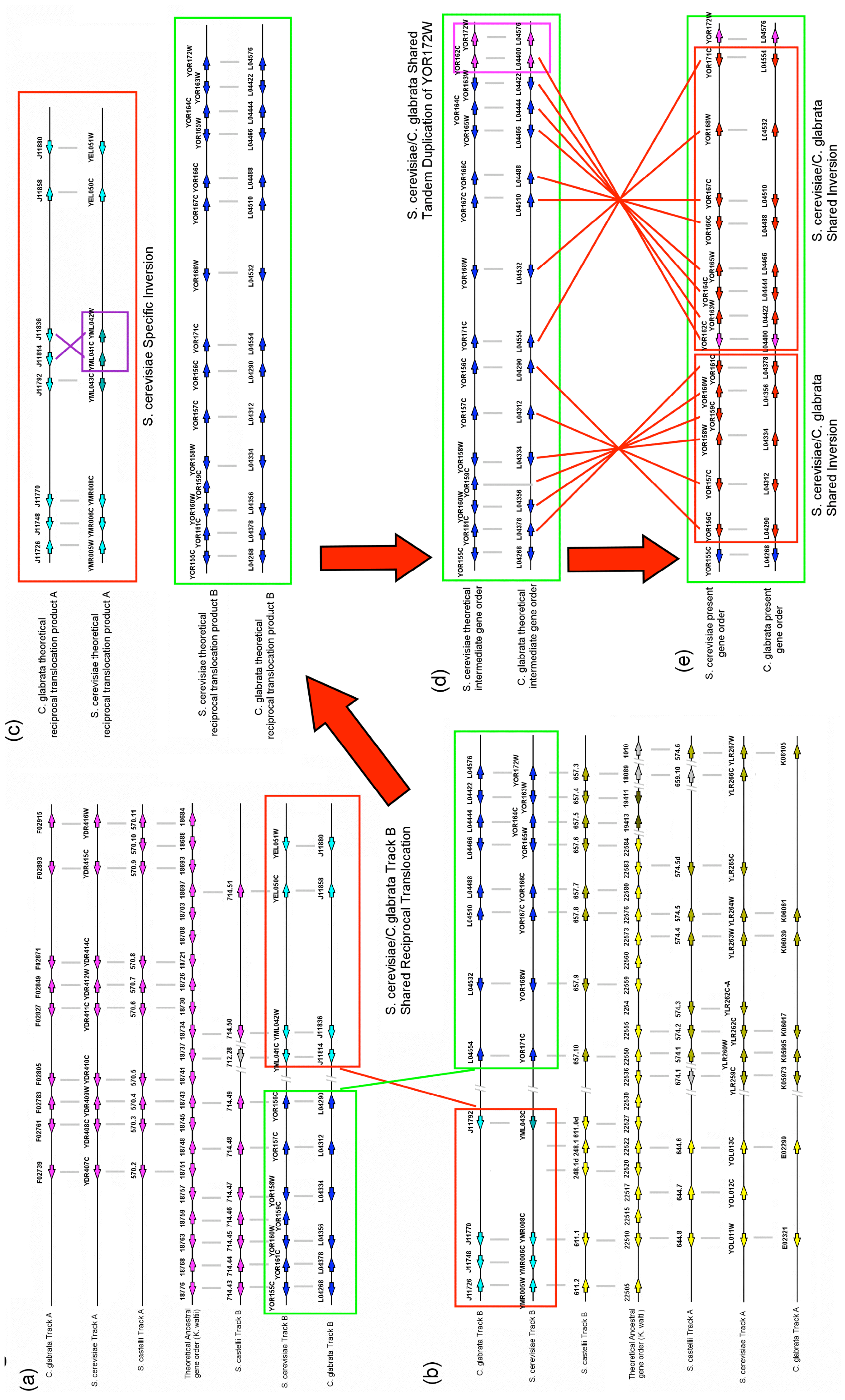

(a)
C. glabrata Track A
S. cerevisiae Track A
S. castellii Track A
Theoretical Ancestral gene order (K. waltii)
S. castellii Track B
S. cerevisiae Track B
C. glabrata Track B
(b)
S. cerevisiae/C.glabrata Track B Shared Reciprocal Translocation
(c)
C. glabrata theoretical reciprocal translocation product A
S. cerevisiae theoretical reciprocal translocation product A
S. cerevisiae Specific Inversion
S. cerevisiae theoretical reciprocal translocation product B
C. glabrata theoretical reciprocal translocation product B
(d)
S. cerevisiae theoretical intermediate gene order
C. glabrata theoretical intermediate gene order
S. cerevisiae/C. glabrata Shared Tandem Duplication of YOR172W
(e)
S. cerevisiae present gene order
C. glabrata present gene order
S. cerevisiae/C. glabrata Shared Inversion
S. cerevisiae/C. glabrata Shared Inversion

Figure 5.2 A reciprocal translocation and two inversions that suggest *S. cerevisiae* and *C. glabrata* are more closely related to one another than either is to *S. castellii* The large arrows in this Figure show the route by which the current genomic organizations are inferred to have arisen. A reciprocal translocation occurred between Track B of the region shown in (a), and Track B of the region shown in (b), in the common ancestor of *S. cerevisiae* and *C. glabrata*, after it had diverged from *S. castellii*. In region (a) the gene order in *S. castellii* is colinear with the ancestral order, represented by *K. waltii*. In region (b) there is a species-specific rearrangement in *S. castellii* Track B in the interval between genes 611.0d and 657.10, and there are also breaks in all three post-WGD species in Track A. However, the gene order seen in *K. waltii* through region (b) can be deduced to be ancestral because homologs of 611.0d and 657.10 are also linked in a *Z. rouxii* plasmid clone (data not shown). Two local chromosomal inversions that subsequently occurred in the green region are shown in (e). (A tandem duplication that produced the genes YOR172W and YOR162C, boxed in purple, may have been involved in these events).

Phylogenetic trees reconstructed from loci in Classes 3A, 3B, 3C.

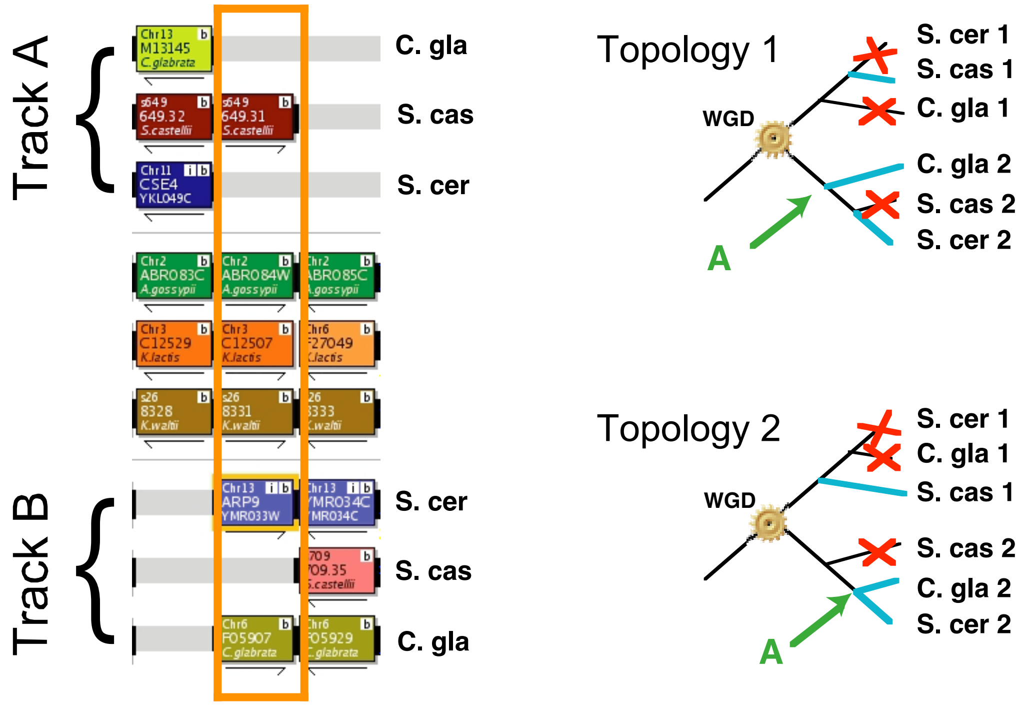


Figure 5.3 *S. cerevisiae* and *C. glabrata* have retained the same copy of the ancestrally duplicated ARP9 gene. Under either species topology they have a common ancestor at ‘A’ while their most recent common ancestor with the S. castellii copy is at the WGD. Therefore, irrespective of the species topology, the *S. cerevisiae* and *C. glabrata* genes should be more similar to each other than either is to the *S. castellii* homolog.

At loci in Class 3 we know the true topology of the gene tree, independent of the species tree, because it is shown by the high-quality synteny evidence (Figure 5.3). One of the remaining genes is a paralog of the other two, so it must be the outgroup. The only situation in which this assumption could be invalid is if one gene copy in a species overwrites the other by gene conversion, but it is unlikely that gene conversion could produce the systematically biased results we observe.

We drew trees for all Class 3 loci (3A, 3B and 3C) using *K. lactis* as an outgroup. For each locus we compared the topology indicated by synteny information (i.e., the tree we know to be the true tree because reliable synteny data shows one sequence to be a paralog of the other two) to the topology obtained from the sequence data using a variety of phylogenetic analysis methods (Table 5.1). For example, using the maximum-likelihood method, out of 28 loci in Class 3B (where the *S. cerevisiae* gene is a paralog of the *S. castellii* and *C. glabrata* genes, and so should appear as the outgroup), only 14 of the ML trees correctly recovered *S. cerevisiae* as the outgroup; 9 incorrectly identified *C. glabrata* as outgroup, and 5 incorrectly identified *S. castelli* as outgroup. Overall, for the ML method, 116 of the 208 gene trees were incorrect according to the synteny data, and in 70 of the 116 cases, the incorrectly proposed outgroup was *C. glabrata*. In the great majority of loci where a conflict is seen between the synteny and sequence trees, *C.glabrata* is the outgroup proposed by the sequence tree, regardless of phylogenetic method used. This suggests that a systematic bias is causing *C. glabrata* to branch too deeply in phylogenetic trees inferred from sequence data.

Table 5.1 Numbers in cells are the number of loci showing a particular combination of Synteny and Sequence topologies. Green highlighting indicates agreement between the Synteny and Sequence trees. Statistical significance was determined by performing a chi-square test of the hypothesis that the conflicting trees are uniformly distributed across the three types of Sequence tree.

		Synteny Outgroup					
Phylogenetic method	**Sequence Outgroup**	**S. castellii (Class 3A)**	**S. cerevisiae (Class 3B)**	**C. glabrata (Class 3C)**	**Total Sequence trees**	**Number of conflicts**	**P-value**
NJ	S. castellii	42	6	9	57	15	
	S. cerevisiae	22	12	10	44	32	
	C. glabrata	67	13	32	112	80	
	Class Totals	131	31	51	213	127	1.7E-12
NJ + 70% Bootstrap	S. castellii	23	2	7	32	9	
	S. cerevisiae	8	8	4	20	12	
	C. glabrata	48	6	23	77	54	
	Class Totals	79	16	34	129	75	1.0E-11
Parsimony	S. castellii	44	8	11	63	19	
	S. cerevisiae	16	10	4	30	20	
	C. glabrata	64	11	33	108	75	
	Class Totals	124	29	48	201	114	1.8E-12
Quartet Puzzling	S. castellii	38	5	12	55	17	
	S. cerevisiae	19	12	7	38	26	
	C. glabrata	48	11	26	85	59	
	Total in each	105	28	45	178	102	5.6E-07
Maximum Likelihood	S. castellii	49	5	11	65	16	
	S. cerevisiae	19	14	11	44	30	
	C. glabrata	61	9	29	99	70	
	Class Totals	129	28	51	208	116	1.0E-09

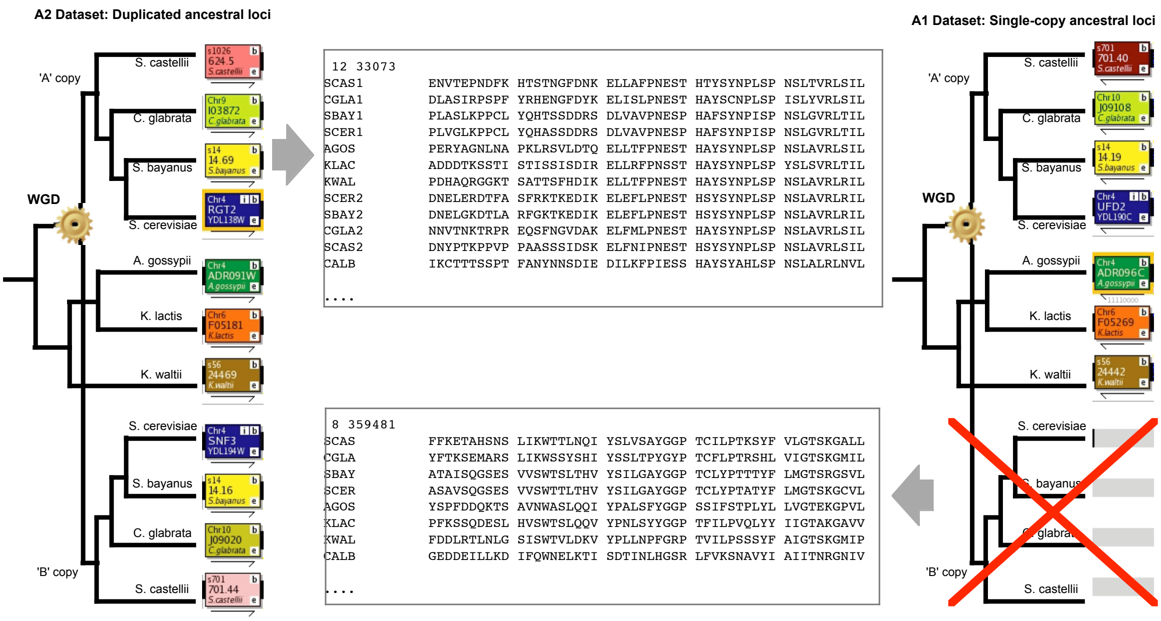
Appendix III Estimation of relative timing of speciation events between *S. cerevisiae*, *C. glabrata* and *S. castellii*

Phylogenetic trees drawn using ancestral loci at which single-copy syntenic orthologs have been retained in all post-WGD species (Class 4 in Figure 2.2 and Figure 5.4 at right), can be used to determine the relative timing of post-WGD speciation events. Ancestral loci that have retained duplicates (Class 0 in Figure 2.2 and Figure 5.4 left) are not suitable for this purpose as they may undergo a period of relaxed selection following duplication (Kondrashov *et al.*, 2002, Nembaware *et al.*, 2002), thus violating the assumptions of the molecular clock. They can be used however to estimate the time of divergence of duplicates created by WGD (at the common ancestor of the 'A' and 'B' copies; Figure 5.4 left).

This supplemental material describes a procedure to merge information from trees of duplicated and single-copy ancestral loci to produce a linear time-scale, on which 0 indicates the initial time of duplicate divergence and the timing of post-WGD speciation events are expressed as a proportions of the total time from duplicate divergence to *S. cerevisiae*.

Alignments of duplicated and single-copy syntenic ancestral loci

We used YGOB (http://wolfe.gen.tcd.ie/ygob/) to assemble sets of ancestral loci at which all post-WGD species had either retained two gene copies (Figure 5.4 left), or had retained the same syntenic copy (Figure 5.4 right). We discarded ribosomal proteins, ancestral loci at which one or more pre-WGD species possessed no ortholog and any ancestral loci for which no unambiguous *C. albicans* ortholog could be detected by reciprocal best BLAST hits with the *K. lactis* protein. The remaining 88 duplicated and 909 single-copy loci were aligned with ClustalW (default parameters), stripped of gapped columns and then merged to produce two super-alignments. The alignment of single-copy loci (referred to as A1 in this appendix) consists of 359,481 sites and the alignment of duplicated loci (A2) consists of 33,073 sites.



```
 12 33073
SCAS1        ENVTEPNDFK HTSTNGFDNK ELLAFPNEST HTYSYNPLSP NSLTVRLSIL
CGLA1        DLASIRPSPF YRHENGFDYK ELISLPNEST HAYSCNPLSP ISLYVRLSIL
SBAY1        PLASLKPPCL YQHTSSDDRS DLVAVPNESP HAFSYNPISP NSLGVRLTIL
SCER1        PLVGLKPPCL YQHASSDDRS DLVAVPNESP HAFSYNPISP NSLGVRLTIL
AGOS         PERYAGNLNA PKLRSVLDTQ ELLTFPNEST HAYSYNPLSP NSLAVRLSIL
KLAC         ADDDTKSSTI STISSISDIR ELLRFPNSST HAYSYNPLSP YSLSVRLTIL
KWAL         PDHAQRGGKT SATTSFHDIK ELLTFPNEST HAYSYNPLSP NSLAVRLRIL
SCER2        DNELERDTFA SFRKTKEDIK ELEFLPNEST HSYSYNPLSP NSLAVRLRIL
SBAY2        DNELGKDTLA RFGKTKEDIK ELEFLPNEST HSYSYNPLSP NSLAVRLRIL
CGLA2        NNVTNKTRPR EQSFNGVDAK ELFMLPNEST HAYSYNPLSP NSLAVRLSIL
SCAS2        DNYPTKPPVP PAASSSIDSK ELFNIPNEST HSYSYNPLSP NSLAVRLSIL
CALB         IKCTTTSSPT FANYNNSDIE DILKFPIESS HAYSYAHLSP NSLALRLNVL

....
```

```
 8 359481
SCAS         FFKETAHSNS LIKWTTLNQI YSLVSAYGGP TCILPTKSYF VLGTSKGALL
CGLA         YFTKSEMARS LIKWSSYSHI YSSLTPYGYP TCFLPTRSHL VIGTSKGMIL
SBAY         ATAISQGSES VVSWTSLTHV YSILGAYGGP TCLYPTTTYF LMGTSRGSVL
SCER         ASAVSQGSES VVSWTTLTHV YSILGAYGGP TCLYPTATYF LMGTSKGCVL
AGOS         YSPFDDQKTS AVNWASLQQI YPALSFYGGP SSIFSTPLYL LVGTEKGPVL
KLAC         PFKSSQDESL HVSWTSLQQV YPNLSYYGGP TFILPVQLYY IIGTAKGAVV
KWAL         FDDLRTLNLG SISWTVLDKV YPLLNPFGRP TVILPSSSYF AIGTSKGMIP
CALB         GEDDEILLKD IFQWNELKTI SDTINLHGSR LFVKSNAVYI AIITNRGNIV

....
```

Figure 5.4 Assembly of alignments of ancestral loci that are still duplicated in all post-WGD species (left) and ancestral that have retained single-copy syntenic orthologs in all post-WGD species (right) using YGOB. The tree topologies on which these alignments are later evaluated are also shown.

Residue matching to construct comparable alignments of columns from duplicated and unduplicated ancestral loci

In order to merge information from trees drawn from duplicated and single-copy loci, we derived two new alignments (A1' and A2') by selecting pairs of columns from A1 and A2 that share the same amino acids in the pre-WGD taxa *K. waltii, K. lactis* and *A. gossypii* (Figure 5.5). 71 columns of 33,073 in A2 (0.21%) could not be paired with columns in A1 and were excluded (Red columns in Figure 5.5). A1' and A2' are therefore exactly the same length and consist of sites that (with the exception of duplication in some taxa) have similar evolutionary trajectories. Because A1' and A2' are large (33,002 sites) stochastic errors due to the residue-matching procedure should be negligible and the unduplicated regions of trees drawn from these alignments should be almost identical.

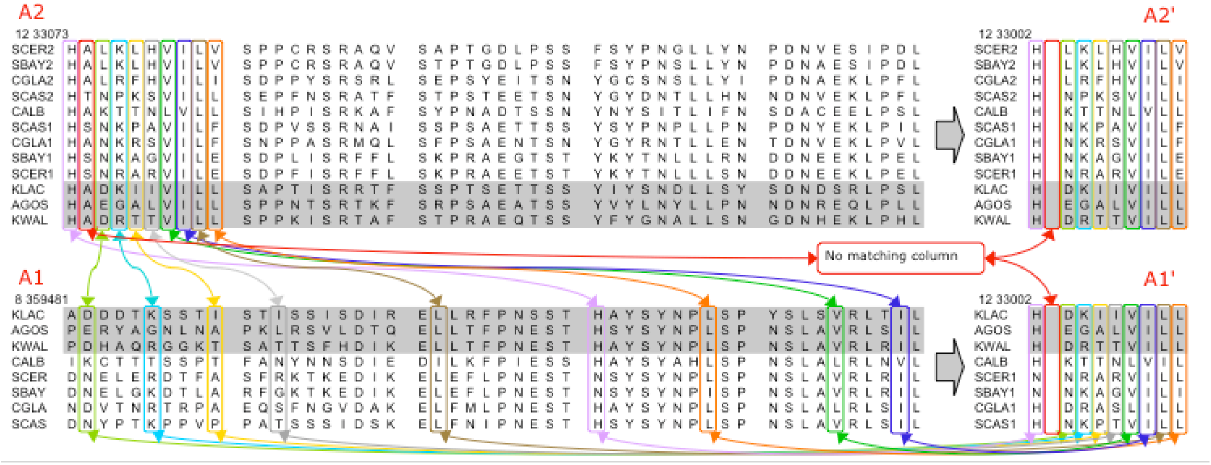
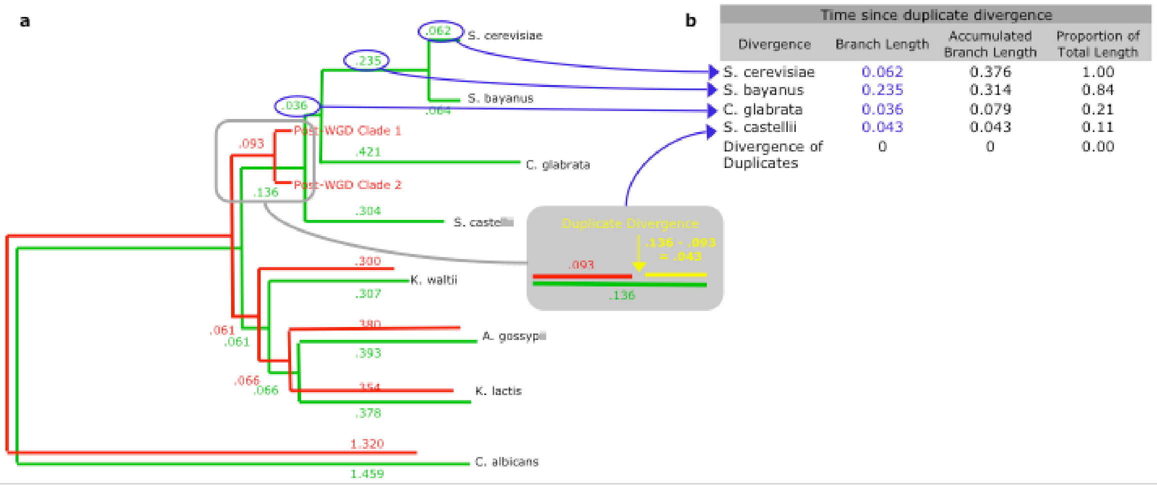


Figure 5.5 Fifty column example of the residue-matching procedure to construct comparable alignments (A1' and A2') of columns from ancestral loci that have been retained in duplicate in all post-WGD species and columns from ancestral loci that have retained single-copy syntenic orthologs in all post-WGD species. The taxa used for residue-matching are shaded in grey. The 10 boxed columns in A1 and A2 that are joined by arrows are examples of columns that have been "matched" between the two alignments. The column in A2 boxed in red could not be matched to a column in A1 (there is no 'AAA' in the 50 columns shown) and so has been omitted from the derived alignments A2' and A1'.

Timing of post-WGD speciation events since duplicate divergence

Maximum likelihood branch-length estimation was carried out for A1' (tree T1; green tree in Figure 5.6a) and A2' (tree T2; red tree in Figure 5.6a) under the topologies shown in Figure 5.4. As expected, unduplicated regions of T1 and T2 are very similar: In the pre-WGD clade T2 branches are on average 97.6% (range 93%-100%) of the length of the equivalent T1 branch, compared to 83.4% (range 79%-92%) for trees drawn from A1 and A2. The internal branches in the pre-WGD clade are exactly the same length.

Because the unduplicated regions of T1 and T2 are almost identical, we use the branch on T2 immediately prior to duplicate divergence to partition the branch on T1 between the divergence of the pre-WGD clade and the divergence of S. castellii into "pre-duplication" and "post-duplication" sections (Figure 5.6a, grey box). On this basis, the initial divergence of duplicates created by WGD occurred at a time equivalent to 4.3% amino acid divergence prior to the divergence of S. castellii. We use this figure, and the interspeciation branches on the post-WGD section of T1 (circled in blue), to estimate the relative timing of speciation events (Figure 5.6b).



Time since duplicate divergence			
Divergence	Branch Length	Accumulated Branch Length	Proportion of Total Length
S. cerevisiae	0.062	0.376	1.00
S. bayanus	0.235	0.314	0.84
C. glabrata	0.036	0.079	0.21
S. castellii	0.043	0.043	0.11
Divergence of Duplicates	0	0	0.00

Figure 5.6 (a) Maximum likelihood trees, T1 and T2, drawn using A1' (green) and A2' (red; duplicated clades have been omitted for clarity). Model selection was performed using ProtTest (Abascal *et al.*, 2005) and the model WAG + G + I + F was selected for all analyses. The gamma distribution was approximated with 8 rate classes. Trees were constrained to the topologies shown in Figure 5.4 and evaluated using Tree-Puzzle (Schmidt *et al.*, 2002). The topology of the post-WGD clade was determined as described in Appendix II. The topology and existence of the pre-WGD clade was inferred from additional trees drawn with A1 (data not shown). (b) Construction of a linear time-scale along the lineage from duplicate divergence to *S. cerevisiae*.

Confidence estimation for inferred speciation times

We calculated errors-bars for speciation time estimates by generating 100 bootstrap replicates of A2 and then performing the residue-matching procedure described above on each pseudo-replicate. Because there are 10 times more sites in A1 than A2, but only the number that can be paired are used, we are effectively also bootstrapping A1. The table below reports the mean and standard deviation for each of the branches on the lineage from duplicate divergence to *S. cerevisiae*. In all cases, the standard deviation is small but greater than the difference between the mean and real data, suggesting that our estimates are likely to be robust.

Table 5.2 Summary statistics for pseudo-replicates.

Divergence	% Time (Real Data)	% Time (Mean of Bootstraps)	Standard Deviation
S. cerevisiae	1.00	1.00	0.00
S. bayanus	0.84	0.84	0.01
C. glabrata	0.21	0.21	0.02
S. castellii	0.11	0.11	0.02
Duplicate Divergence	0.00	0.00	0.00

Appendix IV Model-based estimation of the number of genes still duplicated at phylogenetic nodes

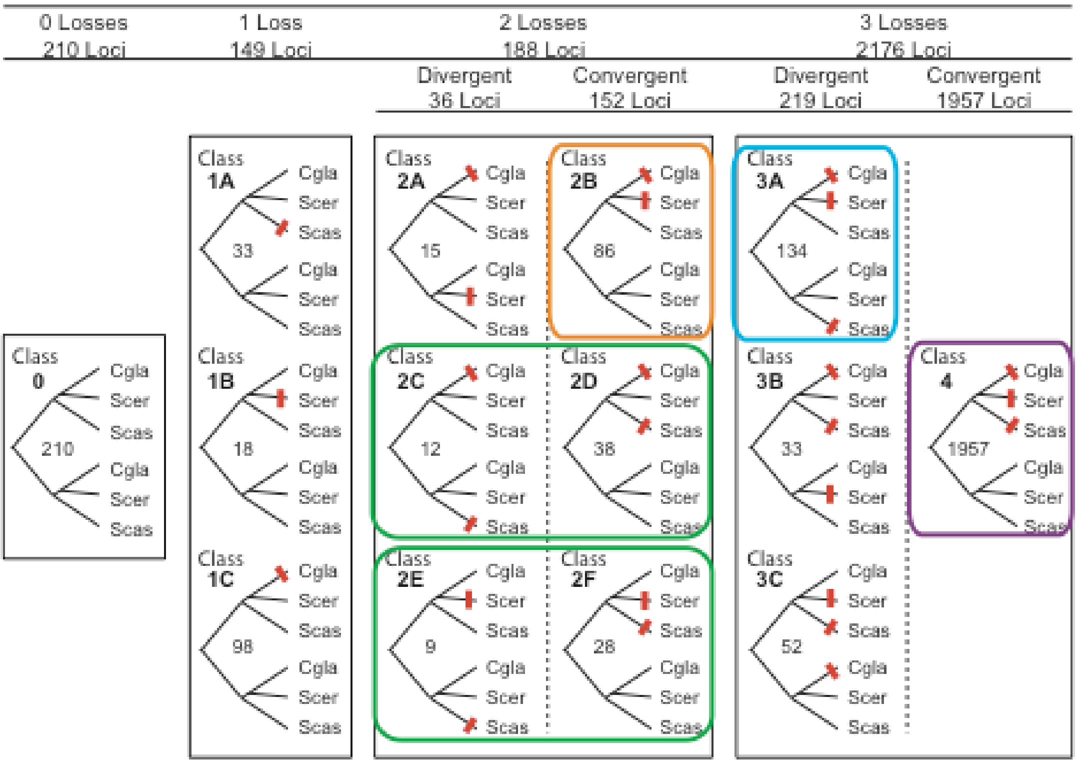


Figure 5.7 Modified version of Figure 2.2 with certain (pairs of) classes highlighted. Green: Double loss classes where the two gene losses must have been independent. Orange: Double loss class where some losses may have occurred on a branch shared by two species (i.e., losses in the common ancestor of *S. cerevisiae* and *C. glabrata*). Blue: Triple loss class where some losses may have occurred on a branch shared by two species. Purple: Triple loss class where some losses may have occurred on a branch shared by two species, and some losses may have occurred on a branch shared by three species.

Estimation of the proportion of convergent losses attributable to selection

If *S. castellii* diverged from the lineage leading to *S. cerevisiae* before *C. glabrata* all the loss classes highlighted in green (Figure 5.7) must have arisen by multiple independent losses. If this is the case, and for all losses the choice of which copy becomes lost is random, we would expect equal frequencies of 2C and 2D and also equal frequencies of 2E and 2F. This is not observed however (P < .05 in both cases) suggesting that selection favored a particular copy. The proportion of ancestrally duplicated loci that are resolved either under selection or neutrally can be estimated from the frequencies of either 2C and 2D, or 2E and 2F (Table 5.3). We use ϕ to denote the proportion of duplicated loci that are resolved neutrally.

Table 5.3 Estimates of the proportion of ancestrally duplicated loci that were resolved neutrally. See the 'Equations' section below for formulae and derivation.

	Class 2 Divergent losses	**Class 2 Convergent losses**	**ϕ**
2C:2D	12	38	.480
2E:2F	9	28	.486

Estimation of the number of apparent double losses that occurred on a shared branch

Some of the losses in Class 2B (orange in Figure 5.7) may be attributable to single losses on the shared branch leading to *S. cerevisiae* and *C. glabrata*. We can estimate the number of these by subtracting the number of convergent losses that we expect to find in Class 2B if all losses are independent, from the observed total of Classes 2A and 2B. From equation XI (in 'Equations', below) we therefore expect that of the 86 losses in Class 2B, 38.9 occurred on the shared branch leading to *S. cerevisiae* and *C. glabrata* and the remaining 47.1 were due to convergent losses after the speciation. This is calculated using equation XI as $SB_2 = 38.9 = (86+15) – 2*15/\phi$, where ϕ is estimated to be 0.483 by comparing Class 2C to 2D, and 2E to 2F (Table 5.3).

Estimation of the number of apparent triple losses that occurred on a shared branch either before the first speciation or before the second speciation

The process for partitioning apparent triple losses into those that occurred immediately after WGD (Speciation 0), after the first speciation (Speciation 1) or after the second speciation (Speciation 2) is identical to that just described for double losses. It is outlined in Figure 5.8 below.

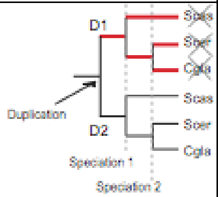
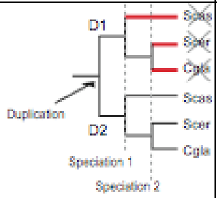
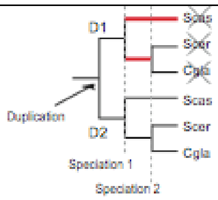
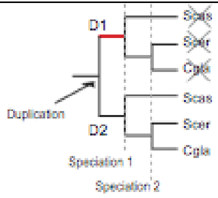
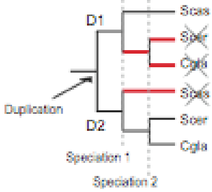
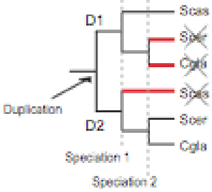
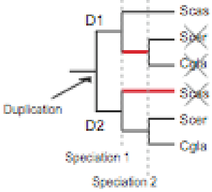
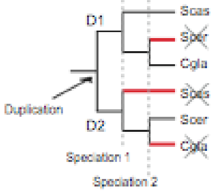
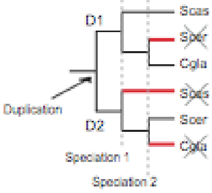
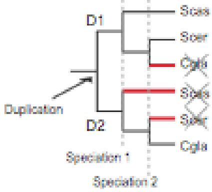
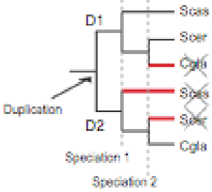
Class	Topology		3 Losses		2 Losses		1 Loss
4	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2	=	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2	+	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2	+	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2
3A	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2	=	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2	+	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2		
3B	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2	=	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2				
3C	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2	=	 D1 Scas Scer Cgla Scas Scer Cgla Duplication D2 Speciation 1 Speciation 2				
Neutral Estimator			Average(3B, 3C) * 4		(3A - Average(3B, 3C)) * 2		
Neutral loci			170		183		
Selected Estimator			Neutral*(1 - ϕ) / ϕ		Neutral*(1 - ϕ) / ϕ		
Selected loci			182		196		
Total loci	2176		352		379		1445
Timing (After)			Speciation 2		Speciation 1		Speciation 0

Figure 5.8 Assigning convergent losses from triple loss classes (3A, 3B, 3C and 4) to time periods delimited by speciation events.

Assumptions

1) We assume that selection on copy number (whether due to dosage, neofunctionalization or subfunctionalization) and selective differences between duplicates are independent. We ignore the former.

2) Selective differences between duplicate pairs we treat as either negligible (duplicates are functionally indistinguishable; $\Delta s_{duplicates} = 0$), in which case alternative copies may be retained in different lineages, or absolute (one of the duplicates is 'superior' to the other in all lineages; $\Delta s_{duplicates} = 1$), in which case a particular copy may be lost repeatedly in independent lineages. In the former case we consider duplicates to be resolved neutrally (N in Table 5.5 below) and in the latter case to be resolved under the influence of selection (S in Table 5.5 below).

3) We assume that ϕ, the fraction of duplicate pairs for which $\Delta s_{duplicates} = 0$, is a constant.

4) We classify the pattern of loss at loci where two or more losses have occurred as convergent if all single-copy lineages have retained the same syntenic copy. If alternative copies have been retained in different lineages the pattern of loss is considered to be divergent.

Duplicate Resolution

Under the assumptions above, the total number of loci in each loss class (defined by number of losses: 0-3) is fixed, but the frequencies of subclasses may be distorted due to preferential retention of one or other copy ($\Delta s_{duplicates} = 1$). This will be observed as an excess of convergent losses over divergent losses: Compare Classes 2A and 2B, 2C and 2D, or 2E and 2F (Table 5.4).

Table 5.4 Gene loss classes, their component classes, and paired divergent/convergent subclasses.

Gene Loss Class	Total	Component classes	Divergent/Convergent Pairs
0 (no losses)	210	0	n/a
1 (one loss)	149	1A, 1B, 1C	n/a
2 (double losses)	188	2A, 2B, 2C, 2D, 2E, 2F	(2A, 2B), (2C, 2D), (2E, 2F)
3 (triple losses)	2176	3A, 3B, 3C, 4	(3A+3B+3C, 4)

Also, under the assumptions above, different paralogs may not be selectively favored in different lineages. All incidences of divergent resolution must therefore be due to neutral loss of alternative copies and SD in Table 5.5 must always be 0.

Table 5.5 Duplicate resolution and pattern of loss. Note: Convergent and divergent losses are observed. Neutral resolution and resolution under selection must be inferred.

Pattern of loss	**Resolution**	
	Neutral (N)	**Selection (S)**
Convergent (C)	NC	SC
Divergent (D)	ND	SD (=0)

If no losses occurred on shared branches, then (where subscripts denote loss class and *n* refers to any class):

$NC_n+ND_n+SC_n+SD_n = Total_n$ *from model (assumptions 1,2,4)* Eqn 0

$(NC_n+ND_n)/Total_n = \phi$ *by definition (assumption 3)* Eqn 1

$D_n = ND_n$ *since $SD_n = 0$ for all n* Eqn 2

$ND_2/NC_2 = 1$ *see Figure 5.9* Eqn 3

$ND_3/NC_3 = 3$ *see Figure 5.10* Eqn 4

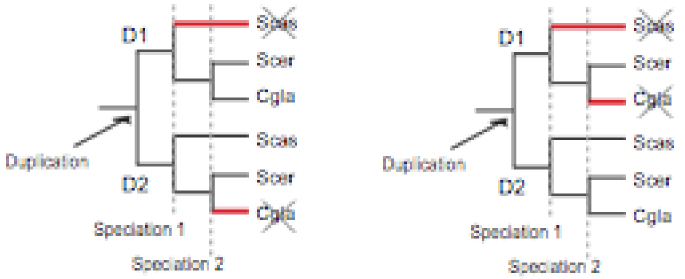


Figure 5.9 Classes 2C and 2D. These outcomes and the other pairs of convergent/divergent losses in Table 5.4 are equally likely if two random losses occur (assuming that both copies of a gene may not be lost and that there are no shared branches).

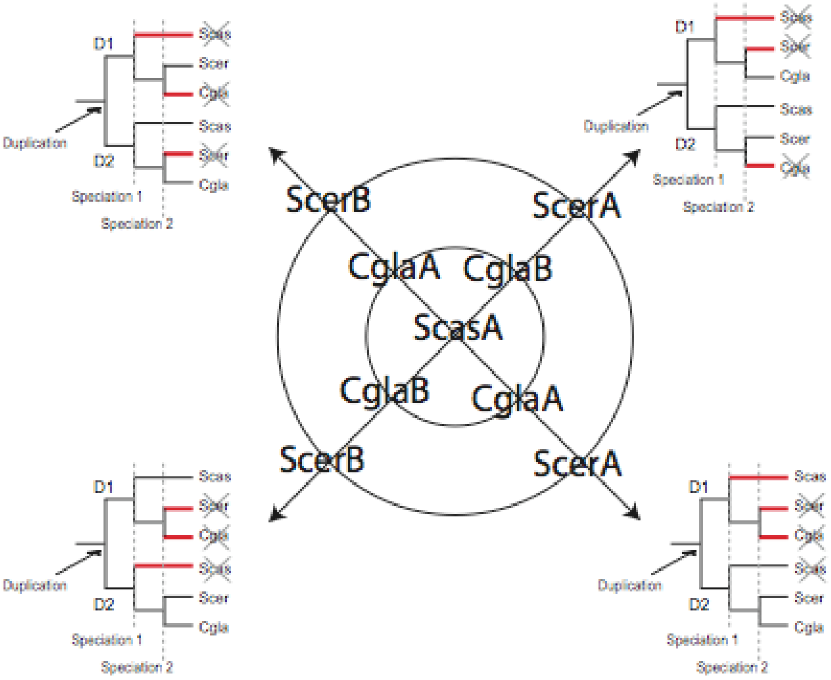


Figure 5.10 Four outcomes are equally likely if three random losses occur (assuming that both copies of a gene may not be lost and that there are no shared branches). These correspond to Classes 3A, 3B, 3C and 4.

Equations

1. ϕ from pairs of convergent/divergent double loss loci (e.g., 2E and 2F)

$(NC_n+ND_n)/(NC_n+ND_n+SC_n+SD_n) = \phi$ *from Eqns 0,1*

$(NC_2+ND_2)/(NC_2+ND_2+SC_2+SD_2) = \phi$ *for class 2 loci* I

$D_2 = ND_2 = NC_2$ *from Eqn 2 and Eqn 3* II

$2*D_2 / (SC_2 + 2*D_2) = \phi$ *from I and II* III

$SC_2 = C_2 - D_2$ *from $C_2 = SC_2 + NC_2$ and Eqn II* IV

$\phi = 2*D_2 / (C_2 + D_2)$ *from III and IV* V

ϕ in terms of observed classes

2. Selected convergent losses from ϕ and the number of neutral divergent losses

(a) Double Loss Loci

$D_2 = ND_2 = NC_2$	*from Eqn 2 and Eqn 3*	*VI*
$SC_2 = 2*ND_2*(1 - \phi) / \phi$	*from I and II*	*VII*

(b) Triple Loss Loci

$D_3 = ND_3 = 3*NC_3$	*from Eqn 2 and Eqn 4*	*VIII*
$SC_3 = (4/3)*ND_3*(1 - \phi) / \phi$	*from I and VIII*	*IX*

3. Shared branch (SB) losses for double loss loci, assuming *S. castellii* to be the outgroup

$NC_2+ND_2+SC_2+SD_2 + SB_2 = \text{Total}_{2A+2B}$	*Eqn 0 modified*	*X*
$SC_2 = 2*ND_2*(1 - \phi) / \phi$	*from VII*	
$SB_2 = \text{Total}_{2A+2B} -$	*from X*	
$D_2 -$	*from II*	
$D_2 -$	*from II*	
$[2*D_2*(1 - \phi) / \phi]$	*from VII and II*	
$SB_2 = \text{Total}_{2A+2B} - 2*D_2 / \phi$	*SB_2 in terms of observed classes*	*XI*

Appendix V Overrepresentation of slowly evolving genes and genes involved in highly conserved biological processes among genes that underwent reciprocal gene loss

In Table 5.6 we show that genes involved in various aspects of RNA metabolism, especially ribosome biogenesis and maturation are over-represented amongst RGL loci. We also show that genes that underwent RGL are slower evolving on average than those in other gene loss classes (Table 5.7). Finally, we use partial correlations to show that these effects are independent (Table 5.8).

Overrepresentation of genes involved in RNA related processes among loci that underwent RGL.

Overrepresentation of genes implicated in RNA related processes among RGL loci was assessed by Fisher's exact tests against control sets of genes. RGL for snoRNA genes was defined exactly as described for protein coding genes and determined by searching genomic DNA with *S. cerevisiae* snoRNA gene sequences downloaded from ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/rna (May 2004). For the snoRNA comparison the control was the proportion of RGL among snoRNA genes versus the proportion of RGL among protein coding genes in Classes 3 and 4 (i.e., 12/52 versus 219/2176). For all other annotations in Table 5.6, Class 3 loci (none of which have undergone RGL) were compared to Class 4 loci (all of which have undergone RGL) as defined in Figure 2.2.

Table 5.6 Overrepresentation of genes involved in RNA related processes among loci that underwent RGL.

Annotation	**RGL loci**			**Non-RGL loci**			**Significance**
	With Annotation	Total	Proportion	With Annotation	Total	Proportion	
YPD: All RNA Terms[§]	40	173	23%	215	1438	15%	5×10^{-3}
RNA binding complexes[†]	36	219	16%	142	1956	7%	2×10^{-5}
Nucleolar Localization[¶]	22	170	13%	57	1451	4%	7×10^{-6}
snoRNA genes[¥]	12	-	-	40	-	-	3×10^{-4}

§ Genes annotated by the Yeast Proteome Database whose annotations contain the term 'RNA', compared to all others except those annotated as 'Biological process unknown'.

† Membership of an RNA-binding complex as determined by (Krogan *et al.*, 2004), compared to genes that were not found to be members of an RNA-binding complex.

¶ Nucleolus-localized proteins as determined by (Huh *et al.*, 2003) compared to all other proteins with a known alternative subcellular localization.

¥ The proportion of RGL loci among snoRNA genes was compared to the proportion of RGL among protein coding genes in Classes 3 and 4 as described above.

Evidence that slowly evolving genes are overrepresented among RGL loci.

Representative K_A values (Davis and Petrov, 2004) were calculated for each ancestral locus using ungapped alignments *K. lactis* and *A. gossypii* orthologs and yn00 in the PAML package. Each ancestral locus is represented once regardless of whether or not it is still duplicated. As can be seen from Table 5.7 and Figure 5.11 the representative K_A of Class 3 (RGL) loci is on average less than that of genes from different gene loss classes.

Table 5.7 Proportion of loci in *K. lactis vs A. gossypii* K_A bins accounted for by gene loss classes.

Gene Loss	K_A Bin				Median K_A
Class	0.0 - 0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	for class
0	0.059	0.074	0.092	0.081	0.471
1	0.147	0.042	0.038	0.045	0.358
2	0.106	0.071	0.065	0.054	0.412
3	0.180	0.098	0.061	0.045	0.331
4	0.507	0.714	0.744	0.775	0.477
Number of Loci	339	742	739	445	

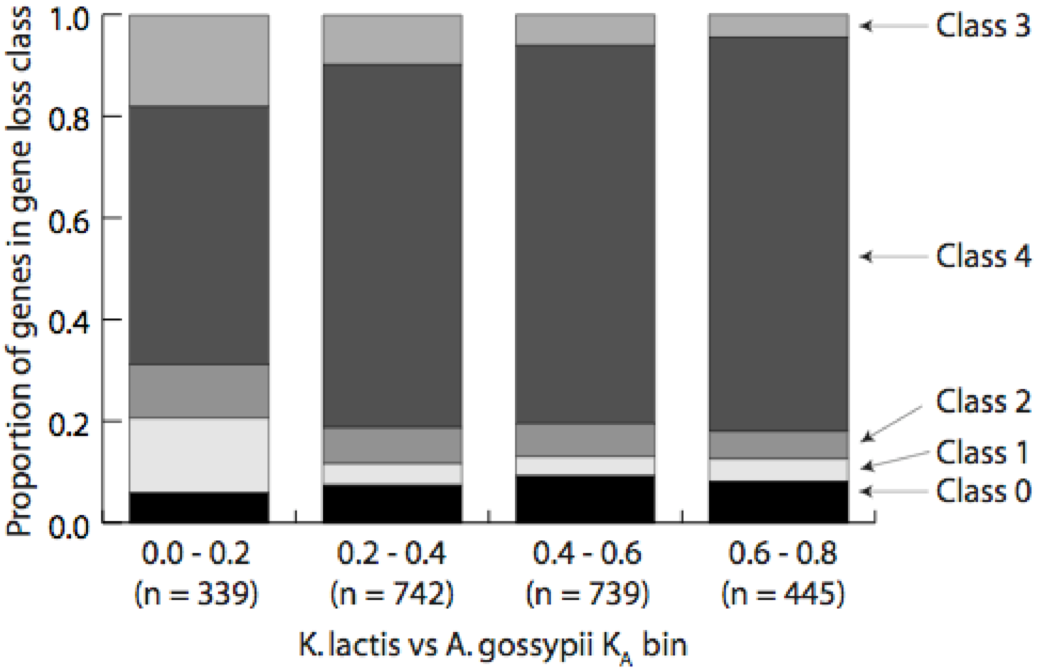


Figure 5.11 The contribution of Class 3 (RGL) loci declines with increasing rate class. The comparatively large contribution of Class 1 loci in the slowest rate bin is due to enrichment for ribosomal proteins.

Evolutionary rate and functional class contribute independently to the pattern of gene loss.

Non-parametric partial correlations (described below) were used to investigate the relationship between the following factors:

"RGL status": Whether a locus has undergone RGL or not (coded as 1 or 0).

" K_A ": Extent of nonsynonymous substitution in the same locus compared between *K. lactis* and *A. gossypii*.

"RNA": Locus is involved in RNA-related biological processes according to YPD annotation, or not (coded as 1 or 0).

Table 5.8 Evolutionary rate and functional class contribute independently to the pattern of gene loss.

Nonparametric correlation				**Nonparametric partial correlation**		
Factor 1	**Factor 2**	**Spearman's rho**	**P**	**Controlling for**	**Partial correlation**	**P**
RGL status	Ka	0.17	4.97E-11	RNA	0.17	9.17E-11
RGL status	RNA	-0.10	1.17E-04	Ka	-0.10	2.18E-04

The correlation between RGL status and K_A does not change when gene involvement in RNA-related functions is controlled for (Table 5.8). Likewise, the correlation between RGL status and RNA-related gene functions does not change when K_A is controlled for. The dataset is 1417 loci of which 171 are RGL loci and 1246 are Class 4.

We also examined the relationship between RGL status, K_A, and protein abundance (Table 5.9). Here, "Exp" is protein abundance data for *S. cerevisiae* from (Ghaemmaghami *et al.*, 2003). The correlations of RGL status with K_A and protein abundance are not independent. The dataset is 1086 loci of which 132 are RGL loci and 956 are Class 4.

Table 5.9 Evolutionary rate and protein abundance do not contribute independently to the pattern of gene loss.

Nonparametric correlation				**Nonparametric partial correlation**		
Factor 1	**Factor 2**	**Spearman's rho**	**P**	**Controlling for**	**Partial correlation**	**P**
RGL status	Ka	0.14	1.73E-06	Exp	0.09	3.62E-04
RGL status	Exp	-0.14	5.08E-06	Ka	-0.08	1.55E-03

Appendix VI Phylogeny of the 'Saccharomyces complex'

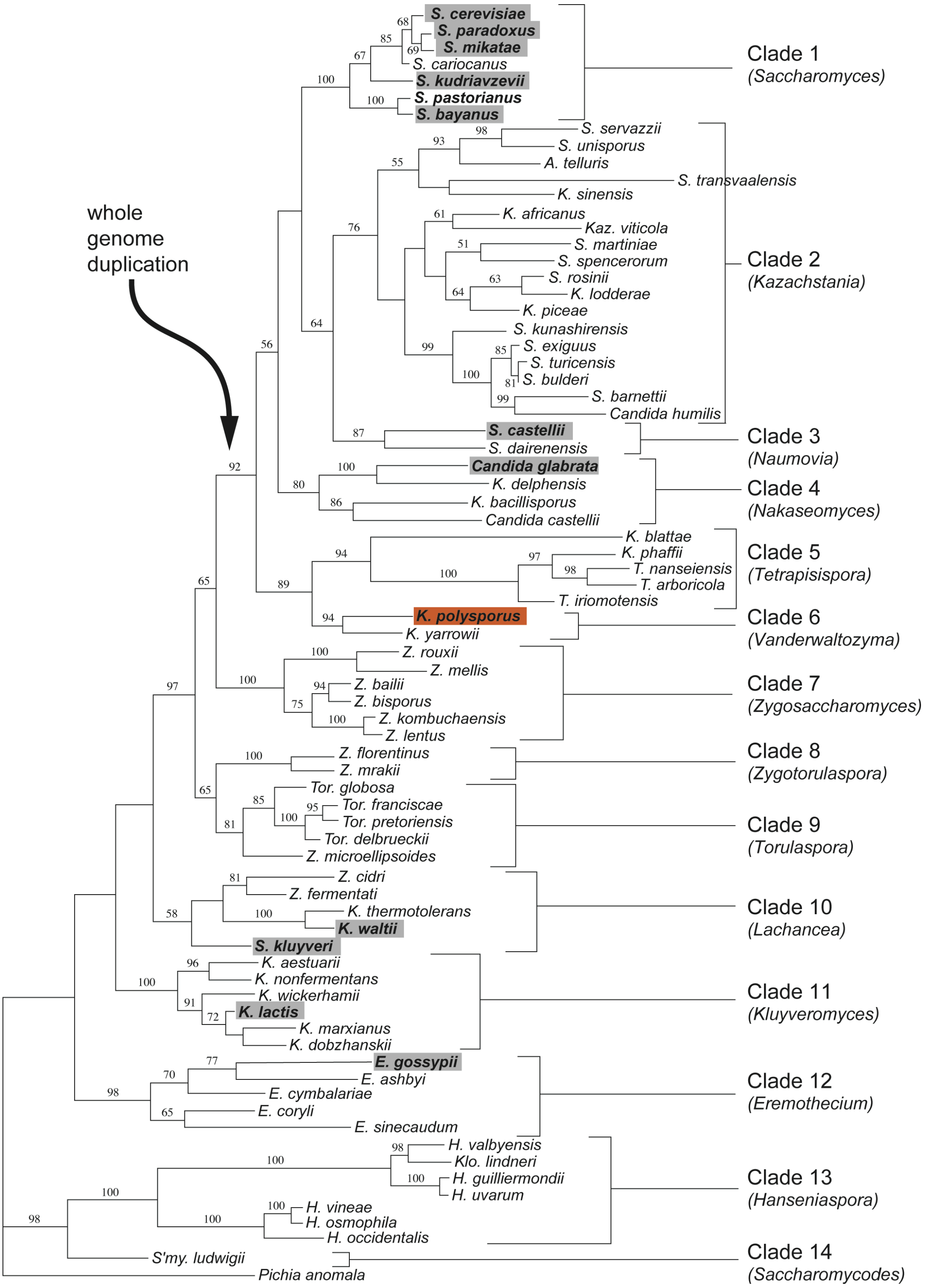


Figure 5.12 Phylogenetic tree of the 14 clades of hemiascomycetes, redrawn from Kurtzman and Robnett (Kurtzman and Robnett, 2003, Kurtzman, 2003). Species with sequenced genomes are highlighted and the inferred position of the WGD is indicated.

Appendix VII K. polysporus scaffolds

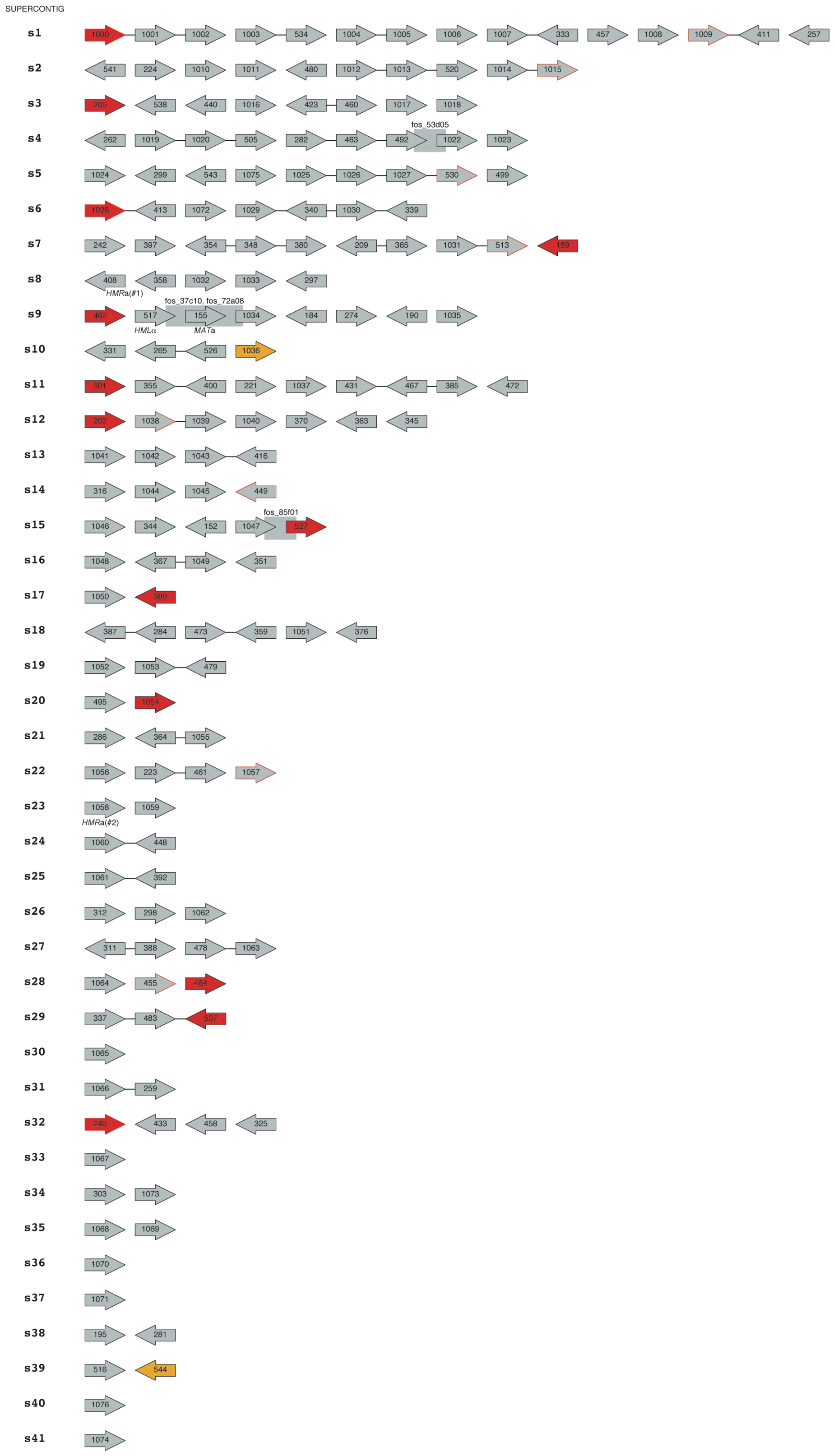
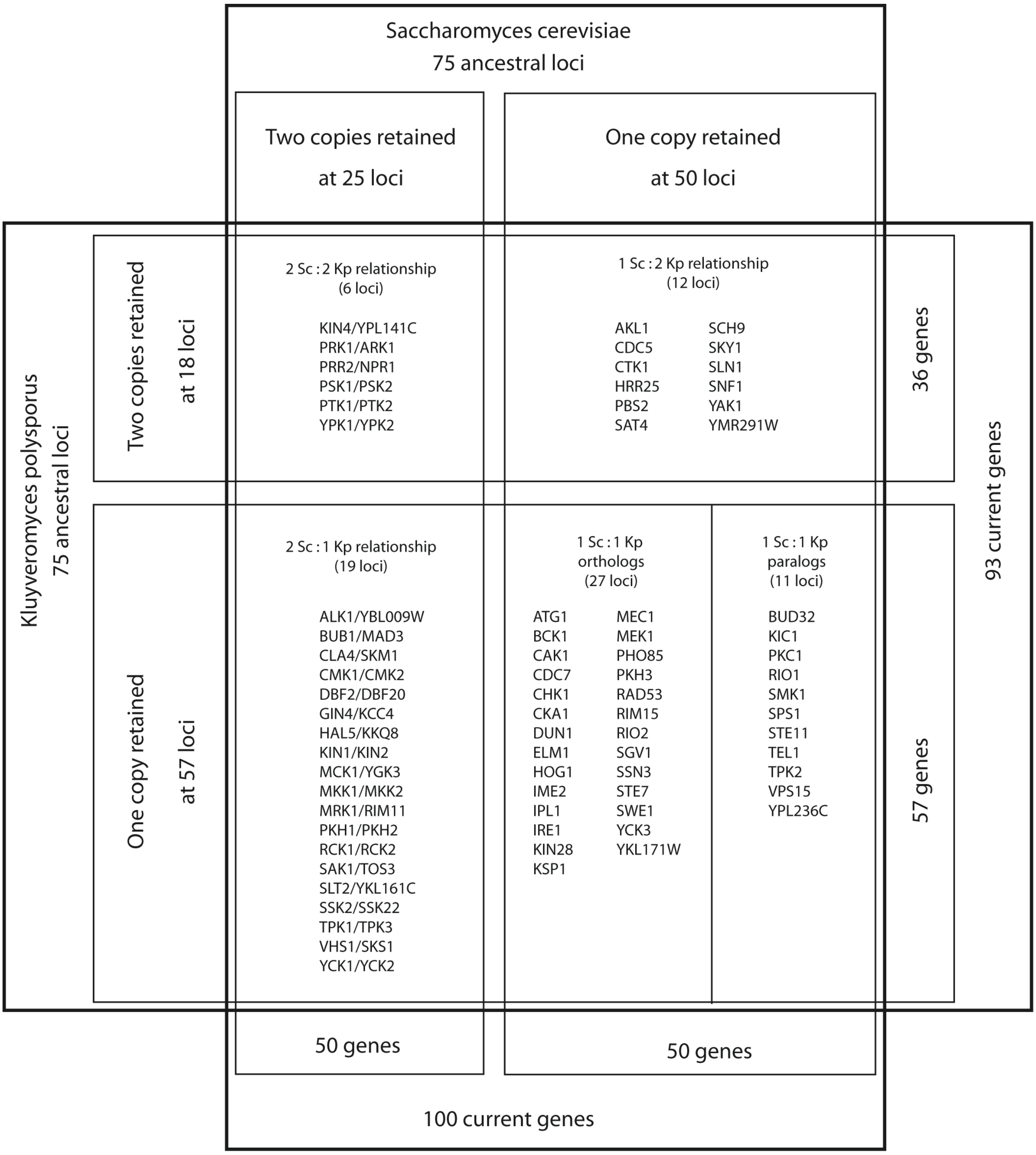


Figure 5.13 Schematic representation of the 41 *K. polysporus* supercontigs. Each row represents a supercontig, and each arrow represents a contig. Contigs with numbers >1000 consist of merged

smaller contigs, based on fosmid read-pair information and gene order information. Solid lines connect contigs between which gene order is consecutive, but where there is at least one gene missing (as compared to the non-WGD species *A. gossypii*, *K. waltii* and *K. lactis*). The order and orientation of unconnected contigs within a supercontig is based on fosmid read-pair information only. Gray rectangles indicate the positions of four fosmid clones that we completely sequenced in addition to the whole-genome shotgun phase. The locations of the *MAT*, *HML* and two *HMR* loci are shown. Red contigs contain telomeric repeats, contigs with red outline contain subtelomeric-type genes (*EXG2* exo-1,3-beta-glucanase homologs), and orange contigs contain rDNA.

Appendix VIII Patterns of gene loss among ancestrally duplicated kinases in *S. cerevisiae* and *K. polysporus*



Saccharomyces cerevisiae
75 ancestral loci

Two copies retained
at 25 loci

One copy retained
at 50 loci

Kluyveromyces polysporus
75 ancestral loci

Two copies retained
at 18 loci

One copy retained
at 57 loci

2 Sc : 2 Kp relationship
(6 loci)

KIN4/YPL141C
PRK1/ARK1
PRR2/NPR1
PSK1/PSK2
PTK1/PTK2
YPK1/YPK2

1 Sc : 2 Kp relationship
(12 loci)

AKL1
CDC5
CTK1
HRR25
PBS2
SAT4

SCH9
SKY1
SLN1
SNF1
YAK1
YMR291W

36 genes

2 Sc : 1 Kp relationship
(19 loci)

ALK1/YBL009W
BUB1/MAD3
CLA4/SKM1
CMK1/CMK2
DBF2/DBF20
GIN4/KCC4
HAL5/KKQ8
KIN1/KIN2
MCK1/YGK3
MKK1/MKK2
MRK1/RIM11
PKH1/PKH2
RCK1/RCK2
SAK1/TOS3
SLT2/YKL161C
SSK2/SSK22
TPK1/TPK3
VHS1/SKS1
YCK1/YCK2

1 Sc : 1 Kp
orthologs
(27 loci)

ATG1
BCK1
CAK1
CDC7
CHK1
CKA1
DUN1
ELM1
HOG1
IME2
IPL1
IRE1
KIN28
KSP1

MEC1
MEK1
PHO85
PKH3
RAD53
RIM15
RIO2
SGV1
SSN3
STE7
SWE1
YCK3
YKL171W

1 Sc : 1 Kp
paralogs
(11 loci)

BUD32
KIC1
PKC1
RIO1
SMK1
SPS1
STE11
TEL1
TPK2
VPS15
YPL236C

57 genes

93 current genes

50 genes

50 genes

100 current genes

Figure 5.14 Differential resolution of protein kinase gene pairs in *K. polysporus* and *S. cerevisiae*. Genes are identified by their *S. cerevisiae* names. The set of genes is based on (Hunter and Plowman, 1997). Protein kinases that are not listed could not be scored on both tracks in both species, due to sequence gaps or lack of synteny.

Appendix IX Gene Ontology (GO) terms that are significantly under- or over-represented among loci retained in duplicate since the WGD in *K. polysporus* or *S. cerevisiae*

GO terms over-represented in *K. polysporus* ohnologs, relative to single-copy genes.

	Ohnologs		Singletons		Corrected
Gene Ontology Term	Count	Percentage	Count	Percentage	P-value
Death	17	3.78%	16	0.57%	3.87E-07
cell death	16.5	3.67%	16	0.57%	1.42E-06
regulation of biological process	90	20.00%	295.5	10.55%	3.21E-06
Cytosol	49.5	11.00%	129.5	4.62%	5.41E-06
Aging	14	3.11%	14	0.50%	6.37E-06
regulation of physiological process	87.5	19.44%	290.5	10.37%	7.68E-06
regulation of cellular physiological process	84	18.67%	282	10.06%	1.15E-05
regulation of cellular process	84	18.67%	282	10.06%	1.15E-05
Cytosolic ribosome (sensu Eukaryota)	26.5	5.89%	51	1.82%	1.66E-05
Golgi-associated vesicle	16.5	3.67%	21.5	0.77%	2.24E-05
cell aging	13.5	3.00%	14	0.50%	2.28E-05
COPII vesicle coat	6	1.33%	1	0.04%	4.51E-05
ER to Golgi transport vesicle membrane	6	1.33%	1	0.04%	4.51E-05
Vesicle	20.5	4.56%	36.5	1.30%	5.51E-05
cytoplasmic vesicle	20.5	4.56%	36.5	1.30%	5.51E-05
cytoplasmic membrane-bound vesicle	20.5	4.56%	36.5	1.30%	5.51E-05
membrane-bound vesicle	20.5	4.56%	36.5	1.30%	5.51E-05
RNA processing	11.5	2.56%	216.5	7.73%	7.14E-05
G1/S transition of mitotic cell cycle	12	2.67%	13.5	0.48%	7.91E-05
Interphase	19	4.22%	34.5	1.23%	7.98E-05
interphase of mitotic cell cycle	19	4.22%	34.5	1.23%	7.98E-05
Cytosolic small ribosomal subunit (sensu Eukaryota)	13	2.89%	18	0.64%	0.000135
eukaryotic 48S initiation complex	13	2.89%	18	0.64%	0.000135
replicative cell aging	10.5	2.33%	10	0.36%	0.000136
eukaryotic 43S preinitiation complex	15	3.33%	24	0.86%	0.000138
positive regulation of cellular process	15	3.33%	24	0.86%	0.000138
positive regulation of cellular physiological process	15	3.33%	24	0.86%	0.000138
positive regulation of physiological process	15	3.33%	24	0.86%	0.000138
positive regulation of transcription	14	3.11%	21	0.75%	0.000139
carbohydrate metabolism	31.5	7.00%	81	2.89%	0.000162
ER to Golgi transport vesicle	9.5	2.11%	7.5	0.27%	0.000169
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	14	3.11%	22	0.79%	0.000199
positive regulation of transcription, DNA-dependent	13	2.89%	19	0.68%	0.000199
organellar ribosome	0	0.00%	61	2.18%	0.000212
mitochondrial ribosome	0	0.00%	61	2.18%	0.000212
positive regulation of biological process	16.5	3.67%	29	1.03%	0.000245
RNA metabolism	21.5	4.78%	296	10.56%	0.000254
regulation of progression through cell cycle	25.5	5.67%	61	2.18%	0.000261
regulation of cell cycle	25.5	5.67%	61	2.18%	0.000261
cell wall organization and biogenesis	25	5.56%	60.5	2.16%	0.000261
external encapsulating structure organization and biogenesis	25	5.56%	60.5	2.16%	0.000261
cellular carbohydrate metabolism	29.5	6.56%	74.5	2.66%	0.000276
positive regulation of cellular metabolism	14	3.11%	23	0.82%	0.000279

positive regulation of metabolism	14	3.11%	23	0.82%	0.000279
protein amino acid O-linked glycosylation	5.5	1.22%	1	0.04%	0.00028
coated vesicle	17.5	3.89%	32.5	1.16%	0.000295
regulation of metabolism	59	13.11%	202.5	7.23%	0.00034
Golgi apparatus	29	6.44%	77.5	2.77%	0.00037
response to oxidative stress	12.5	2.78%	19	0.68%	0.000586
regulation of cellular metabolism	54.5	12.11%	189	6.75%	0.00063
oxygen and reactive oxygen species metabolism	12.5	2.78%	20	0.71%	0.000819
transport vesicle membrane	6	1.33%	4	0.14%	0.000934
Golgi-associated vesicle membrane	6	1.33%	4	0.14%	0.000934
glucose metabolism	13	2.89%	23.5	0.84%	0.00103
mitotic cell cycle	36.5	8.11%	111.5	3.98%	0.001044
monosaccharide metabolism	17	3.78%	36.5	1.30%	0.001062
mRNA processing	3	0.67%	91	3.25%	0.001145
hexose metabolism	16	3.56%	33.5	1.20%	0.001406
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	47.5	10.56%	164.5	5.87%	0.001616
response to chemical stimulus	31	6.89%	96	3.43%	0.001744
small nuclear ribonucleoprotein complex	0	0.00%	46	1.64%	0.001885
DNA binding	28.5	6.33%	85	3.03%	0.002318
transcription factor activity	10	2.22%	17	0.61%	0.002444
transport vesicle	9.5	2.11%	13.5	0.48%	0.002578
regulation of transcription	42.5	9.44%	147	5.25%	0.002614
mitochondrial envelope	11.5	2.56%	173	6.17%	0.002733
plasma membrane	25	5.56%	74.5	2.66%	0.003303
bud neck	18	4.00%	47	1.68%	0.003468
response to abiotic stimulus	38	8.44%	133	4.75%	0.003509
cell cycle	55	12.22%	211	7.53%	0.003532
phosphotransferase activity, alcohol group as acceptor	25	5.56%	76	2.71%	0.003606
organelle lumen	40.5	9.00%	413.5	14.76%	0.003963
membrane-enclosed lumen	40.5	9.00%	413.5	14.76%	0.003964
mitochondrion	58.5	13.00%	554.5	19.79%	0.004202
kinase activity	28.5	6.33%	89	3.18%	0.004268
bud	22	4.89%	64	2.28%	0.004297
polysome	4	0.89%	2	0.07%	0.004511
positive regulation of gene expression, epigenetic	4	0.89%	1.5	0.05%	0.004512
loss of chromatin silencing	4	0.89%	1.5	0.05%	0.004513
regulation of translational fidelity	4	0.89%	2	0.07%	0.004515
progressive alteration of chromatin during cell aging	4	0.89%	1.5	0.05%	0.004516
translation elongation factor activity	4	0.89%	2	0.07%	0.004517
Rho GTPase activator activity	4	0.89%	2	0.07%	0.004518
development	51	11.33%	194.5	6.94%	0.004553
Golgi membrane	11	2.44%	23	0.82%	0.005099
specific RNA polymerase II transcription factor activity	8	1.78%	12.5	0.45%	0.005396
vesicle coat	8	1.78%	13	0.46%	0.005397
alcohol metabolism	23.5	5.22%	67.5	2.41%	0.005429
bud tip	10.5	2.33%	20	0.71%	0.00588
enzyme regulator activity	27	6.00%	89	3.18%	0.006631
ribosome biogenesis	8	1.78%	127.5	4.55%	0.006799
macromolecule biosynthesis	65	14.44%	268.5	9.58%	0.006933
antioxidant activity	5	1.11%	5	0.18%	0.007282
phosphatase regulator activity	5	1.11%	5	0.18%	0.007284
protein phosphatase regulator activity	5	1.11%	5	0.18%	0.007286
GTPase activator activity	8	1.78%	13.5	0.48%	0.007469
cytoplasmic vesicle membrane	8	1.78%	14	0.50%	0.007471
vesicle membrane	8	1.78%	14	0.50%	0.007473

coated vesicle membrane	8	1.78%	14	0.50%	0.007475
spliceosome complex	1	0.22%	52	1.86%	0.007565
positive regulation of transcription from RNA polymerase II promoter	9	2.00%	17.5	0.62%	0.008829
regulation of mitosis	9	2.00%	17.5	0.62%	0.008831
cell wall glycoprotein biosynthesis	4	0.89%	3	0.11%	0.009425
cell wall mannoprotein biosynthesis	4	0.89%	3	0.11%	0.009427
mannoprotein biosynthesis	4	0.89%	3	0.11%	0.00943
mannoprotein metabolism	4	0.89%	3	0.11%	0.009432
age-dependent general metabolic decline	4	0.89%	3	0.11%	0.009434
mitochondrial membrane	10.5	2.33%	151.5	5.41%	0.009475
signal transduction	24.5	5.44%	79	2.82%	0.009812
regulation of glycolysis	3	0.67%	1	0.04%	0.009867
rDNA binding	3	0.67%	1	0.04%	0.00987
RNA splicing, via transesterification reactions	3	0.67%	71.5	2.55%	0.010173
major (U2-dependent) spliceosome	0	0.00%	34	1.21%	0.010942
reproductive physiological process	27	6.00%	93	3.32%	0.011191
reproductive cellular physiological process	27	6.00%	93	3.32%	0.011194
monosaccharide catabolism	7.5	1.67%	12	0.43%	0.011288
sphingolipid metabolism	7.5	1.67%	12	0.43%	0.011291
vacuolar transport	1	0.22%	49	1.75%	0.011339
translational elongation	5	1.11%	6	0.21%	0.011905
mRNA catabolism, deadenylylation-dependent decay	5	1.11%	6	0.21%	0.011908
nuclear lumen	27.5	6.11%	288	10.28%	0.012326
ribosome	33	7.33%	120	4.28%	0.012517
cell wall	11.5	2.56%	24.5	0.87%	0.012622
external encapsulating structure	11.5	2.56%	24.5	0.87%	0.012625
cell wall (sensu Fungi)	11.5	2.56%	24.5	0.87%	0.012629
nucleoplasm	13	2.89%	164	5.85%	0.012925
cell communication	26	5.78%	88	3.14%	0.013328
membrane coat	8	1.78%	16	0.57%	0.013386
coated membrane	8	1.78%	16	0.57%	0.013389
rRNA processing	6	1.33%	101.5	3.62%	0.014194
nuclear mRNA splicing, via spliceosome	3	0.67%	68	2.43%	0.014494
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	3	0.67%	69	2.46%	0.014541
regulation of transcription, DNA-dependent	37.5	8.33%	141.5	5.05%	0.015548
carbohydrate kinase activity	4.5	1.00%	4	0.14%	0.016912
regulation of cyclin dependent protein kinase activity	4	0.89%	4	0.14%	0.016917
glucose catabolism	6.5	1.44%	10	0.36%	0.017169
hexose catabolism	6.5	1.44%	10	0.36%	0.017174
carbohydrate catabolism	8.5	1.89%	17	0.61%	0.01736
cellular carbohydrate catabolism	8.5	1.89%	17	0.61%	0.017364
actin cortical patch	8	1.78%	17	0.61%	0.017369
organellar large ribosomal subunit	0	0.00%	32	1.14%	0.01747
mitochondrial large ribosomal subunit	0	0.00%	32	1.14%	0.017475
rRNA metabolism	7	1.56%	105.5	3.77%	0.017625
hydrolase activity, hydrolyzing O-glycosyl compounds	5	1.11%	7	0.25%	0.018206
regulation of mRNA stability	5	1.11%	7	0.25%	0.01821
glycolysis	5.5	1.22%	7	0.25%	0.018215
regulation of RNA stability	5	1.11%	7	0.25%	0.01822
cytosolic large ribosomal subunit (sensu Eukaryota)	11	2.44%	28	1.00%	0.018383
transferase activity, transferring hexosyl groups	13.5	3.00%	35	1.25%	0.01902
regulation of endocytosis	2	0.44%	0	0.00%	0.02
protein phosphatase inhibitor activity	2	0.44%	0	0.00%	0.020005
positive regulation of glycolysis	2	0.44%	0	0.00%	0.02001

ligase activity, forming carbon-carbon bonds	2	0.44%	0	0.00%	0.020016
proton-transporting ATP synthase, catalytic core (sensu Eukaryota)	2	0.44%	0	0.00%	0.020021
proton-transporting ATP synthase, catalytic core	2	0.44%	0	0.00%	0.020026
protein desumoylation	2	0.44%	0	0.00%	0.020031
eukaryotic translation elongation factor 1 complex	2	0.44%	0	0.00%	0.020036
re-entry into mitotic cell cycle	2.5	0.56%	0	0.00%	0.020042
glutathione peroxidase activity	2	0.44%	0	0.00%	0.020047
ubiquitin-like-protein-specific protease activity	2	0.44%	0	0.00%	0.020052
re-entry into mitotic cell cycle after pheromone arrest	2.5	0.56%	0	0.00%	0.020057
SUMO-specific protease activity	2	0.44%	0	0.00%	0.020062
phosphatase inhibitor activity	2	0.44%	0	0.00%	0.020068
1,3-beta-glucan synthase complex	2	0.44%	0	0.00%	0.020073
protein biosynthesis	57.5	12.78%	242.5	8.65%	0.020323
alcohol catabolism	7.5	1.67%	13.5	0.48%	0.020497
site of polarized growth	21	4.67%	68.5	2.44%	0.020629
glycoprotein biosynthesis	13.5	3.00%	36	1.28%	0.020783
reproduction	33.5	7.44%	127	4.53%	0.020855
response to stimulus	62.5	13.89%	268.5	9.58%	0.021549
programmed cell death	3	0.67%	2	0.07%	0.02246
loss of chromatin silencing during replicative cell aging	3	0.67%	1.5	0.05%	0.022466
apoptosis	3	0.67%	2	0.07%	0.022472
carbohydrate transporter activity	3	0.67%	2	0.07%	0.022477
progressive alteration of chromatin during replicative cell aging	3	0.67%	1.5	0.05%	0.022483
response to reactive oxygen species	3	0.67%	2	0.07%	0.022489
glycoprotein metabolism	13.5	3.00%	37	1.32%	0.022938
small GTPase regulator activity	10	2.22%	24.5	0.87%	0.024216
actin filament organization	10.5	2.33%	24.5	0.87%	0.024223
intracellular signaling cascade	17	3.78%	54.5	1.95%	0.026041
regulation of RNA metabolism	5	1.11%	8	0.29%	0.026491
tRNA modification	0	0.00%	28	1.00%	0.026764
spindle checkpoint	4	0.89%	5	0.18%	0.027464
chronological cell aging	4.5	1.00%	5	0.18%	0.027471
nuclear nucleosome	4	0.89%	5	0.18%	0.027478
mitotic spindle checkpoint	4	0.89%	5	0.18%	0.027486
nucleosome	4	0.89%	5	0.18%	0.027493
mitotic checkpoint	4	0.89%	5	0.18%	0.0275
RNA splicing	4.5	1.00%	82	2.93%	0.027527
GTPase regulator activity	12	2.67%	32.5	1.16%	0.028193
DNA-directed RNA polymerase II, holoenzyme	2	0.44%	54	1.93%	0.029519
condensed chromosome	2	0.44%	53.5	1.91%	0.029527
protein kinase activity	18.5	4.11%	59	2.11%	0.030446
endocytosis	12	2.67%	33.5	1.20%	0.030515
response to stress	47	10.44%	199.5	7.12%	0.030805
budding cell bud growth	6	1.33%	12	0.43%	0.031601
non-developmental growth	6	1.33%	12	0.43%	0.03161
cysteine-type peptidase activity	6	1.33%	12	0.43%	0.031618
signal transducer activity	10.5	2.33%	26.5	0.95%	0.031816
growth	18	4.00%	59.5	2.12%	0.031846
biopolymer glycosylation	12	2.67%	35	1.25%	0.033459
protein amino acid glycosylation	12	2.67%	35	1.25%	0.033467
enzyme activator activity	12	2.67%	35	1.25%	0.033476
endomembrane system	38	8.44%	156	5.57%	0.035082
cellular lipid metabolism	29	6.44%	112	4.00%	0.036473
small GTPase mediated signal transduction	9	2.00%	23	0.82%	0.036688

regulation of protein kinase activity	5	1.11%	9	0.32%	0.036818
regulation of kinase activity	5	1.11%	9	0.32%	0.036828
COPI-coated vesicle	5	1.11%	9	0.32%	0.036838
cyclin-dependent protein kinase regulator activity	5	1.11%	9	0.32%	0.036847
regulation of transferase activity	5	1.11%	9	0.32%	0.036857
RNA modification	1	0.22%	38	1.36%	0.037176
sporulation	16	3.56%	53.5	1.91%	0.039203
age-dependent response to oxidative stress	3	0.67%	3	0.11%	0.040754
age-dependent general metabolic decline during chronological cell aging	3	0.67%	3	0.11%	0.040765
age-dependent response to oxidative stress during chronological cell aging	3	0.67%	3	0.11%	0.040776
regulation of translation	6.5	1.44%	12.5	0.45%	0.041094
regulation of protein biosynthesis	6.5	1.44%	12.5	0.45%	0.041105
ER-associated protein catabolism	6	1.33%	12.5	0.45%	0.041115
tRNA metabolism	4	0.89%	71	2.53%	0.041167
biosynthesis	93	20.67%	443.5	15.83%	0.041357
condensed nuclear chromosome	2	0.44%	49.5	1.77%	0.041625
biopolymer methylation	0	0.00%	25	0.89%	0.042012
mitochondrial small ribosomal subunit	0	0.00%	26	0.93%	0.04285
organellar small ribosomal subunit	0	0.00%	26	0.93%	0.042862
outer membrane	3	0.67%	60.5	2.16%	0.042875
organelle outer membrane	3	0.67%	60.5	2.16%	0.042886
mitochondrial outer membrane	3	0.67%	60.5	2.16%	0.042897
lipid metabolism	30	6.67%	120	4.28%	0.043283
main pathways of carbohydrate metabolism	11.5	2.56%	31.5	1.12%	0.044875
cellular polysaccharide metabolism	8.5	1.89%	18.5	0.66%	0.046224
translation factor activity, nucleic acid binding	8	1.78%	19	0.68%	0.046236
polysaccharide metabolism	8.5	1.89%	18.5	0.66%	0.046248
actin cytoskeleton organization and biogenesis	15.5	3.44%	50	1.78%	0.047829
nucleic acid binding	54.5	12.11%	242.5	8.65%	0.048341

GO terms over-represented in *S. cerevisiae* ohnologs, relative to single-copy genes.

	Ohnologs		Singletons		Corrected
Gene Ontology Term	Count	Percentage	Count	Percentage	P-value
cytosolic ribosome (sensu Eukaryota)	42.5	9.82%	35	1.24%	4.83E-17
cytosol	65	15.01%	114	4.04%	6.31E-14
cytosolic large ribosomal subunit (sensu Eukaryota)	23	5.31%	16	0.57%	4.78E-11
eukaryotic 48S initiation complex	19	4.39%	12	0.43%	8.77E-10
cytosolic small ribosomal subunit (sensu Eukaryota)	19	4.39%	12	0.43%	8.77E-10
structural constituent of ribosome	42	9.70%	81	2.87%	1.33E-08
ribosome	47	10.85%	106	3.76%	6.46E-08
eukaryotic 43S preinitiation complex	19	4.39%	20	0.71%	1.41E-07
RNA processing	7	1.62%	221	7.84%	4.53E-07
organelle lumen	27	6.24%	427	15.15%	2.06E-06
membrane-enclosed lumen	27	6.24%	427	15.15%	2.06E-06
RNA metabolism	15.5	3.58%	302	10.71%	4.89E-06
ribosome biogenesis	2.5	0.58%	133	4.72%	1.84E-05
macromolecule biosynthesis	74.5	17.21%	259	9.19%	3.16E-05
phosphotransferase activity, alcohol group as acceptor	29	6.70%	72	2.55%	5.69E-05
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	0	0.00%	72	2.55%	5.91E-05
mRNA processing	1	0.23%	93	3.30%	5.91E-05
biosynthesis	108.5	25.06%	428	15.18%	5.98E-05
cellular carbohydrate metabolism	30	6.93%	74	2.63%	6.15E-05
protein kinase activity	24.5	5.66%	53	1.88%	6.23E-05
large ribosomal subunit	23	5.31%	48	1.70%	6.25E-05
carbohydrate metabolism	31.5	7.27%	81	2.87%	7.33E-05
structural molecule activity	50	11.55%	161	5.71%	8.28E-05
cellular biosynthesis	98.5	22.75%	384	13.62%	8.49E-05
nuclear lumen	18.5	4.27%	297	10.54%	8.57E-05
kinase activity	32.5	7.51%	85	3.02%	9.35E-05
nuclear mRNA splicing, via spliceosome	0	0.00%	71	2.52%	9.75E-05
small ribosomal subunit	19	4.39%	38	1.35%	0.000109
cell wall organization and biogenesis	25.5	5.89%	60	2.13%	0.000151
external encapsulating structure organization and biogenesis	25.5	5.89%	60	2.13%	0.000151
plasma membrane	27.5	6.35%	72	2.55%	0.000286
nucleoplasm	8	1.85%	169	6.00%	0.000295
mitochondrial ribosome	0	0.00%	61	2.16%	0.000347
organellar ribosome	0	0.00%	61	2.16%	0.000347
rRNA processing	2.5	0.58%	105	3.72%	0.000384
protein biosynthesis	63	14.55%	237	8.41%	0.000471
cell wall	13	3.00%	23	0.82%	0.000528
external encapsulating structure	13	3.00%	23	0.82%	0.000528
cell wall (sensu Fungi)	13	3.00%	23	0.82%	0.000528
biopolymer biosynthesis	7.5	1.73%	6	0.21%	0.000641
polysaccharide biosynthesis	7.5	1.73%	6	0.21%	0.000642
RNA splicing, via transesterification reactions	0.5	0.12%	74	2.63%	0.000704
mRNA metabolism	5	1.15%	124	4.40%	0.000715
protein amino acid phosphorylation	18	4.16%	41	1.45%	0.000719
spliceosome complex	0	0.00%	53	1.88%	0.000811
protein serine/threonine kinase activity	14.5	3.35%	28	0.99%	0.000828
rRNA metabolism	3.5	0.81%	109	3.87%	0.000929
biopolymer metabolism	91.5	21.13%	883	31.32%	0.001091
cyclin-dependent protein kinase regulator activity	7	1.62%	7	0.25%	0.001136
phosphorylation	23.5	5.43%	63	2.23%	0.00116
signal transduction	26.5	6.12%	77	2.73%	0.001247

energy reserve metabolism	8.5	1.96%	10	0.35%	0.001319
glycogen biosynthesis	4.5	1.04%	1	0.04%	0.001449
glucan biosynthesis	5.5	1.27%	3	0.11%	0.001714
regulation of cyclin dependent protein kinase activity	5	1.15%	3	0.11%	0.001715
cellular polysaccharide metabolism	10	2.31%	17	0.60%	0.001824
polysaccharide metabolism	10	2.31%	17	0.60%	0.001824
35S primary transcript processing	0	0.00%	49	1.74%	0.002013
regulation of cell redox homeostasis	3.5	0.81%	0	0.00%	0.002424
cell redox homeostasis	3.5	0.81%	0	0.00%	0.002425
glucan metabolism	8	1.85%	12	0.43%	0.002963
small nuclear ribonucleoprotein complex	0	0.00%	46	1.63%	0.003123
transferase activity, transferring phosphorus-containing groups	38.5	8.89%	137	4.86%	0.003191
cell communication	27	6.24%	87	3.09%	0.003253
RNA splicing	2.5	0.58%	84	2.98%	0.00352
nucleus	116.5	26.91%	1047	37.14%	0.003801
alcohol metabolism	22	5.08%	69	2.45%	0.005108
regulation of cellular process	69	15.94%	297	10.54%	0.005213
regulation of cellular physiological process	69	15.94%	297	10.54%	0.005215
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	84	19.40%	772	27.39%	0.005518
carbohydrate biosynthesis	11.5	2.66%	25	0.89%	0.006134
response to abiotic stimulus	36	8.31%	135	4.79%	0.006136
reproductive cellular physiological process	27	6.24%	93	3.30%	0.006547
reproductive physiological process	27	6.24%	93	3.30%	0.006549
endocytosis	13.5	3.12%	32	1.14%	0.00657
regulation of transferase activity	6	1.39%	8	0.28%	0.006723
glycogen metabolism	6.5	1.50%	8	0.28%	0.006725
regulation of protein kinase activity	6	1.39%	8	0.28%	0.006727
regulation of kinase activity	6	1.39%	8	0.28%	0.006728
transcription factor activity	9	2.08%	18	0.64%	0.006842
RNA splicing factor activity, transesterification mechanism	0	0.00%	38	1.35%	0.007156
small GTPase mediated signal transduction	10	2.31%	22	0.78%	0.007418
regulation of physiological process	70	16.17%	308	10.93%	0.007517
phosphorus metabolism	26.5	6.12%	88	3.12%	0.007759
phosphate metabolism	26.5	6.12%	88	3.12%	0.007761
cytoplasm organization and biogenesis	9.5	2.19%	149	5.29%	0.007765
ribosome biogenesis and assembly	9.5	2.19%	149	5.29%	0.007767
condensed chromosome	0.5	0.12%	55	1.95%	0.007979
DNA-directed RNA polymerase II, holoenzyme	1	0.23%	55	1.95%	0.007981
regulation of biological process	71.5	16.51%	314	11.14%	0.00803
pyrimidine base metabolism	4	0.92%	3	0.11%	0.008174
UDP-glucosyltransferase activity	4	0.92%	3	0.11%	0.008177
chromosome	12	2.77%	167	5.92%	0.008494
enzyme regulator activity	26	6.00%	90	3.19%	0.008505
oxidoreductase activity, acting on the CH-CH group of donors, quinone or related compound as acceptor	3	0.69%	1	0.04%	0.0088
succinate dehydrogenase (ubiquinone) activity	3	0.69%	1	0.04%	0.008802
thiol-disulfide exchange intermediate activity	3	0.69%	1	0.04%	0.008804
intracellular membrane-bound organelle	233	53.81%	1903	67.51%	0.008919
membrane-bound organelle	233	53.81%	1903	67.51%	0.008921
ribonucleoprotein complex	51	11.78%	215	7.63%	0.00975
protein complex	92.5	21.36%	827	29.34%	0.009848
vacuolar transport	1	0.23%	49	1.74%	0.011299
condensed nuclear chromosome	0.5	0.12%	51	1.81%	0.011526
endomembrane system	14	3.23%	180	6.39%	0.011606

reproduction	33.5	7.74%	127	4.51%	0.013068
G1/S transition of mitotic cell cycle	8.5	1.96%	17	0.60%	0.013835
organelle organization and biogenesis	67	15.47%	614	21.78%	0.013866
mitochondrial lumen	6	1.39%	104	3.69%	0.014064
mitochondrial matrix	6	1.39%	104	3.69%	0.014068
intracellular signaling cascade	17.5	4.04%	54	1.92%	0.014292
DNA recombination	1.5	0.35%	63	2.23%	0.014337
bud tip	9.5	2.19%	21	0.74%	0.014462
lipid metabolism	31	7.16%	119	4.22%	0.014874
ribonucleotide biosynthesis	7	1.62%	14	0.50%	0.016603
response to chemical stimulus	27	6.24%	100	3.55%	0.017144
organellar large ribosomal subunit	0	0.00%	32	1.14%	0.017259
mitochondrial large ribosomal subunit	0	0.00%	32	1.14%	0.017263
major (U2-dependent) spliceosome	0	0.00%	34	1.21%	0.018314
ATP-dependent helicase activity	0	0.00%	34	1.21%	0.018318
septin ring assembly	2	0.46%	0	0.00%	0.01838
thioredoxin peroxidase activity	2	0.46%	0	0.00%	0.018385
glycogen synthase kinase 3 activity	2	0.46%	0	0.00%	0.018389
tRNA-pseudouridine synthase activity	2	0.46%	0	0.00%	0.018394
regulation of glycogen catabolism	2	0.46%	0	0.00%	0.018399
septin ring organization	2	0.46%	0	0.00%	0.018403
ligase activity, forming carbon-carbon bonds	2	0.46%	0	0.00%	0.018408
regulation of glycogen biosynthesis	2.5	0.58%	0	0.00%	0.018413
small GTPase regulator activity	10.5	2.42%	24	0.85%	0.019155
helicase activity	1.5	0.35%	58	2.06%	0.019815
transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer	3	0.69%	2	0.07%	0.019981
pyrimidine base biosynthesis	3	0.69%	2	0.07%	0.019986
disulfide oxidoreductase activity	3	0.69%	2	0.07%	0.019991
organelle membrane	32.5	7.51%	333	11.81%	0.021585
protein kinase regulator activity	8	1.85%	19	0.67%	0.02237
glucosyltransferase activity	4	0.92%	5	0.18%	0.023811
proteolysis	7	1.62%	106	3.76%	0.023904
covalent chromatin modification	1	0.23%	43	1.53%	0.024994
chromosome, pericentric region	0.5	0.12%	43	1.53%	0.025001
histone modification	1	0.23%	43	1.53%	0.025007
regulation of progression through cell cycle	19.5	4.50%	67	2.38%	0.02567
regulation of cell cycle	19.5	4.50%	67	2.38%	0.025676
interphase of mitotic cell cycle	13.5	3.12%	40	1.42%	0.025728
interphase	13.5	3.12%	40	1.42%	0.025735
monosaccharide metabolism	13.5	3.12%	40	1.42%	0.025741
regulation of enzyme activity	6.5	1.50%	12	0.43%	0.026287
signal transducer activity	10	2.31%	27	0.96%	0.026421
phosphoric monoester hydrolase activity	12	2.77%	34	1.21%	0.027044
protein amino acid acetylation	0	0.00%	30	1.06%	0.027405
nuclear envelope-endoplasmic reticulum network	4.5	1.04%	85	3.02%	0.02766
proteolysis during cellular protein catabolism	5	1.15%	86	3.05%	0.027671
ribonucleotide metabolism	7	1.62%	16	0.57%	0.027761
meiotic recombination	0	0.00%	31	1.10%	0.02868
transcription factor complex	3.5	0.81%	77	2.73%	0.029567
protein serine/threonine phosphatase activity	5	1.15%	9	0.32%	0.031296
phosphoric ester hydrolase activity	12	2.77%	37	1.31%	0.034093
hexose metabolism	12.5	2.89%	37	1.31%	0.034102
development	45.5	10.51%	200	7.09%	0.034667
DNA metabolism	28	6.47%	283	10.04%	0.034691
nucleolus	11	2.54%	139	4.93%	0.035721

Ras protein signal transduction	4.5	1.04%	6	0.21%	0.035763
antioxidant activity	4	0.92%	6	0.21%	0.035772
oxidoreductase activity, acting on the CH-CH group of donors	4	0.92%	6	0.21%	0.035782
actin cap	4	0.92%	6	0.21%	0.035791
regulation of translational fidelity	3	0.69%	3	0.11%	0.036319
response to salt stress	3	0.69%	3	0.11%	0.036329
translation elongation factor activity	3	0.69%	3	0.11%	0.036338
mitochondrial transport	3	0.69%	3	0.11%	0.036348
kinetochore	0.5	0.12%	40	1.42%	0.037222
ubiquitin-dependent protein catabolism	5	1.15%	84	2.98%	0.038345
modification-dependent protein catabolism	5	1.15%	84	2.98%	0.038355
cytoplasm	311	71.82%	1713	60.77%	0.0398
cortical cytoskeleton	9.5	2.19%	25	0.89%	0.040261
cortical actin cytoskeleton	9.5	2.19%	25	0.89%	0.040272
nuclear chromosome	12	2.77%	144	5.11%	0.040654
methyltransferase activity	1.5	0.35%	52	1.84%	0.04162
mitochondrial small ribosomal subunit	0	0.00%	26	0.92%	0.042086
organellar small ribosomal subunit	0	0.00%	26	0.92%	0.042097
ubiquitin ligase complex	0	0.00%	26	0.92%	0.042108
growth	17.5	4.04%	60	2.13%	0.042478
transferase activity, transferring one-carbon groups	1.5	0.35%	53	1.88%	0.042772
purine ribonucleotide biosynthesis	6	1.39%	14	0.50%	0.043797
specific RNA polymerase II transcription factor activity	6.5	1.50%	14	0.50%	0.043808
generation of precursor metabolites and energy	25.5	5.89%	99	3.51%	0.045675
energy derivation by oxidation of organic compounds	22.5	5.20%	86	3.05%	0.046304
cellular protein catabolism	5.5	1.27%	89	3.16%	0.046985

Appendix X Notes on the gene content of *K. polysporus*

Mating type loci

The life cycle of *K. polysporus* has been described in detail (van der Walt, 1956, Roberts and van der Walt, 1959). It is homothallic, and we identified a homolog (*Kpol_1054.32*) of the *HO* endonuclease gene, which catalyzes mating-type switching in *S. cerevisiae*. *K. polysporus* has been reported to grow primarily as a haploid (zygotes do not bud but instead sporulate soon after formation) (Roberts and van der Walt, 1959), but our sequenced isolate was either diploid or contained a mixture of *MAT***a** and *MAT*α haploid cells. We identified eight clones in our fosmid library with ~40 kb inserts spanning the *MAT* locus (in supercontig s9; Appendix VII), of which five contained a *MAT***a** allele and three contained a *MAT*α allele, as determined by sequencing the fosmids with a primer flanking the *MAT* locus. We completely sequenced the inserts in one *MAT*α fosmid (fos_37c10) and one *MAT***a** fosmid (fos_72a08) and found that they had no sequence differences other than the α-specific and **a**-specific "Y" regions of the *MAT* locus. Unusually, the *K. polysporus* genome sequence includes three silent copies of mating-type information: two *HMR***a**-like loci (in supercontigs s8 and s23) and one *HML*α-like locus (in supercontig s9, 100 kb from the *MAT* locus). Like *Candida glabrata* (Fabre *et al.*, 2005), the genome of *K. polysporus* does not contain a homolog of the *S. cerevisiae* silencing gene *SIR1*, although *SIR2*, *SIR3* and *SIR4* homologs are present. (The *K. polysporus* ohnolog pair *Kpol_1032.18* and *Kpol_479.28* corresponds to the *S. cerevisiae* ohnolog pair *SIR2* and *HST1*; the pair *Kpol_1001.11* and *Kpol_520.35* corresponds to the pair *SIR3* and *ORC1*; *Kpol_269.1* is an ortholog of *SIR4.*)

Genes for pheromones and their receptors

K. polysporus has two copies (ohnologs) of the α-pheromone gene. One copy (*Kpol_1002.67*) codes for five identical repeats of the peptide WHWLELDNGQPIY, and the other (*Kpol_1033.32*) codes for four identical repeats of the peptide WHWLRLRYGEPIY. The 9/13 amino acid match between these two putative pheromone peptides is surprisingly low. Interestingly, *K. polysporus* retains two ohnolog copies of the *STE2* α-pheromone receptor (*Kpol_1011.19* and *Kpol_1058.22*), so it is possible that there are two separately interacting pheromone/receptor pairs in this species. The only **a**-

pheromone genes in *K. polysporus* (*Kpol_1039.70*, *Kpol_1039.70a*, and *Kpol_1039.70b*) are in a triple tandem repeat at a locus that is in a paralogous relationship (reciprocal gene loss after WGD) with *S. cerevisiae MFA2*. *K. polysporus* retains a single ortholog of the *STE3* **a**-factor receptor gene (*Kpol_1022.10*).

Subtelomeric regions

The subtelomeric regions of the *K. polysporus* genome contain multiple genes (at least 19 copies) for exo-1,3-beta-glucanase, an enzyme that degrades the cell wall polymer beta-glucan. In *S. cerevisiae* there are only three exo-1,3-beta-glucanase genes (*SPR1*, *EXG1* and *EXG2*), and they function in cell wall assembly and spore wall morphogenesis (Muthukumar *et al.*, 1993, Esteban *et al.*, 1999). The amplification of this family in *K. polysporus* is possibly related to its multi-spored phenotype.

Protein complexes

Protein complexes and genes coding for their components tend to be lost and gained relatively rarely during evolution. However, we noticed that the genes coding for all three subunits (*SSY1*, *SSY5* and *PTR3*) of the SPS extracellular amino acid sensor system (Forsberg and Ljungdahl, 2001), and several subunits of dynein and dynactin (discussed in main text) are absent from the genome of *K. polysporus*, as are genes for enzymes of the DAL pathway (*DAL1*, *DAL2*, *DAL3*, *DAL4*, *DAL7* and *DCG1*; these are not known to form a complex) (Wong and Wolfe, 2005). In addition, six (*SFB3*, *SEC13*, *SEC16*, *SEC23*, *SEC31* and *SEC24*/*SFB2*) of the seven genes coding for subunits of the COPII vesicle complex are retained as ohnolog pairs in *K. polysporus*. Only *SEC24*/*SFB2* is present in duplicate in *S. cerevisiae* and *SAR1* is duplicated in neither species. COPII proteins coat and direct the formation of vesicles that transport proteins from the ER to the golgi and may also have a role in 'cargo' protein selection (Kirchhausen, 2000). Genes coding for COPII subunits are evolutionarily well conserved and most have single orthologs in mammals (Kirchhausen, 2000). Three interacting subunits of the F_1 portion of the mitochondrial F_1F_0-ATPase (*ATP1*, *ATP2* and *ATP5*) have also been retained as ohnolog pairs in *K. polysporus* but not in other post-WGD yeasts.

Species-specific genes

The *K. polysporus* genome contains some multicopy gene families that have no homologs in other yeasts. A similar situation exists in *S. castellii* (Cliften *et al.*, 2006). Representative members of *K. polysporus*-specific families are *Kpol_489.2* and *Kpol_1035.52*. Other *K. polysporus* gene families, such as those represented by *Kpol_387.6* and *Kpol_487.8*, lack homologs in *S. cerevisiae* but are also multigene families in other yeasts such as *S. castellii* or *C. glabrata*. None of these genes have functionally characterized homologs in any other organism. We also noticed that *K. polysporus* has a gene (*Kpol_520.25*) coding for a protein in the Argonaute family. Argonaute proteins bind small RNAs and usually function in gene silencing. Although present in most eukaryotes, including the filamentous euascomycetes and *Schizosaccharomyces pombe*, there are no Argonaute homologs in *S. cerevisiae*. The *K. polysporus* Argonaute gene has a WGD-derived paralog in *S. castellii* (*Scas_719.65)* but not in any of the other species (post-WGD or pre-WGD) in YGOB. There is also an Argonaute homolog in *C. albicans* (Nakayashiki *et al.*, 2006).

Transposable elements

We identified at least 39 LTR (long terminal repeat) retrotransposons, similar to the Ty elements of *S. cerevisiae.* The exact number of retroelements is uncertain because many of them cause gaps between contigs. We named the elements Tkp1, Tkp3, Tkp4 and Tkp5, following the nomenclature of ref. (Neuveglise *et al.*, 2002), of which the most common type of solo LTR is Tkp5. Although most retroelements are inserted near tRNA or rRNA genes or in telomeric regions, there are two cases where a Tkp5 element interrupts an otherwise intact protein coding gene (*Kpol_1036.28* and *Kpol_1047.49*), suggesting that the insertions are recent and that Tkp5 is an active element.

Appendix XI Measuring the effect of the ortholog-paralog bias in YGOB's tracking algorithm

YGOB uses an algorithm based on shared gene content in a local (41 locus) sliding window to assign orthology of the sister genomic regions (tracks) among different post-WGD species (Byrne and Wolfe, 2005), but the high levels of independent gene loss that have occurred between *K. polysporus* and the other post-WGD yeasts make this assignment difficult in most parts of the genome. In the region shown in Figure 3.1, for example, there are two places where YGOB's algorithm 'changes its mind' about how orthology and paralogy are assigned between *K. polysporus* and *S. cerevisiae* chromosomes. We refer to the process of identifying orthologous chromosomal regions between species as 'tracking'.

In the whole-genome comparison of the 3252 ancestral loci that could be reliably scored as present or absent in both *K. polysporus* and *S. cerevisiae*, YGOB scored 44.7% of loci as single-copy orthologs and 34.6% as single-copy paralogs (reciprocal gene losses) (Table 3.1). Because YGOB's algorithm works on the principle that orthologous regions should have higher similarity of gene content than paralogous regions, and because it operates on a local window, it has a built-in bias that will cause it to overestimate the number of orthologs in situations where the true numbers of orthologs and paralogs are similar.

We measured the effect of this bias by using the YGOB engine to create and score 100 *K. polysporus* pseudo-genomes in which any possible signal of shared ancestry with *S. cerevisiae* was obliterated. While scoring the real *K. polysporus* genome against the ancestral gene order ('Real genome' columns in Table 5.10) we created 100 pseudo-genomes where at every locus with a syntenic *K. polysporus* presence on one track and a syntenic *K. polysporus* absence on the other track, we swapped the syntenic gene from its chromosome into the syntenic gap in the chromosome on the other track with a probability of 0.5. This procedure means that the pseudo-genomes must, on average, contain equal numbers of orthologs and paralogs of the *S. cerevisiae* single-copy genes. We then used the YGOB engine to score these 100 pseudo-genomes, calculating a mean and standard deviation for each locus class (Table 5.10). As would be expected due to the randomizations' breaking of chromosomes into smaller syntenic fragments, the number of scoreable loci in the pseudo-genomes is less than in the real genome. Nevertheless the average proportions of single-copy orthologs (43.42% ± s.d. 2.23%) and paralogs (33.80% ± s.d. 2.64%) reported in the pseudo-genomes are the same as in the real data, instead of being equal to each other.

Thus, the reported excess of orthologs over paralogs in Table 5.10 may be due to YGOB's bias towards reporting orthologs. These results fail to reject the null hypothesis of no shared gene losses on the phylogenetic branch between the WGD and the common ancestor of *K. polysporus* and *S. cerevisiae*, such as would occur if they had undergone completely independent WGD events. However, modeling gene losses using a likelihood approach does reveal a signal of shared ancestry (Appendix XIV).

Table 5.10 Percentages of loci in different retention classes between *S. cerevisiae* and the real *K. polysporus* genome, and in 100 pseudo-genomes where the tracking of *K. polysporus* single-copy genes was randomized.

Locus class (K. pol.:S. cer.)	**Real genome**		**Pseudo-genomes**			
			Number of loci		**Percent**	
	Number of loci	**Percent**	**Mean**	**S.D.**	**Mean**	**S.D.**
2:2	212	6.52%	209.17	1.81	7.60%	0.87%
2:1	238	7.32%	234.90	1.65	8.53%	0.70%
1:2	221	6.80%	183.27	5.00	6.66%	2.73%
1:1 orthologs	1455	44.74%	1195.72	26.65	43.42%	2.23%
1:1 paralogs	1126	34.62%	930.61	24.61	33.80%	2.64%
Total	3252		2753.67			
Proportion of paralogs among 1:1 loci		44%			44%	1%

Appendix XII Relationship between the estimated fraction of paralogous single-copy genes and the confidence of YGOB's orthologous track assignment between *K. polysporus* and *S. cerevisiae*

Our estimate that 44.7% of single-copy loci in *K. polysporus* and *S. cerevisiae* are paralogs (Table 3.1) is based on scoring all 3252 ancestral loci that can be compared between the two species, using the YGOB engine (Byrne and Wolfe, 2005). The accuracy of this estimate depends on the accuracy with which YGOB identifies, in any genomic region, the correct overall orthology and paralogy relationships among the two *K. polysporus* genomic tracks (K1 and K2 in Figure 5.15) and the two *S. cerevisiae* genomic tracks (S1 and S2). We refer to this identification process as 'tracking'. If the tracking of a particular genomic region is incorrect, individual single-copy loci within that region will be mis-called (orthologs will be misidentified as paralogs, and *vice versa*).

We were concerned that our estimate of the proportion of paralogs in the genome might be inflated by the inclusion of mis-tracked genomic regions in the analysis. However, using a heuristic measure of the confidence of tracking, we show below that there are few regions of the genome where the percentage of single-copy loci that are paralogs is less than 20%, and that the fraction of paralogs is at least 30% in the half of the genome that is most confidently tracked.

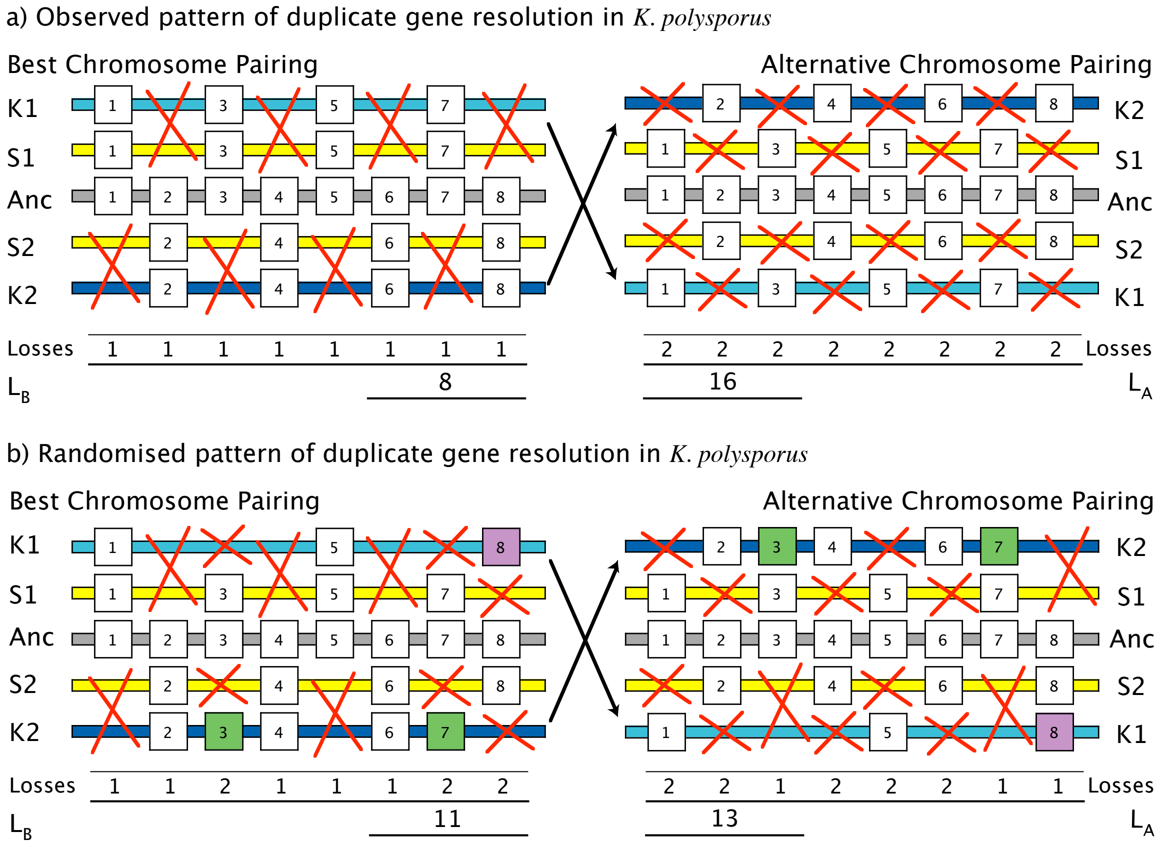

a) Observed pattern of duplicate gene resolution in *K. polysporus*
Best Chromosome Pairing
Alternative Chromosome Pairing
K1
S1
Anc
S2
K2
Losses 1 1 1 1 1 1 1 1
L_B 8
Losses 2 2 2 2 2 2 2 2
L_A 16
b) Randomised pattern of duplicate gene resolution in *K. polysporus*
Best Chromosome Pairing
Alternative Chromosome Pairing
Losses 1 1 2 1 1 1 2 2
L_B 11
Losses 2 2 1 2 2 2 1 1
L_A 13

Figure 5.15 Method for estimating confidence of orthologous track assignment. See text for details. 'Anc' represents the ancestral gene order before WGD.

We used YGOB to find pairs of homologous chromosomal segments in the genomes of both *S. cerevisiae* and *K. polysporus* that have remained unrearranged since the WGD and where no sequence gaps exist in the *K. polysporus* assembly. We retrieved 98 such 'blocks' (a pair of contiguous homologous chromosomal segments from *S. cerevisiae* and the corresponding pair of regions from *K. polysporus*), ranging in length from 10 to 73 genes, and containing a total of 1765 ancestral loci.

For each block we considered the two possible orthologous chromosomal pairings between the *S. cerevisiae* and *K. polysporus* segments (*i.e.*, S1 orthologous to K1 and S2 orthologous to K2, or S1 orthologous to K2 and S2 orthologous to K1). We counted the number of gene losses, *L*, required to account for the observed pattern of gene loss in each case. We assumed that all gene losses were of single genes (Byrnes *et al.*, 2006) and that where a gene is missing from an orthologous locus (in the context of the pairing being considered) in both species, it was lost in the common ancestor. We refer to the chromosomal pairing that requires the fewest gene losses (L_B in Figure 8a) as the 'best' pairing and the other possible pairing as the 'alternative' pairing (which requires L_A losses).

$D = L_A - L_B$ gives the number of loci that support the best pairing over the alternative pairing and has a value between 0 and the length of the block.

If there are many more single-copy orthologs (which can be explained by single gene losses in the common ancestor of *S. cerevisiae* and *K. polysporus*) in the best chromosomal pairing than in the alternative pairing, D is large and parsimony favors the best pairing as the true orthologous pairing (in the example in Figure 5.15a, $D = 8$). By contrast, if the numbers of single-copy orthologs in the best and alternative pairings are approximately equal, D will be close to zero and neither chromosomal pairing is well supported. We assigned significance to D by comparing the observed value of D for the best pairing (D_{Real}) to a null distribution obtained by calculating D for randomized blocks (D_{Rand}). Randomizations preserved the number of genes retained in each genome but randomized the pattern of duplicate gene resolution by reassigning genes from *K. polysporus* segment K1 to the paralogous locus on *K. polysporus* segment K2 with a probability of 0.5 (compare loci 3, 7, and 8 between panels a and b in Figure 5.15). The percentage of randomized datasets for which D_{Rand} is less than D_{Real} is a measure of our confidence that the best pairing reflects a correct assignment of orthologous tracks.

We found that orthologous chromosomes can be inferred with reasonable confidence in some regions of the genome, but that in others (even where relatively large contiguous regions exist in both *S. cerevisiae* and *K. polysporus*) the pattern of gene loss is not significantly different from that predicted by independent WGD events (*i.e.,* no shared history). For instance, although block 91 is 57 genes long, the best chromosome pairing requires only 3 fewer losses to explain than the alternative, which is better than only 25% of randomized datasets. By contrast, for block 43 (15 genes long) the best pairing involves 9 fewer losses than the alternative, which is better than 99% of randomizations.

We stratified blocks according to intervals of our confidence statistic (Table 5.11) and calculated the percentage of single-copy orthologs and single-copy paralogs in each stratum. The estimated proportion of orthologs decreases as the tracking confidence decreases. This is as expected, because a block with a high content of orthologs should be easy to track. No matter what the average proportion of orthologs is across the whole genome, we would expect there to be some regional variation (purely by chance) resulting in some blocks with confident tracking and high ortholog content, and other blocks with lower tracking confidence and lower ortholog content.

Table 5.11 indicates that, even in the most confidently-tracked blocks in the genome (containing 12.7% of the studied loci), 17.4% of single-copy loci are paralogs between *K. polysporus* and *S. cerevisiae*. Among the best-tracked 55.8% of loci (the top four strata), the estimated fraction of paralogs is 31.7%. Similar to YGOB's estimate for the whole genome (Table 5.11), we estimate that among all 98 blocks considered here the proportion of single-copy loci that are paralogs is 38.9%.

Table 5.11 Estimated proportions of orthologous and paralogous loci between *K. polysporus* and *S. cerevisiae*, in 98 genomic blocks stratified according to confidence of track assignment.

Tracking confidence percentile	**Total Blocks**	**Total Loci**	**Number of single-copy loci (as %)**			**Cumulative percentage* of**		
			Total	**Orthologs**	**Paralogs**	**Orthologs**	**Paralogs**	**Loci**
81-100	12	225	109	90 (82.6)	19 (17.4)	82.6	17.4	12.7
61-80	17	274	134	96 (71.6)	38 (28.4)	76.5	23.5	28.3
41-60	10	170	82	55 (67.1)	27 (32.9)	74.2	25.8	37.9
21-40	17	315	155	87 (56.1)	68 (43.9)	68.3	31.7	55.8
1-20	10	228	107	58 (54.2)	49 (45.8)	65.8	34.2	68.7
0	32	553	298	155 (52.0)	143 (48)	61.1	38.9	100.0

* Cumulative precentage calculated across the confidence percentile intervals 81-100%, 61-100%, 41-100%, 21-100%, 1-100% and 0-100%.

Appendix XIII Calculating the expected number of shared ohnolog pairs between *S. cerevisiae* and *K. polysporus*

The high level of paralogy (~44.7%) among genes that are single-copy in both *S. cerevisiae* and *K. polysporus* indicates that the fates of most duplicated loci were not determined at the time of divergence of these two species. Indeed, our model indicates that 79% of loci were still duplicated and in the U ('undecided') state at this time (Figure 3.2; Appendix XIV). Since 47% of loci that are currently duplicated in *K. polysporus* are also present in duplicate in *S. cerevisiae* (212 of 450, among the 3252 loci studied in Table 3.1), this suggests substantial convergent preservation of duplicates. We estimated the number of duplicate genes that were preserved convergently in two different ways.

Method 1: Assuming negligible shared ancestry

Because *S. cerevisiae* and *K. polysporus* diverged very soon after the WGD we estimated the number of loci that would be preserved in duplicate under the assumption of negligible shared ancestry (*i.e*., the length of the shared evolutionary branch after WGD is effectively zero) and in the absence of selection. Although this is a very naïve calculation it serves as an estimate of the number of duplicate pairs that will be shared due to chance alone. In the genomes of *S. cerevisiae* and *K. polysporus* 13% and 14% of loci respectively are present in duplicate and the expected number of shared duplicate loci is therefore 0.13 * 0.14 * 3252 = 60 loci. Since the observed number of shared duplicates is 212 (approximately 3.5 times the expected), this represents an excess of 152 loci.

Method 2: Accounting for the shared evolutionary branch

Using the model described in Appendix XIV it is possible to estimate the number of loci that were preserved in duplicate in the common ancestor of *S. cerevisiae* and *K. polysporus.* Note that the model estimates were calculated on a reduced dataset of 2299 loci, which contains exactly 169 loci (7.35%) in each of three configurations: duplicated in *S. cerevisiae* only; duplicated in *K. polysporus* only; and duplicated in both species. The model estimates that 1.93% of loci (44.4 loci) were fixed in duplicate prior to the divergence of *S. cerevisiae* and *K. polysporus,* and 5.42% of loci (7.35% - 1.93% = 5.42%; 124.6 loci) must therefore have been preserved in duplicate convergently.

Using the same approach as in Method 1 (above) it is now possible to calculate how many loci were preserved in duplicate convergently in excess of that expected by chance. At the time of divergence between *S. cerevisiae* and *K. polysporus* 1808 loci (79% of the original total) were still duplicated and in the **U** ('undecided') state and 16.24% ((169+124.6)/1808 = 0.1624) of these were preserved in duplicate in each lineage after this time. We therefore expect 0.1624 * 0.1624 * 1808 = 47.7 loci to be preserved in duplicate in both lineages by chance alone. The total expected number of shared duplicates is therefore 92.1 loci (44.4 on the shared branch and 47.7 due to sampling) and the ratio of the observed to the expected is 169/92.1 = 1.84-fold. This represents an excess of 76.9 loci and suggests that a significant number of loci have been independently preserved in duplicate in *S. cerevisiae* and *K. polysporus*.

We tested whether the observed excess of shared ohnolog pairs was statistically significant using a hypergeometric probability. Considering only the 124.6 duplicate pairs inferred to have been preserved in duplicate convergently on the *S. cerevisiae* and *K. polysporus* lineages, we calculated the probability of observing this number or greater by chance given that 293.6 (= 124.6 + 169) duplicate pairs were preserved independently on each lineage and that 1808 duplicate pairs in total were available for preservation. The probability of observing this by chance is effectively zero ($P = 2.4 \times 10^{-33}$).

Appendix XIV Modeling the resolution of genome duplication

We developed a mathematical model of the loss or fixation of duplicated genes after WGD. This model is significantly more powerful and flexible than the approach we took in ref. (Scannell *et al.*, 2006a). Our model assumes that the observed genomic sequences are related to each other by an (unknown) bifurcating phylogenetic topology. It attempts to explain the observed frequencies of duplicates and of the shared or divergent losses of duplicates among the five genomes (*K. polysporus, S. castellii, C. glabrata, S. cerevisiae* and *S. bayanus*). Thus, we create an 'alignment' of five species. Each site in this alignment represents an ancestral locus was duplicated in the WGD. For each species, we used YGOB to determine if that locus is still duplicated (state D_O) or had lost the first copy of the duplicate pair (S_1) or the second copy (S_2). We excluded from our analysis sites where both duplicates appear to have been lost. We use YGOB to assign consistent definitions of S_1 and S_2 across the five species (Byrne and Wolfe, 2005, Scannell *et al.*, 2006a).

Our model (DL-SUBF) is in the spirit of likelihood models of character state evolution proposed by Lewis (Lewis, 2001). We assume that a pair of loci formed by WGD can be in one of 6 possible states, and that transitions between states are possible (with rates specified by the parameters α,β and γ) as summarized in Figure 5.16A.

Initially all genes are assumed to be duplicated (i.e. $P(U|t_0)$=1.0). The instantaneous transition probabilities given in Figure 5.16 were used to construct a system of linear differential equations, which were symbolically solved using *Mathematica* 5.2. The probability of observing each state for each ancestral locus after a given time t is thus given by:

$$P(U \rightarrow U \mid t) = e^{-(2+2\beta+\gamma)\alpha t}$$

$$P(U \rightarrow S_1 \mid t) = \frac{(1+2\beta)\cdot(1+\beta+\gamma) - (1+\beta)\cdot(1+\gamma)\cdot e^{-(2+2\beta+\gamma)\alpha t} - \beta(2+2\beta+\gamma)\cdot e^{-(1+\gamma)\alpha t}}{(1+2\beta)\cdot(1+\gamma)\cdot(2+2\beta+\gamma)}$$

$$P(U \rightarrow F \mid t) = \frac{\gamma\cdot\left((1+2\beta)\cdot(1+2\beta+\gamma) - (1+\gamma)\cdot e^{-(2+2\beta+\gamma)\alpha t} - 2\beta\cdot(2+2\beta+\gamma)\cdot e^{-(1+\gamma)\alpha t}\right)}{(1+2\beta)\cdot(1+\gamma)\cdot(2+2\beta+\gamma)}$$

$$P(U \rightarrow C_1 \mid t) = \frac{\beta\cdot\left(e^{-(1+\gamma)\alpha t} - e^{-(2+2\beta+\gamma)\alpha t}\right)}{1+2\beta}$$

$$P(C_1 \rightarrow C_1 \mid t) = e^{-(1+\gamma)\alpha t}$$

$$P(C_1 \rightarrow S_1 \mid t) = \frac{1 - e^{-(1+\gamma)\alpha t}}{1+\gamma}$$

$$P(C_1 \rightarrow F \mid t) = \frac{\gamma\cdot\left(1 - e^{-(1+\gamma)\alpha t}\right)}{1+\gamma}$$

Here *U* is a state where both duplicates are present and redundant (meaning that the loss of one or the other is selectively equivalent). When one copy of a duplicate is lost, the locus transitions to state S_1 or S_2. Note that these two states are completely symmetrical and hence that equations for state S_2 are not shown above. Duplicates can also be fixed: once in state *F* neither copy of a duplicate pair can be lost.

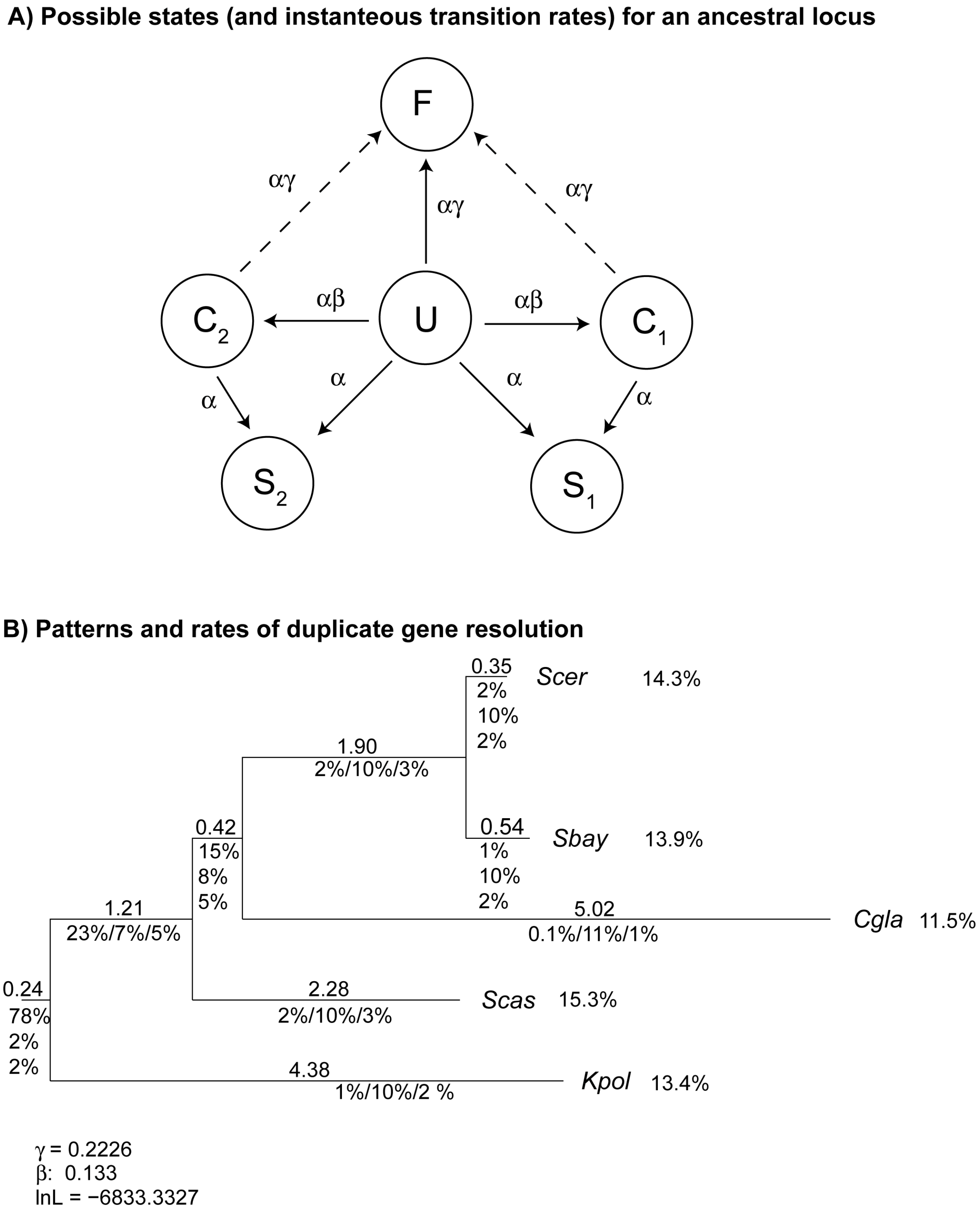


Figure 5.16 Modeling the resolution of WGD. **(A)** The 6 model states and the rates of the possible transitions between them (see equations above). (**B)** Maximum likelihood phylogeny for the 5 species under this model inferred from 2299 conservative sites identified by YGOB. Numbers above branches are branch lengths (see text). Numbers below the branches are the percentages of the original duplicate pairs that are in states *U*, *F*, and C_1+C_2, respectively.

Our previous analysis suggested that there is an excess of convergent losses of duplicated genes (cases where two species share a loss pattern than cannot be attributed to common ancestry) (Scannell *et al.*, 2006a). We incorporated this feature into the model by creating states C_1 and C_2. Genes in these states are duplicated, but if a loss is to occur from this state it will always be to state S_1 or S_2, respectively. Such loci can alternatively become fixed. Thus, an initial partial loss of function mutation in the second copy of a gene predisposes that duplicate to be lost (entering state C_1). If further mutations accumulate, that copy is lost (transition to state S_1). If the first copy instead undergoes a partial loss of function, the two copies can be fixed by subfunctionalization, with each performing a subset of the ancestral functions (state F). Because these convergent duplicated states can be inherited, they allow us to explain the observation of convergent losses. Note that states F, C_1, C_2, and U are degenerate with respect to our data – we can only identify observed duplicate gene pairs D_O, so for each such pair we sum over the likelihood of the four possible duplicated states in the model. By partitioning states S_1 and S_2 into separate states for convergent and non-convergent losses, we can also infer what proportion of losses along any branch are convergent. A similar approach can be taken for the fixed duplicates to determine if they were directly fixed from state U or by first passing through states C_1 or C_2.

Given a bifurcating phylogenetic topology τ, values of β and γ and of the $2n-1$ branch lengths (αt above, where n is the number of taxa in our analysis), we can calculate the likelihood of the data using our own implementation of the tree-transversal algorithm of Felsenstein (Felsenstein, 1981). We then use standard numerical optimization (Press *et al.*, 1992) to find maximum likelihood estimates of the branch lengths and of β and γ. Note that because this model is not time-reversible, our inferences are performed on rooted topologies. In practice, we infer the phylogenetic relationship of the genomes in question with an exhaustive search across all possible topologies τ, retaining the topology with the highest likelihood. The results of applying this model to our data are shown in Figure 5.16B. Above each branch is given the branch length in terms of $x = (2 + 2\beta + \gamma)\alpha t$. Taking e^{-x} gives the probability of a duplicate gene remaining in state U along that branch. Below each branch are the percentages of the total set of genes duplicated at WGD that are still in the duplicated states U, F, and C_1+C_2, respectively. We simulate data under the inferred maximum likelihood tree to estimate the statistical error associated with the model parameters. Doing this constitutes an implicit hypothesis test of the topology shown in

Figure 5.16B. We find that this topology is strongly supported (99% confidence intervals do not overlap zero on any branch).

Degenerate forms of the above model can also be constructed so as to disallow certain evolutionary possibilities. Thus, duplicate fixation can be forbidden by setting $\gamma = 0$ (DL-C); likewise convergence by setting $\beta = 0$ (DL-F). Subfunctionalization can be precluded by letting γ and β be nonzero but forbidding transitions from C_1 and C_2 to F (*i.e.*, removing the dashed lines in Figure 5.16A, DL-CF). Of course fixation and convergence can also be simultaneously disallowed by setting both γ and β to zero (DL). By simulating data under these more simple models, we can test the hypotheses that duplicate fixation, convergence, and subfunctionalization are statistically significant effects. In all four cases (alternative and null models DL-F and DL, DL-C and DL, DL-CF and DL-F, and DL-SUBF and DL-CF, respectively), we find the alternative models with these effects fit the data significantly better than the null models (**P** < 0.001).

The model DL-SUBF assumes that the instantaneous rate of duplicate loss and fixation from states C_1 and C_2 (C_x) is the same as that rate from state U. It is possible to relax this assumption, allowing more or less rapid rates of this processes after entering state C_x. Upon applying this more complex model (DL-SUBF-2) we found that while it offered a higher likelihood than the DL-SUBF model (2ΔlnL=135.8), it was not significantly better than a model where the *U*-*F* transition was forbidden (DL-SUBF-2 vs. DL-SUBF-C, 2ΔlnL =1.4). Effectively, the DL-SUBF-C model thus requires all fixations to pass through states C_x. Both model DL-SUBF-C and model DL-SUBF-2 have transition probabilities that are significantly more complicated than DL-SUBF. Moreover, the improvements seen using these two models are no longer significant if *C. glabrata* and *S. bayanus* are removed from the analysis (data not shown). For reasons of clarity we have therefore chosen to report our results in terms of the simpler model. We note that our general conclusions are not altered by using these more complex models.

One hypothesis of interest is whether the whole-genome duplication observed in *K. polysporus* is actually the same event as those seen in the other four species. Were they different events, the length of the root branch, which separates *K. polysporus* from the other four taxa, would have length 0. We can test if the inferred length of this branch in Figure 5.16 is significantly different from zero by simulating data under the hypothesis that this branch has length zero and using a likelihood ratio test to compare the null to the alternative hypothesis. When we do so, we find strong evidence that this branch has non-zero length and hence that all five species underwent the same duplication event (**P** < 0.001).

Our analysis uses YGOB (Byrne and Wolfe, 2005) to infer orthology between the duplicated regions of these five genomes. There are occasions, however, when this inference can be problematic. In some cases, data may be missing from the genome sequence of one organism, making it impossible to determine whether a particular WGD locus is retained in duplicate in that species. There are also cases where single copy genes in a species cannot be confidently assigned as either orthologs or a paralogs of the corresponding WGD loci in the other species (for instance if that gene resides alone on its contig). We omit all such ambiguous sites in the estimates presented here. However, adding data where one or more species is ambiguous at certain sites produces essentially identical results (data not shown).

The problem of determining whether single copy genes in one species are true orthologs to their homologs in other species is especially pronounced in *K. polysporus* due to this species' early divergence from the other four species. Given this fact, it is possible that our scoring approach using YGOB could tend to over or under-estimate the proportion of shared gene losses at the root of the tree in Figure 5.16B above (further details are given in Appendix XI and Appendix XII). We can test whether this problem is misleading us by discarding the information as to which copy (S_1 or S_2) is present in *K. polysporus* and treating all single copy loci in this species as ambiguous with respect to the remaining four species (S_x). When we re-estimate the model parameters by maximum likelihood, the probability of each single copy site in *K. polysporus* is the sum of the probability for states S_1 and S_2 above. Doing so actually increases the inferred number of shared losses on the root branch of the tree in Figure 5.16B, suggesting that our original analysis is conservative in its estimate of the degree of shared ancestry between *S. cerevisiae* and *K. polysporus*. To test whether we would observe such a long root branch were the genome duplication not shared between the five species, we simulated data under the assumption of no shared ancestry between *K. polysporus* and the other four taxa. We then discarded information on which single copy genes were present for *K. polysporus* (creating the same ambiguities as above) and optimized the resulting datasets under the assumption of a zero length root branch and without this constraint. None of these simulated datasets showed an improvement in likelihood after constraint relaxation that was as large as seen in the real data (**P** < 0.001). This is strong evidence that our scoring approach has not misled us into inferring a single duplication event. It is also an encouraging signal that many of our other conclusions would be robust to incorrect tracking.

Appendix XV Direct comparison of representative K_A values between convergently and divergently resolved loci

To exclude the possibility that the result shown in Figure 3.4 could be caused by a general trend towards resolving slower-evolving loci at later time points, we tested whether loci undergoing convergent loss at later time points tended to be biased towards slower-evolving loci, in the same way as loci undergoing RGL are biased.

We assembled sets of loci at which either convergent gene loss (orthologs lost in two independent lineages; single-copy orthologs retained) or divergent gene loss (paralogs lost in two independent lineages; single-copy paralogs retained) have occurred between *S. cerevisiae* and *K. polysporus*. We excluded the possibility that loci in our convergent gene loss dataset were products of a single gene loss on a shared branch by requiring that the missing gene copy be still present in either *S. castellii* or *C. glabrata.* Although divergent gene loss at an ancestrally duplicated locus cannot be explained by a single gene loss on a shared branch, we imposed the same phylogenetic criterion when selecting convergently and divergently resolved loci so the two datasets could be compared directly.

In brief, we used YGOB to select loci at which one gene copy from each duplicate clade was retained in at least one of *S. cerevisiae*, *C. glabrata* or *S. castellii* (Figure 5.17 panel 1). All loci selected on this basis must have been retained in duplicate on the lineage leading to *S. cerevisiae* until at least the divergence of *S. castellii* (t_2 in panel 1). We then discarded any loci at which duplicates have been retained in either *S. cerevisiae* or *K. polysporus* (panel 2) and partitioned the remaining loci into those at which single-copy orthologs (167 loci) and single-copy paralogs (111 loci) were retained between *S. cerevisiae* and *K. polysporus* (panel 3).

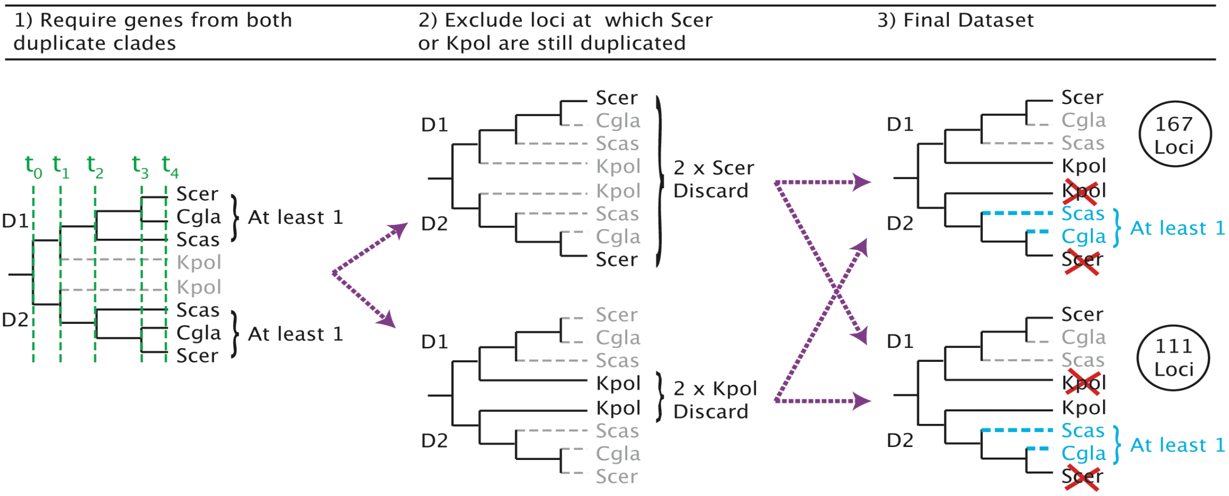


Figure 5.17 Method of selection of sets of genes that have either been convergently or divergently resolved between *S. cerevisiae* and *K. polysporus*. Because all of these loci were retained in duplicate on the *S. cerevisiae* lineage until at least the divergence of *S. castellii*, they must all have involved at least two independent gene losses: one on the *K. polysporus* lineage in the interval between t_1 and t_4 and one on the *S. cerevisiae* lineage between t_2 and t_4.

For each locus in both datasets we calculated 'representative' K_A values between the orthologous genes in *K. lactis* and *A. gossypii, $K_{A(Klac\text{-}Agos)}$* (Scannell *et al.*, 2006a), because this provides a measure of the intrinsic rate of evolution of the gene unaffected by any possible rate acceleration after gene duplication (Davis and Petrov, 2004). We find that the median $K_{A(Klac\text{-}Agos)}$ in single-copy orthologs is significantly greater than that amongst single-copy paralogs (0.3732 vs. 0.3315; $P = 0.006$ by one-sided Wilcoxon rank-sum test), indicating that RGL occurs preferentially at slow-evolving loci.

Although we used the same procedure to select loci for our single-copy ortholog and single-copy paralog datasets, it is possible that these datasets may be enriched for loci with different patterns of gene loss in *S. castellii* and *C. glabrata* and that it may therefore not be appropriate to compare them directly. To exclude this possibility we paired loci between our single-copy ortholog and single-copy paralog datasets whose patterns of gene loss were identical in all species except that the single-copy ortholog had retained the same (syntenic ortholog) gene copy in both *S. cerevisiae* and *K. polysporus* while the single copy paralog had retained alternative gene copies in these species. This produced 106 locus pairs whose only systematic difference is that one locus in each pair had lost orthologous gene copies independently in *S. cerevisiae* and *K. polysporus* and the second locus had independently lost paralogous gene copies. We performed this matching procedure 100 times and found that in 79 of 100 replicates, the $K_{A(Klac\text{-}Agos)}$ values for single-copy paralogs were significantly lower ($P < 0.05$ by one-sided Wilcoxon rank-sum test) than those for single-copy orthologs.

These results are consistent with hypothesis that RGL is more likely to occur at loci where duplicates are functionally interchangeable (Scannell *et al.*, 2006a) and that this condition is more likely to be met by slowly evolving loci.

Appendix XVI The proportion of partisan gene losses increases on successive branches after the WGD

As shown in Figure 3.2 the percentage of partisan losses (**C→S** transitions) as a fraction of all gene loss events (**U→S** and **C→S** transitions) inferred by our model of gene loss increases on successive branches after the WGD. It rises from 1% on the earliest branch after the WGD to 40% on the terminal *S. cerevisiae* branch. Because neutral losses (**U→S** transitions) arise from state **U** (which initially contains 100% of loci and must therefore decrease) while partisan losses arise from state **C** (which initially contains 0% of loci and must therefore decrease), we wanted to exclude the possibility that the increasing prevalence of partisan loss relative to neutral loss was a trivial consequence of the structure of our model. We therefore used a method that does not rely on the model to estimate the proportions of neutral and partisan gene losses at two different timepoints after the WGD and verified that the fraction of partisan gene losses is significantly higher at the later timepoint.

A simple method to estimate the proportion of neutral and partisan losses using gene loss data from post-WGD genome trios is described in ref. (Scannell *et al.*, 2006a). Because any three post-WGD genomes can be resolved into a pair of ingroup genomes and a single outgroup genome, it is possible to identify loci that have been returned to single-copy independently in the outgroup genome and one of the in-group genomes by selecting loci that are still duplicated in the second ingroup genome (See Figure 2.2, Classes 2C – 2F). We can then compare the proportions of loci at which orthologous and paralogous gene copies (using synteny information to distinguish syntenic orthologs from non-syntenic paralogs) have been retained between the single-copy outgroup and ingroup genomes. Moreover, since any excess of orthologous over paralogous gene losses must be attributable to events on the shared evolutionary branch between the WGD and the divergence of the three species of interest, we can examine the effect of the time since duplicate gene divergence by selecting genome trios whose common ancestor existed at different timepoints after the WGD.

We used a genome trio composed of *(K. polysporus, (S. castellii, S. cerevisiae))* and one composed of *(S. castellii, (C. glabrata, S. cerevisiae))* to identify sets of genes that were resolved independently in two lineages after the divergence of *K. polysporus* (Kpol-Trio) or *S. castellii* (Scas-Trio) from the *S. cerevisiae* lineage respectively. Following exclusion of any loci that did not satisfy the synteny quality criteria required by the Yeast Gene Order Browser (Byrne and Wolfe, 2005), we obtained 130 loci from the Kpol-Trio and 83 loci from the Scas-Trio for which independent resolution of gene duplicates in two lineages could be inferred with confidence. As can be seen from Table 5.12, the proportion of orthologous and paralogous gene losses is close to equal for the Kpol-Trio (77 orthologous gene losses compared to 53 paralogous gene losses in the combined dataset) but very skewed for the Scas-Trio (65 orthologous gene losses compared to 18 paralogous gene losses in the combined dataset). These are significantly different in a chi-squared test of homogeneity (P = 0.006) indicating that the proportion of orthologous and paralogous gene losses depends on the time since the WGD. In addition, the direction of the change in the relative proportions of orthologous and paralogous gene losses (increase in the former relative to the latter at the later timepoint) is consistent with the idea that proportion of orthologous gene losses (and hence partisan losses; Table 5.12) increases with time since the duplication. These data indicate that the conclusion that the proportion of partisan gene losses is higher at later timepoints is not solely due to the structure of our likelihood model but is a property of the data.

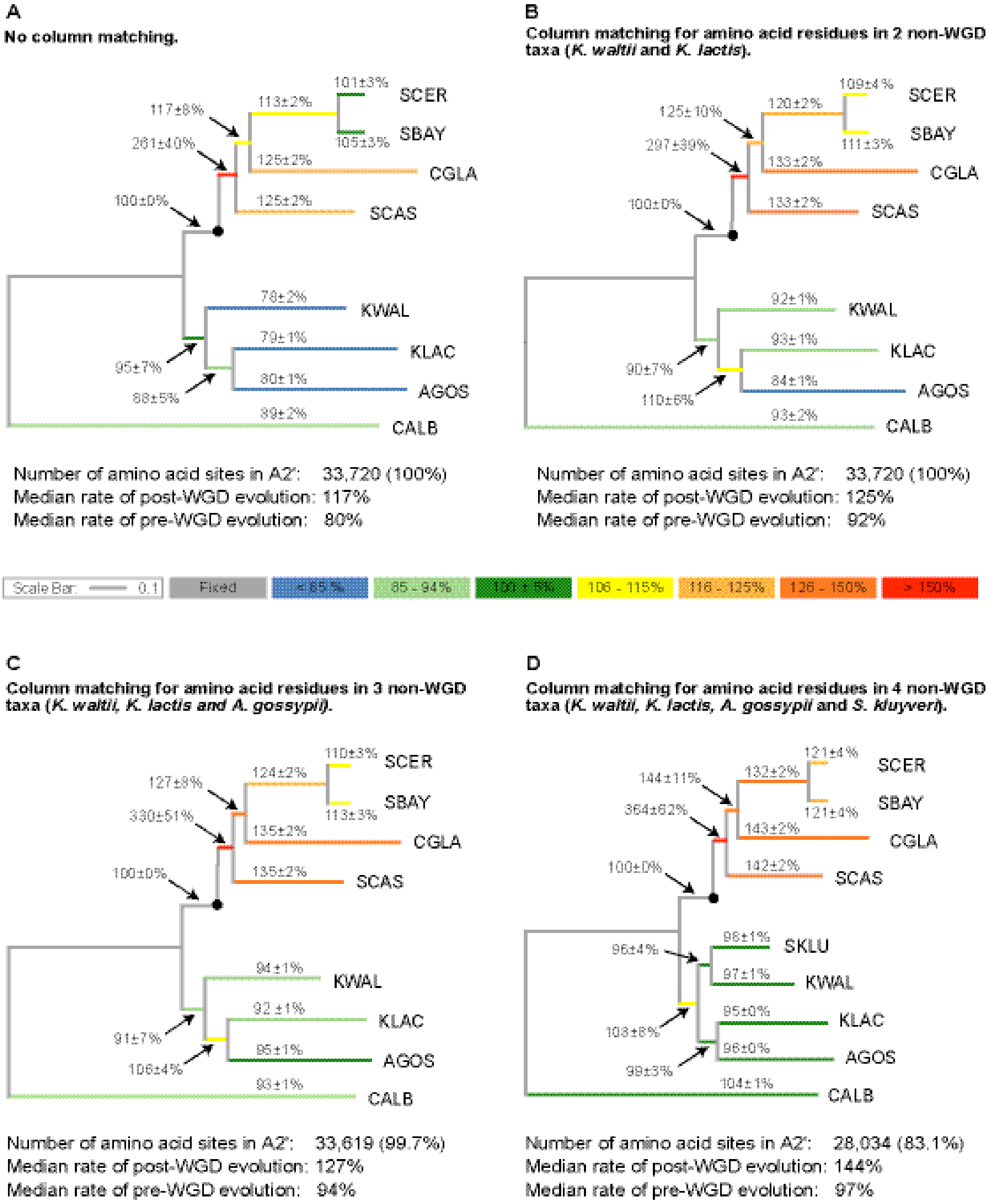
Table 5.12 Estimated percentage of partisan gene losses at two different timepoints based on counts of orthologous and paralogous gene losses from two genome trios.

Outgroup	**Ingroup**		**Gene Losses**					
Single-copy	**Single-copy**	**Double-copy**	**Orthologous losses**	**Paralogous losses**	**Total**	**Neutral losses***	**Partisan losses***	**% Partisan losses**
Kpol	Scer	Scas	47	28	75	56	19	25.3%
Kpol	Scas	Scer	30	25	55	50	5	9.1%
	Combined		77	53	130	106	24	18.5%
Scass	Scer	Cgla	26	9	35	18	17	48.6%
Scas	Cgla	Scer	39	9	48	18	30	62.5%
	Combined		65	18	83	36	47	56.6%

* The number of neutral gene losses was estimated as twice the number of paralogous gene losses and the number of partisan gene losses was calculated as the number of orthologous gene losses minus the number of paralogous gene losses. See ref. (Scannell *et al.*, 2006a) for justification. Note that because of the method by which these loci were selected (duplicates were required in at least one species) the proportions of orthologous and paralogous (or neutral and partisan) losses are not the same as those estimated by the

model (Figure 3.2). The latter are based on a much larger and less biased dataset and should be more accurate.

Appendix XVII Dependency of column-matching procedure on the number of non-WGD taxa


A
No column matching.

Number of amino acid sites in A2': 33,720 (100%)
Median rate of post-WGD evolution: 117%
Median rate of pre-WGD evolution: 80%

B
Column matching for amino acid residues in 2 non-WGD taxa (*K. waltii* and *K. lactis*).

Number of amino acid sites in A2': 33,720 (100%)
Median rate of post-WGD evolution: 125%
Median rate of pre-WGD evolution: 92%

C
Column matching for amino acid residues in 3 non-WGD taxa (*K. waltii, K. lactis and A. gossypii*).

Number of amino acid sites in A2': 33,619 (99.7%)
Median rate of post-WGD evolution: 127%
Median rate of pre-WGD evolution: 94%

D
Column matching for amino acid residues in 4 non-WGD taxa (*K. waltii, K. lactis, A. gossypii* and *S. kluyveri*).

Number of amino acid sites in A2': 28,034 (83.1%)
Median rate of post-WGD evolution: 144%
Median rate of pre-WGD evolution: 97%

Figure 5.18 Effect of the number of non-WGD species used for column-matching on the inferred rates of protein sequence evolution in super-alignment A2' relative to A1'. Panels A, B, C and D show the results of column-matching using 0, 2, 3, and 4 non-WGD species respectively. Branch lengths indicate the length of a branch in tree T2', expressed as a percentage of the corresponding

branch length in T1', and are the means (± s.d.) of 100 bootstrap replicates. The median values for post-WGD and non-WGD species are shown below each panel, with the number of columns from A2 that matched columns in A1 and so were retained to form super-alignment A2' (see *Methods*). In (A), no column-matching was done; 33,720 columns were randomly sampled from A1 to produce A1', and all columns in A2 were retained to form A2'. To reduce computation time, the trees in this figure were produced using the WAG+G(8)+F model, in contrast to Figure 1 of the main text which used the WAG+G(8)+I+F; this change accounts for the minor differences in branch lengths between panel C and Figure 1B, both of which use column-matching on three non-WGD species.

Appendix XVIII Accelerated evolution of double-copy sequences is not an artifact of the column-matching procedure

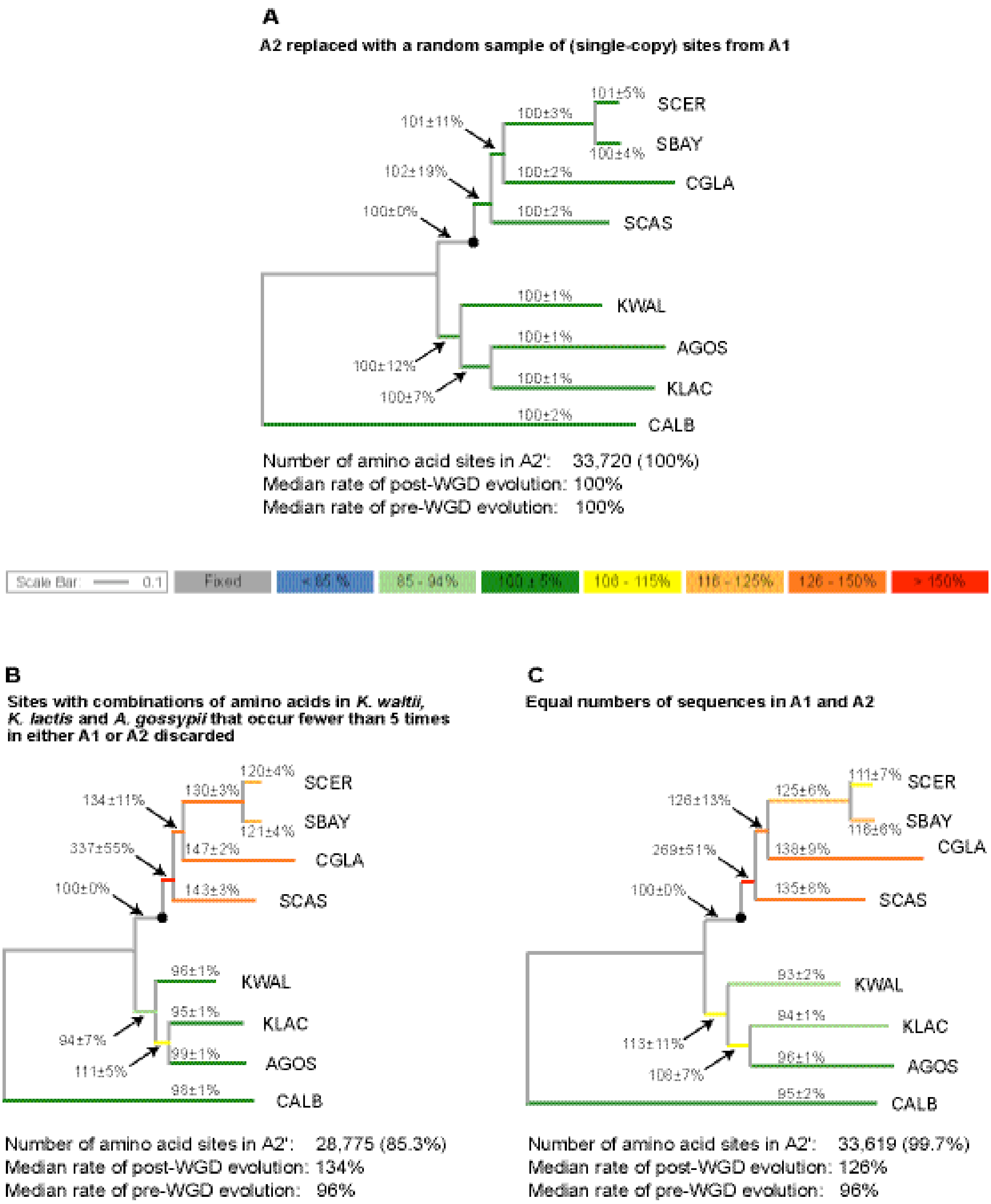

A
A2 replaced with a random sample of (single-copy) sites from A1
101±5%
SCER
100±3%
101±11%
100±4%
SBAY
102±19%
100±2%
CGLA
100±0%
100±2%
SCAS
100±1%
KWAL
100±1%
AGOS
100±12%
100±1%
100±7%
KLAC
100±2%
CALB
Number of amino acid sites in A2': 33,720 (100%)
Median rate of post-WGD evolution: 100%
Median rate of pre-WGD evolution: 100%
Scale Bar: 0.1
Fixed
< 85 %
85 - 94%
100 ± 5%
106 - 115%
116 - 125%
126 - 150%
> 150%
B
Sites with combinations of amino acids in *K. waltii*, *K. lactis* and *A. gossypii* that occur fewer than 5 times in either A1 or A2 discarded
120±4%
SCER
130±3%
134±11%
SBAY
121±4%
337±55%
147±2%
CGLA
100±0%
143±3%
SCAS
96±1%
KWAL
95±1%
KLAC
94±7%
99±1%
AGOS
111±5%
98±1%
CALB
Number of amino acid sites in A2': 28,775 (85.3%)
Median rate of post-WGD evolution: 134%
Median rate of pre-WGD evolution: 96%
C
Equal numbers of sequences in A1 and A2
111±7%
SCER
125±6%
126±13%
SBAY
116±6%
269±51%
138±9%
CGLA
100±0%
135±8%
SCAS
93±2%
KWAL
94±1%
KLAC
113±11%
96±1%
AGOS
108±7%
95±2%
CALB
Number of amino acid sites in A2': 33,619 (99.7%)
Median rate of post-WGD evolution: 126%
Median rate of pre-WGD evolution: 96%

Figure 5.19 The construction of the super-alignments and the interpretation of results are given in the main text. Branch-lengths, coloring and other details are as for Figure 1B. (A) A random sample of single-copy sites was substituted for A2. (B) Sites with combinations of amino acids in

K. waltii, K. lactis and *A. gossypii* that occur fewer than five times in either A1 or A2 were discarded. (C) Sequences corresponding to one of the two duplicate clades were removed from A2.

References

Abascal, F., Zardoya, R. & Posada, D. (2005) Prottest: selection of best-fit models of protein evolution. *Bioinformatics*, doi:10.1093/bioinformatics/bti263.

Adams, K. L., Daley, D. O., Qiu, Y.-L., Whelan, J. & Palmer, J. D. (2000) Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. *Nature,* 408**,** 354-357.

Amores, A., Force, A., Yan, Y. L., Joly, L., Amemiya, C., Fritz, A., Ho, R. K., Langeland, J., Prince, V., Wang, Y. L., Westerfield, M., Ekker, M. & Postlethwait, J. H. (1998) Zebrafish hox clusters and vertebrate genome evolution. *Science,* 282**,** 1711-4.

Bailey, J. A. & Eichler, E. E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet,* 7**,** 552-64.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science,* 290**,** 972-7.

Barns, S. M., Lane, D. J., Sogin, M. L., Bibeau, C. & Weisburg, W. G. (1991) Evolutionary relationships among pathogenic *Candida* species and relatives. *J Bacteriol,* 173**,** 2250-2255.

Benovoy, D. & Drouin, G. (2006) Processed pseudogenes, processed genes, and spontaneous mutations in the Arabidopsis genome. *J Mol Evol,* 62**,** 511-22.

Betran, E., Thornton, K. & Long, M. (2002) Retroposed new genes out of the X in Drosophila. *Genome Res,* 12**,** 1854-9.

Blanc, G., Hokamp, K. & Wolfe, K. H. (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res,* 13**,** 137-144.

Blanc, G. & Wolfe, K. H. (2004a) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell,* 16**,** 1679-1691.

Blanc, G. & Wolfe, K. H. (2004b) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell,* 16**,** 1667-78.

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.

Bon, E., Neuveglise, C., Lepingle, A., Wincker, P., Artiguenave, F., Gaillardin, C. & Casaregola, S. (2000) Genomic Exploration of the Hemiascomycetous Yeasts: 6. *Saccharomyces exiguus*. *FEBS Lett,* 487**,** 42-46.

Bond, U., Neal, C., Donnelly, D. & James, T. C. (2004) Aneuploidy and copy number breakpoints in the genome of lager yeasts mapped by microarray hybridisation. *Curr Genet,* 45**,** 360-370.

Bonnaud, B., Beliaeff, J., Bouton, O., Oriol, G., Duret, L. & Mallet, F. (2005) Natural history of the ERVWE1 endogenous retroviral locus. *Retrovirology,* 2**,** 57.

Brem, R. B., Storey, J. D., Whittle, J. & Kruglyak, L. (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature,* 436**,** 701-3.

Brunet, F. G., Crollius, H. R., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., Laudet, V. & Robinson-Rechavi, M. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol,* 23**,** 1808-16.

Butler, G., Kenny, C., Fagan, A., Kurischko, C., Gaillardin, C. & Wolfe, K. H. (2004) Evolution of the *MAT* locus and its Ho endonuclease in yeast species. *Proc Natl Acad Sci U S A,* 101**,** 1632-1637.

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.

Byrnes, J. K., Morris, G. P. & Li, W. H. (2006) Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. *Mol Biol Evol,* 23**,** 1136-43.

Cai, J. J., Woo, P. C., Lau, S. K., Smith, D. K. & Yuen, K. Y. (2006) Accelerated Evolutionary Rate May Be Responsible for the Emergence of Lineage-Specific Genes in Ascomycota. *J Mol Evol*.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., De Bono, B., Della Gatta, G., Di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., Mcwilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., et al. (2005) The transcriptional landscape of the mammalian genome. *Science,* 309**,** 1559-63.

Casaregola, S., Lepingle, A., Bon, E., Neuveglise, C., Nguyen, H., Artiguenave, F., Wincker, P. & Gaillardin, C. (2000) Genomic Exploration of the Hemiascomycetous Yeasts: 7. *Saccharomyces servazzii*. *FEBS Lett,* 487**,** 47-51.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol,* 17**,** 540-52.

Chambers, S. R., Hunter, N., Louis, E. J. & Borts, R. H. (1996) The mismatch repair system reduces meiotic homeologous recombination and stimulates recombination-dependent chromosome loss. *Mol Cell Biol,* 16**,** 6110-20.

Chien, C. T., Buck, S., Sternglanz, R. & Shore, D. (1993) Targeting of SIR1 protein establishes transcriptional silencing at HM loci and telomeres in yeast. *Cell,* 75**,** 531-41.

Chopra, V. S. & Mishra, R. K. (2005) To SIR with Polycomb: linking silencing mechanisms. *Bioessays,* 27**,** 119-21.

Ciccarelli, F. D., Von Mering, C., Suyama, M., Harrington, E. D., Izaurralde, E. & Bork, P. (2005) Complex genomic rearrangements lead to novel primate gene function. *Genome Res,* 15**,** 343-51.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A. & Johnston, M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science,* 301**,** 71-76.

Cliften, P. F., Fulton, R. S., Wilson, R. K. & Johnston, M. (2006) After the duplication: gene loss and adaptation in Saccharomyces genomes. *Genetics,* 172**,** 863-72.

Cliften, P. F., Hillier, L. W., Fulton, L., Graves, T., Miner, T., Gish, W. R., Waterston, R. H. & Johnston, M. (2001) Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res,* 11**,** 1175-86.

Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet,* 26**,** 183-6.

Comai, L. (2005) The advantages and disadvantages of being polyploid. *Nat Rev Genet,* 6**,** 836-46.

Conant, G. C. & Wagner, A. (2003) Asymmetric sequence divergence of duplicate genes. *Genome Res,* 13**,** 2052-8.

Conant, G. C., Wagner, G. P. & Stadler, P. F. (2006) Modeling amino acid substitution patterns in orthologous and paralogous genes. *Mol Phylogenet Evol.*

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

Coyne, J. A. & Orr, H. A. (2004) *Speciation,* Sunderland, MA, Sinauer.

Cusack, B. P. & Wolfe, K. H. (2006) Not born equal: Increased rate asymmetry in relocated

and retrotransposed rodent gene duplicates. *Mol Biol Evol,* Submitted.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W. & Steinmetz, L. M. (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A,* 103**,** 5320-5325.

Davis, J. C. & Petrov, D. A. (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol,* 2**,** E55.

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

Delneri, D., Colson, I., Grammenoudi, S., Roberts, I. N., Louis, E. J. & Oliver, S. G. (2003) Engineering evolution to study speciation in yeasts. *Nature,* 422**,** 68-72.

Derisi, J. L., Iyer, V. R. & Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science,* 278**,** 680-686.

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

Dietrich, F. S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pohlmann, R., Luedi, P., Choi, S., Wing, R. A., Flavier, A., Gaffney, T. D. & Philippsen, P. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science,* 304**,** 304-307.

Dobzhansky, T. (1933) On the sterility of the interracial hybrids in *Drosophila pseudoobscura*. *Proc Natl Acad Sci U S A,* 19**,** 397-403.

Domergue, R., Castano, I., De Las Penas, A., Zupancic, M., Lockatell, V., Hebel, J. R., Johnson, D. & Cormack, B. P. (2005) Nicotinic acid limitation regulates silencing of *Candida* adhesins during UTI. *Science,* 308**,** 866-70.

Doolittle, W. F. (1999) Phylogenetic classification and the universal tree. *Science,* 284**,** 2124-9.

Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L. T., Wu, X., Reith, M., Cavalier-Smith, T. & Maier, U. G. (2001) The highly reduced genome of an enslaved algal nucleus. *Nature,* 410**,** 1091-6.

Drummond, D. A., Raval, A. & Wilke, C. O. (2006) A Single Determinant Dominates the Rate of Yeast Protein Evolution. *Mol Biol Evol,* 23**,** 327-337.

Dudley, A. M., Janse, D. M., Tanay, A., Shamir, R. & Church, G. M. (2005) A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol Syst Biol,* 1**,** 2005 0001.

Dujon, B. (1996) The yeast genome project: what did we learn? *Trends Genet,* 12**,** 263-270.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, 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., Boisramé, 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., Wésolowski-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.

Duret, L., Chureau, C., Samain, S., Weissenbach, J. & Avner, P. (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science,* 312**,** 1653-5.

Eddy, S. R., Mitchison, G. & Durbin, R. (1995) Maximum discrimination hidden Markov models of sequence consensus. *J Comput Biol,* 2**,** 9-23.

Ericson, E., Pylvanainen, I., Fernandez-Ricaud, L., Nerman, O., Warringer, J. & Blomberg, A. (2006) Genetic pleiotropy in Saccharomyces cerevisiae quantified by high-resolution phenotypic profiling. *Mol Genet Genomics,* 275**,** 605-14.

Esteban, P. F., Vazquez De Aldana, C. R. & Del Rey, F. (1999) Cloning and characterization of 1,3-beta-glucanase-encoding genes from non-conventional yeasts. *Yeast,* 15**,** 91-109.

Evans, B. J., Kelley, D. B., Melnick, D. J. & Cannatella, D. C. (2005) Evolution of RAG-1 in polyploid clawed frogs. *Mol Biol Evol,* 22**,** 1193-207.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res,* 8**,** 175-85.

Fabre, E., Muller, H., Therizols, P., Lafontaine, I., Dujon, B. & Fairhead, C. (2005) Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing and subtelomeres. *Mol Biol Evol,* 22**,** 856-873.

Fares, M. A., Byrne, K. P. & Wolfe, K. H. (2006) Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. *Mol Biol Evol,* 23**,** 245-53.

Fay, J. C. & Benavides, J. A. (2005a) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet***,** in press.

Fay, J. C. & Benavides, J. A. (2005b) Hypervariable noncoding sequences in Saccharomyces cerevisiae. *Genetics,* 170**,** 1575-87.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol,* 17**,** 368-376.

Fernandes, L., Rodrigues-Pousada, C. & Struhl, K. (1997) Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. *Mol Cell Biol,* 17**,** 6982-93.

Ferris, S. D. & Whitt, G. S. (1977) Loss of duplicate gene expression after polyploidisation. *Nature,* 265**,** 258-60.

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

Fischer, G., James, S. A., Roberts, I. N., Oliver, S. G. & Louis, E. J. (2000) Chromosomal evolution in *Saccharomyces*. *Nature,* 405**,** 451-4.

Fischer, G., Neuvéglise, C., Durrens, P., Gaillardin, C. & Dujon, B. (2001) Evolution of gene order in the genomes of two related yeast species. *Genome Res,* 11**,** 2009-19.

Force, A., Cresko, W. A., Pickett, F. B., Proulx, S. R., Amemiya, C. & Lynch, M. (2005) The origin of subfunctions and modular gene regulation. *Genetics,* 170**,** 433-46.

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.

Forsberg, H. & Ljungdahl, P. O. (2001) Genetic and biochemical analysis of the yeast plasma membrane Ssy1p-Ptr3p-Ssy5p sensor of extracellular amino acids. *Mol Cell Biol,* 21**,** 814-26.

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

Friedman, R. & Hughes, A. L. (2001) Gene duplication and the structure of eukaryotic genomes. *Genome Res,* 11**,** 373-381.

Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., Fitzhugh, W., Ma, L. J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C. B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M. A., Werner-Washburne, M., Selitrennikoff, C. P., Kinsey, J. A., Braun, E. L., Zelter, A., Schulte, U., Kothe, G. O., Jedd, G., Mewes, W., Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Mauceli, E., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, R. L., Perkins, D. D., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R. J., Osmani, S. A., Desouza, C. P., Glass, L., Orbach, M. J., Berglund, J. A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D. O., Alex, L. A., Mannhaupt, G., Ebbole, D. J., Freitag, M., Paulsen, I., Sachs, M. S., Lander, E. S., Nusbaum, C. & Birren, B. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature,* 422**,** 859-68.

Galagan, J. E., Calvo, S. E., Cuomo, C., Ma, L. J., Wortman, J. R., Batzoglou, S., Lee, S. I., Basturkmen, M., Spevak, C. C., Clutterbuck, J., Kapitonov, V., Jurka, J., Scazzocchio, C., Farman, M., Butler, J., Purcell, S., Harris, S., Braus, G. H., Draht, O., Busch, S., D'enfert, C., Bouchier, C., Goldman, G. H., Bell-Pedersen, D., Griffiths-Jones, S., Doonan, J. H., Yu, J., Vienken, K., Pain, A., Freitag, M., Selker, E. U., Archer, D. B., Penalva, M. A., Oakley, B. R., Momany, M., Tanaka, T., Kumagai, T., Asai, K., Machida, M., Nierman, W. C., Denning, D. W., Caddick, M., Hynes, M., Paoletti, M., Fischer, R., Miller, B., Dyer, P., Sachs, M. S., Osmani, S. A. & Birren, B. W. (2005a) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature,* 438**,** 1105-15.

Galagan, J. E., Henn, M. R., Ma, L. J., Cuomo, C. A. & Birren, B. (2005b) Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res,* 15**,** 1620-31.

Gerstein, A. C., Chun, H. J., Grant, A. & Otto, S. P. (2006) Genomic Convergence toward Diploidy in Saccharomyces cerevisiae. *PLoS Genet,* 2.

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.

Gibson, T. J. & Spring, J. (1998) Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet,* 14**,** 46-9.

Gilson, P. R., Su, V., Slamovits, C. H., Reith, M. E., Keeling, P. J. & Mcfadden, G. I. (2006) Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc Natl Acad Sci U S A,* 103**,** 9566-71.

Gimeno, C. J., Ljungdahl, P. O., Styles, C. A. & Fink, G. R. (1992) Unipolar cell divisions in the yeast *S. cerevisiae* lead to filamentous growth: regulation by starvation and RAS. *Cell,* 68**,** 1077-90.

Goffeau, A., Aert, R., Agostini-Carbone, M. L., Ahmed, A., Aigle, M., Alberghina, L., Allen, E., Alt-Mörbe, J., André, B., Andrews, S., Ansorge, W., Antoine, G., Anwar, R., Aparicio, A., Araujo, R., Arino, J., Arnold, F., Arroyo, J., Aviles, E. & Et Al. (1997) The Yeast Genome Directory. *Nature,* 387 (Suppl.)**,** 5-105.

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, 563-567.

Gogarten, J. P. & Townsend, J. P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol,* 3**,** 679-87.

Gojkovic, Z., Knecht, W., Zameitat, E., Warneboldt, J., Coutelis, J. B., Pynyaha, Y., Neuveglise, C., Moller, K., Loffler, M. & Piskur, J. (2004) Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts. *Mol Genet Genomics,* 271**,** 387-93.

Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science,* 185**,** 862-4.

Greig, D., Borts, R. H., Louis, E. J. & Travisano, M. (2002a) Epistasis and hybrid sterility in *Saccharomyces*. *Proc R Soc Lond B Biol Sci,* 269**,** 1167-71.

Greig, D., Louis, E. J., Borts, R. H. & Travisano, M. (2002b) Hybrid speciation in experimental populations of yeast. *Science,* 298**,** 1773-5.

Greig, D., Travisano, M., Louis, E. J. & Borts, R. H. (2003) A role for the mismatch repair system during incipient speciation in Saccharomyces. *J Evol Biol,* 16**,** 429-37.

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.

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.

Guindon, S. & Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol,* 52**,** 696-704.

Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N., Van Helden, J., Lemer, C., Richelles, J., Wodak, S. J., Garcia-Martinez, J., Perez-Ortin, J. E., Michael, H., Kaps, A., Talla, E., Dujon, B., Andre, B., Souciet, J. L., De Montigny, J., Bon, E., Gaillardin, C. & Mewes, H. W. (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res,* 33**,** D364-D368.

Haber, J. E. (1998) Mating-type gene switching in *Saccharomyces cerevisiae*. *Annu Rev Genet,* 32**,** 561-99.

Haber, J. E. & Wolfe, K. H. (2005) Evolution and function of HO and VDE endonucleases in fungi. IN BELFORT, M., DERBYSHIRE, V., STODDARD, B. & WOOD, D. (Eds.) *Inteins and homing endonucleases.* Berlin, Springer-Verlag.

Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res,* 15**,** 1153-60.

Hall, C., Brachat, S. & Dietrich, F. S. (2005) Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot Cell,* 4**,** 1102-15.

Halligan, D. L. & Keightley, P. D. (2006) Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res,* 16**,** 875-84.

Hanada, K., Gojobori, T. & Li, W. H. (2006) Radical amino acid change versus positive selection in the evolution of viral envelope proteins. *Gene*.

Hansen, J. & Piskur, J. (2003) Fungi in brewing: biodiversity and biotechnology perspectives. IN ARORA, D. K. (Ed.) *Handbook of Fungal Biotechnology.* New York, Marcel Dekker.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E. & Young, R. A. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature,* 431**,** 99-104.

Hartl, D. L. & Clark, A. G. (1997) *Principles of Population Genetics,* Sunderland, MA, Sinauer.

He, X. & Zhang, J. (2005a) Gene complexity and gene duplicability. *Curr Biol,* 15**,** 1016-21.

He, X. & Zhang, J. (2005b) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*.

Heck, J. A., Argueso, J. L., Gemici, Z., Reeves, R. G., Bernard, A., Aquadro, C. F. & Alani, E. (2006) Negative epistasis between natural variants of the Saccharomyces cerevisiae MLH1 and PMS1 genes results in a defect in mismatch repair. *Proc Natl Acad Sci U S A,* 103**,** 3256-61.

Heckman, D. S., Geiser, D. M., Eidell, B. R., Stauffer, R. L., Kardos, N. L. & Hedges, S. B. (2001) Molecular evidence for the early colonization of land by fungi and plants. *Science,* 293**,** 1129-33.

Hedges, S. B. (2002) The origin and evolution of model organisms. *Nat Rev Genet,* 3**,** 838-49.

Hirschman, J. E., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hong, E. L., Livstone, M. S., Nash, R., Park, J., Oughtred, R., Skrzypek, M., Starr, B., Theesfeld, C. L., Williams, J., Andrada, R., Binkley, G., Dong, Q., Lane, C., Miyasato, S., Sethuraman, A., Schroeder, M., Thanawala, M. K., Weng, S., Dolinski, K., Botstein, D. & Cherry, J. M. (2006) Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome. *Nucleic Acids Res,* 34**,** D442-5.

Hittinger, C. T., Rokas, A. & Carroll, S. B. (2004) Parallel inactivation of multiple *GAL* pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci U S A,* 101**,** 14144-14149.

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

Hughes, A. L. & Friedman, R. (2003) Parallel evolution by gene duplication in the genomes of two unicellular fungi. *Genome Res,* 13**,** 794-9.

Hughes, T. R., Roberts, C. J., Dai, H., Jones, A. R., Meyer, M. R., Slade, D., Burchard, J., Dow, S., Ward, T. R., Kidd, M. J., Friend, S. H. & Marton, M. J. (2000) Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat Genet,* 25**,** 333-7.

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.

Hull, C. M. & Johnson, A. D. (1999) Identification of a mating type-like locus in the asexual pathogenic yeast Candida albicans. *Science,* 285**,** 1271-5.

Hull, C. M., Raisner, R. M. & Johnson, A. D. (2000) Evidence for mating of the "asexual" yeast Candida albicans in a mammalian host. *Science,* 289**,** 307-10.

Hunter, N., Chambers, S. R., Louis, E. J. & Borts, R. H. (1996) The mismatch repair system contributes to meiotic sterility in an interspecific yeast hybrid. *Embo J,* 15**,** 1726-33.

Hunter, T. & Plowman, G. D. (1997) The protein kinases of budding yeast: six score and more. *Trends Biochem Sci,* 22**,** 18-22.

Ingold, C. T. & Hudson, H. J. (1961) *The Biology of Fungi,* London, Chapmann & Hall.

Ioshikhes, I. P., Albert, I., Zanton, S. J. & Pugh, B. F. (2006) Nucleosome positions predicted through comparative genomics. *Nat Genet,* 38**,** 1210-5.

Iwama, H. & Gojobori, T. (2004) Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network. *Proc Natl Acad Sci U S A,* 101**,** 17156-61.

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.

James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V., Cox, C. J., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J., Lumbsch, H. T., Rauhut, A., Reeb, V., Arnold, A. E., Amtoft, A., Stajich, J. E., Hosaka, K., Sung, G. H., Johnson, D., O'rourke, B., Crockett, M., Binder, M., Curtis, J. M., Slot, J. C., Wang, Z., Wilson, A. W., Schussler, A., Longcore, J. E., O'donnell, K., Mozley-Standridge, S., Porter, D., Letcher, P. M., Powell, M. J., Taylor, J. W., White, M. M., Griffith, G. W., Davies, D. R., Humber, R. A., Morton, J. B., Sugiyama, J., Rossman, A. Y., Rogers, J. D., Pfister, D. H., Hewitt, D., Hansen, K., Hambleton, S., Shoemaker, R. A., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Spotts, R. A., Serdani, M., Crous, P. W., Hughes, K. W., Matsuura, K., Langer, E., Langer, G., Untereiner, W. A., Lucking, R., Budel, B., Geiser, D. M., Aptroot, A., Diederich, P., Schmitt, I., Schultz, M., Yahr, R., Hibbett, D. S., Lutzoni, F., Mclaughlin, D. J., Spatafora, J. W. & Vilgalys, R. (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature,* 443**,** 818-22.

Jeffries, T. W. (2006) Engineering yeasts for xylose metabolism. *Curr Opin Biotechnol,* 17**,** 320-6.

Johnson, A. D. (1995) Molecular mechanisms of cell-type determination in budding yeast. *Curr Opin Genet Dev,* 5**,** 552-8.

Jones, C. D. & Begun, D. J. (2005) Parallel evolution of chimeric fusion genes. *Proc Natl Acad Sci U S A,* 102**,** 11373-8.

Jordan, I. K., Wolf, Y. I. & Koonin, E. V. (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol Biol,* 4**,** 22.

Jorgensen, E. M. & Mango, S. E. (2002) The art and design of genetic screens: caenorhabditis elegans. *Nat Rev Genet,* 3**,** 356-69.

Kassir, Y., Granot, D. & Simchen, G. (1988) *IME1*, a positive regulator gene of meiosis in *S. cerevisiae*. *Cell,* 52**,** 853-62.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K. C., Hallinan, J., Mattick, J., Hume, D. A., Lipovich, L., Batalov, S., Engstrom, P. G., Mizuno, Y., Faghihi, M. A., Sandelin, A., Chalk, A. M., Mottagui-Tabar, S., Liang, Z., Lenhard, B. & Wahlestedt, C. (2005) Antisense transcription in the mammalian transcriptome. *Science,* 309**,** 1564-6.

Katju, V. & Lynch, M. (2003) The structure and early evolution of recently arisen gene duplicates in the Caenorhabditis elegans genome. *Genetics,* 165**,** 1793-803.

Keightley, P. D. & Otto, S. P. (2006) Interference among deleterious mutations favours sex and recombination in finite populations. *Nature,* 443**,** 89-92.

Kellis, M., Birren, B. W. & Lander, E. S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature,* 428**,** 617-624.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature,* 423**,** 241-54.

Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A. & Voytas, D. F. (1998) Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome Res,* 8**,** 464-78.

Kim, S. H. & Yi, S. V. (2006) Correlated Asymmetry of Sequence and Functional Divergence Between Duplicate Proteins of Saccharomyces cerevisiae. *Mol Biol Evol*.

Kirchhausen, T. (2000) Three ways to make a vesicle. *Nat Rev Mol Cell Biol,* 1**,** 187-98.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. (2002) Selection in the evolution of gene duplications. *Genome Biol,* 3**,** RESEARCH0008.

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.

Kurtzman, C. P. (2003) Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Res,* 4**,** 233-45.

Kurtzman, C. P. & Robnett, C. J. (2003) Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. *FEMS Yeast Res,* 3**,** 417-32.

Kwast, K. E., Lai, L. C., Menda, N., James, D. T., 3rd, Aref, S. & Burke, P. V. (2002) Genomic analyses of anaerobically induced genes in *Saccharomyces cerevisiae*: functional roles of Rox1 and other factors in mediating the anoxic response. *J Bacteriol,* 184**,** 250-65.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., Mcewan, P., Mckernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., Mcmurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., Mcpherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001) Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

Landry, C. R., Oh, J., Hartl, D. L. & Cavalieri, D. (2006) Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene,* 366**,** 343-51.

Lee, J. M. & Sonnhammer, E. L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res,* 13**,** 875-82.

Leh-Louis, V., Wirth, B., Potier, S., Souciet, J. L. & Despons, L. (2004) Expansion and contraction of the DUP240 multigene family in Saccharomyces cerevisiae populations. *Genetics,* 167**,** 1611-9.

Lercher, M. J. & Hurst, L. D. (2006) Co-expressed Yeast Genes Cluster Over a Long Range but are not Regularly Spaced. *J Mol Biol*.

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. (2006) Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A,* 103**,** 9935-9.

Lewis, P. O. (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol,* 50**,** 913-25.

Li, W.-H. (1997) *Molecular Evolution,* Sunderland, Mass., Sinauer.

Li, W. H., Wu, C. I. & Luo, C. C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol,* 2**,** 150-74.

Liti, G., Barton, D. B. & Louis, E. J. (2006) Sequence Diversity, Reproductive Isolation and Species Concepts in Saccharomyces. *Genetics*.

Liti, G., Peruffo, A., James, S. A., Roberts, I. N. & Louis, E. J. (2005) Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the Saccharomyces sensu stricto complex. *Yeast,* 22**,** 177-92.

Llorente, B., Durrens, P., Malpertuy, A., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., Dujon, B., De Montigny, J., Lepingle, A., Neuveglise, C., Ozier-Kalogeropoulos, O., Potier, S., Saurin, W., Tekaia, F., Toffano-Nioche, C., Wesolowski-Louvel, M., Wincker, P., Weissenbach, J., Souciet, J. & Gaillardin, C. (2000) Genomic Exploration of the Hemiascomycetous Yeasts: 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*. *FEBS Lett,* 487**,** 122-133.

Logue, M. E., Wong, S., Wolfe, K. H. & Butler, G. (2005) A genome sequence survey shows that the pathogenic yeast Candida parapsilosis has a defective MTLa1 at its mating type locus. *Eukaryot Cell,* 4**,** 1009-1017.

Long, M., Betrán, E., Thornton, K. & Wang, W. (2003) The origin of new genes: glimpses from the young and old. *Nat Rev Genet,* 4**,** 865-875.

Long, M. & Langley, C. H. (1993) Natural selection and the origin of *jingwei,* a chimeric processed functional gene in *Drosophila*. *Science,* 260**,** 91-5.

Long, M., Wang, W. & Zhang, J. (1999) Origin of new genes and source for N-terminal domain of the chimerical gene, jingwei, in Drosophila. *Gene,* 238**,** 135-41.

Lowe, T. M. & Eddy, S. R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res,* 25**,** 955-64.

Lupski, J. R. & Stankiewicz, P. (2005) Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet,* 1**,** e49.

Lynch, M. (2004) Gene duplication and evolution. IN MOYA, A. & E., F. (Eds.) *Evolution: From molecules to ecosystems.* Oxford, Oxford University Press.

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

Lynch, M. & Conery, J. S. (2003) The origins of genome complexity. *Science,* 302**,** 1401-1404.

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

Lynch, M. & Force, A. G. (2000b) The origin of interspecies genomic incompatibility via gene duplication. *Am Nat,* 156**,** 590-605.

Lynch, M. & Katju, V. (2004) The altered evolutionary trajectories of gene duplicates. *Trends Genet,* 20**,** 544-9.

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

Lynch, M., Scofield, D. G. & Hong, X. (2005) The evolution of transcription-initiation sites. *Mol Biol Evol,* 22**,** 1137-46.

Macpherson, S., Larochelle, M. & Turcotte, B. (2006) A fungal family of transcriptional regulators: the zinc cluster proteins. *Microbiol Mol Biol Rev,* 70**,** 583-604.

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

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.

Martens, J. A., Laprade, L. & Winston, F. (2004) Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature,* 429**,** 571-4.

Masly, J. P., Jones, C. D., Noor, M. A., Locke, J. & Orr, H. A. (2006) Gene transposition as a cause of hybrid sterility in Drosophila. *Science,* 313**,** 1448-50.

Mayr, E. (1942) *Systematics and the Origin of Species*.

Mclysaght, A., Hokamp, K. & Wolfe, K. H. (2002) Extensive genomic duplication during early chordate evolution. *Nat Genet,* 31**,** 200-4.

Melnick, L. & Sherman, F. (1993) The gene clusters *ARC* and *COR* on chromosomes 5 and 10, respectively, of *Saccharomyces cerevisiae* share a common ancestry. *J Mol Biol,* 233**,** 372-88.

Miller, M. G. & Johnson, A. D. (2002) White-opaque switching in *Candida albicans* is controlled by mating-type locus homeodomain proteins and allows efficient mating. *Cell,* 110**,** 293-302.

Moller, K., Olsson, L. & Piskur, J. (2001a) Ability for anaerobic growth is not sufficient for development of the petite phenotype in *Saccharomyces kluyveri*. *J Bacteriol,* 183**,** 2485-9.

Moller, K., Sharif, M. Z. & Olsson, L. (2004) Production of fungal alpha-amylase by Saccharomyces kluyveri in glucose-limited cultivations. *J Biotechnol,* 111**,** 311-8.

Moller, K., Tidemand, L. D., Winther, J. R., Olsson, L., Piskur, J. & Nielsen, J. (2001b) Production of a heterologous proteinase A by Saccharomyces kluyveri. *Appl Microbiol Biotechnol,* 57**,** 216-9.

Moore, R. C. & Purugganan, M. D. (2003) The early stages of duplicate gene evolution. *Proc Natl Acad Sci U S A*.

Morett, E., Korbel, J. O., Rajan, E., Saab-Rincon, G., Olvera, L., Olvera, M., Schmidt, S., Snel, B. & Bork, P. (2003) Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat Biotechnol,* 21**,** 790-5.

Mortimer, R. K. (1993a) Carl C. Lindegren: Iconoclastic father of *Neurospora* and yeast genetics. IN HALL, M. N. & LINDER, P. (Eds.) *The Early Days of Yeast Genetics.* New York, Cold Spring Harbor Laboratory Press.

Mortimer, R. K. (1993b) Øjvind Winge: Founder of yeast genetics. IN HALL, M. N. & LINDER, P. (Eds.) *The Early Days of Yeast Genetics.* New York, Cold Spring Harbor Laboratory Press.

Mortimer, R. K. (2000) Evolution and variation of the yeast (*Saccharomyces*) genome. *Genome Res,* 10**,** 403-9.

Muthukumar, G., Suhng, S. H., Magee, P. T., Jewell, R. D. & Primerano, D. A. (1993) The Saccharomyces cerevisiae SPR1 gene encodes a sporulation-specific exo-1,3-beta-glucanase which contributes to ascospore thermoresistance. *J Bacteriol,* 175**,** 386-94.

Nakayashiki, H., Kadotani, N. & Mayama, S. (2006) Evolution and diversification of RNA silencing proteins in fungi. *J Mol Evol,* 63**,** 127-35.

Neafsey, D. E. & Hartl, D. L. (2005) Convergent loss of an anciently duplicated, functionally divergent RH2 opsin gene in the fugu and Tetraodon pufferfish lineages. *Gene,* 350**,** 161-71.

Nelson, C. E., Hersh, B. M. & Carroll, S. B. (2004) The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol,* 5**,** R25.

Nembaware, V., Crum, K., Kelso, J. & Seoighe, C. (2002) Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Res,* 12**,** 1370-6.

Neuveglise, C., Feldmann, H., Bon, E., Gaillardin, C. & Casaregola, S. (2002) Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts. *Genome Res,* 12**,** 930-43.

Nissley, D. V., Boyer, P. L., Garfinkel, D. J., Hughes, S. H. & Strathern, J. N. (1998) Hybrid Ty1/HIV-1 elements used to detect inhibitors and monitor the activity of HIV-1 reverse transcriptase. *Proc Natl Acad Sci U S A,* 95**,** 13905-10.

Noor, M. A. & Feder, J. L. (2006) Speciation genetics: evolving approaches. *Nat Rev Genet,* 7**,** 851-61.

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

Nurminsky, D. I., Nurminskaya, M. V., De Aguiar, D. & Hartl, D. L. (1998) Selective sweep of a newly evolved sperm-specific gene in Drosophila. *Nature,* 396**,** 572-5.

Nussbaum, R. L. (2005) Mining yeast in silico unearths a golden nugget for mitochondrial biology. *J Clin Invest,* 115**,** 2689-91.

Ohno, S. (1970) *Evolution by Gene Duplication,* London, George Allen and Unwin.

Oliver, S. G., Van Der Aart, Q. J., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J. P., Benit, P. & Et Al. (1992) The complete DNA sequence of yeast chromosome III. *Nature,* 357**,** 38-46.

Olson, M. V. (1999) When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet,* 64**,** 18-23.

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

Padiath, Q. S., Saigoh, K., Schiffmann, R., Asahara, H., Yamada, T., Koeppen, A., Hogan, K., Ptacek, L. J. & Fu, Y. H. (2006) Lamin B1 duplications cause autosomal dominant leukodystrophy. *Nat Genet,* 38**,** 1114-23.

Pal, C. & Hurst, L. D. (2003) Evidence for co-evolution of gene order and recombination rate. *Nat Genet,* 33**,** 392-5.

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

Paterson, A. H., Bowers, J. E. & Chapman, B. A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A,* 101**,** 9903-8.

Pfeiffer, T., Schuster, S. & Bonhoeffer, S. (2001) Cooperation and competition in the evolution of ATP-producing pathways. *Science,* 292**,** 504-7.

Phillips, M. J., Delsuc, F. & Penny, D. (2004) Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol,* 21**,** 1455-8.

Piatigorsky, J. & Wistow, G. (1991) The recruitment of crystallins: new functions precede gene duplication. *Science,* 252**,** 1078-9.

Piskur, J. & Langkjaer, R. B. (2004) Yeast genome sequencing: the power of comparative genomics. *Mol Microbiol,* 53**,** 381-9.

Pop, M., Kosack, D. S. & Salzberg, S. L. (2004) Hierarchical scaffolding with Bambus. *Genome Res,* 14**,** 149-59.

Postlethwait, J., Amores, A., Cresko, W., Singer, A. & Yan, Y. L. (2004) Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet,* 20**,** 481-90.

Presgraves, D. C. (2003) A fine-scale genetic analysis of hybrid incompatibilities in Drosophila. *Genetics,* 163**,** 955-72.

Presgraves, D. C., Balagopalan, L., Abmayr, S. M. & Orr, H. A. (2003) Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature,* 423**,** 715-9.

Press, W. H., Teukolsky, S. A., Vetterling, W. A. & Flannery, B. P. (1992) *Numerical Recipes in C,* New York, Cambridge University Press.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., Mccartney, R. R., Schmidt, M. C., Rachidi, N., Lee, S. J., Mah, A. S., Meng, L., Stark, M. J., Stern, D. F., De Virgilio, C., Tyers, M., Andrews, B., Gerstein, M., Schweitzer, B., Predki, P. F. & Snyder, M. (2005) Global analysis of protein phosphorylation in yeast. *Nature,* 438**,** 679-84.

Pupko, T., Pe'er, I., Hasegawa, M., Graur, D. & Friedman, N. (2002) A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. *Bioinformatics,* 18**,** 1116-23.

Ranz, J. M., Ponce, A. R., Hartl, D. L. & Nurminsky, D. (2003) Origin and evolution of a new gene expressed in the Drosophila sperm axoneme. *Genetica,* 118**,** 233-44.

Rizzon, C., Ponger, L. & Gaut, B. S. (2006) Striking Similarities in the Genomic Distribution of Tandemly Arrayed Genes in Arabidopsis and Rice. *PLoS Comput Biol,* 2.

Ro, D. K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., Ho, K. A., Eachus, R. A., Ham, T. S., Kirby, J., Chang, M. C., Withers, S. T., Shiba, Y., Sarpong, R. & Keasling, J. D. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature,* 440**,** 940-3.

Roberts, C. J. & Van Der Walt, J. P. (1959) The life cycle of *Kluyveromyces polysporus*. *Compt Rend Lab Carlsberg,* 31**,** 129-148.

Rokas, A., Williams, B. L., King, N. & Carroll, S. B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature,* 425**,** 798-804.

Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerriere, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M., Dubas, F., Frebourg, T. & Campion, D. (2006) APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet,* 38**,** 24-6.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S., Gong, F., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J., Li, Z., Hynes, R. O., Jones, S. J., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. H. & Lewis, S. (2000) Comparative genomics of the eukaryotes. *Science,* 287**,** 2204-15.

Ruderfer, D. M., Pratt, S. C., Seidel, H. S. & Kruglyak, L. (2006) Population genomic analysis of outcrossing and recombination in yeast. *Nat Genet.*

Salzberg, S. L., White, O., Peterson, J. & Eisen, J. A. (2001) Microbial genes in the human genome: lateral transfer or gene loss? *Science,* 292**,** 1903-6.

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

Scannell, D. R., Frank, A. C., Conant, G. C., Byrne, K. P., Woolfit, M. & K.H., W. (2006b) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci U S A,* In Press.

Scannell, D. R. & Wolfe, K. (2004) Rewiring the transcriptional regulatory circuits of cells. *Genome Biol,* 5**,** 206.

Schacherer, J., De Montigny, J., Welcker, A., Souciet, J. L. & Potier, S. (2005) Duplication processes in Saccharomyces cerevisiae haploid strains. *Nucleic Acids Res,* 33**,** 6319-26.

Schacherer, J., Tourrette, Y., Souciet, J.-L., Potier, S. & De Montigny, J. (2004) Recovery of a Function Involving Gene Duplication by Retroposition in Saccharomyces cerevisiae. *Genome Res,* 14**,** 1291-1297.

Schmidt, H. A., Strimmer, K., Vingron, M. & Von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics,* 18**,** 502-4.

Schranz, M. E. & Mitchell-Olds, T. (2006) Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell,* 18**,** 1152-65.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J. P. & Widom, J. (2006) A genomic code for nucleosome positioning. *Nature,* 442**,** 772-8.

Semon, M. & Duret, L. (2005) No evidence of selection for the clustering of co-expressed genes in the human genome. *Mol Biol Evol*?

Seoighe, C. & Gehring, C. (2004) Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet,* 20**,** 461-4.

Seoighe, C., Johnston, C. R. & Shields, D. C. (2003) Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. *Mol Biol Evol,* 20**,** 484-90.

Seoighe, C. & Wolfe, K. H. (1999a) Yeast genome evolution in the post-genome era. *Curr Opin Microbiol,* 2**,** 548-554.

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

Sheeman, B., Carvalho, P., Sagot, I., Geiser, J., Kho, D., Hoyt, M. A. & Pellman, D. (2003) Determinants of S. cerevisiae dynein localization and activation: implications for the mechanism of spindle positioning. *Curr Biol,* 13**,** 364-72.

Sherlock, G., Rosenzweig, R. F., Levine, R. P., Dunn, B. L. & Schwartz, K. (2006) Directed Evolution and Genomic Analysis of Novel Yeast Species for More Efficient Biomass Conversion. Stanford Global Energy and Climate Project.

Shimodaira, H. & Hasegawa, M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics,* 17**,** 1246-7.

Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M. & Van De Peer, Y. (2002) The hidden duplication past of Arabidopsis thaliana. *Proc Natl Acad Sci U S A,* 99**,** 13627-32.

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.

Stoltzfus, A. (1999) On the possibility of constructive neutral evolution. *J Mol Evol,* 49**,** 169-81.

Storchova, Z., Breneman, A., Cande, J., Dunn, J., Burbank, K., O'toole, E. & Pellman, D. (2006) Genome-wide genetic analysis of polyploidy in yeast. *Nature,* 443**,** 541-7.

Sudarsanam, P., Pilpel, Y. & Church, G. M. (2002) Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. *Genome Res,* 12**,** 1723-31.

Sugino, R. P. & Innan, H. (2005) Estimating the Time to the Whole-Genome Duplication and the Duration of Concerted Evolution via Gene Conversion in Yeast. *Genetics,* 171**,** 63-9.

Taddei, A., Van Houwe, G., Hediger, F., Kalck, V., Cubizolles, F., Schober, H. & Gasser, S. M. (2006) Nuclear pore association confers optimal expression levels for an inducible yeast gene. *Nature,* 441**,** 774-8.

Tanay, A. (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res,* 16**,** 962-72.

Tang, H., Wyckoff, G. J., Lu, J. & Wu, C. I. (2004) A universal evolutionary index for amino acid changes. *Mol Biol Evol,* 21**,** 1548-56.

Taxis, C., Keller, P., Kavagiou, Z., Jensen, L. J., Colombelli, J., Bork, P., Stelzer, E. H. & Knop, M. (2005) Spore number control and breeding in Saccharomyces cerevisiae: a key role for a self-organizing system. *J Cell Biol,* 171**,** 627-40.

Taylor, J. S. & Raes, J. (2004) Duplication and Divergence: The Evolution of New Genes and Old Ideas. *Annu Rev Genet,* 38**,** 615-643.

Taylor, J. S., Van De Peer, Y. & Meyer, A. (2001) Genome duplication, divergent resolution and speciation. *Trends Genet,* 17**,** 299-301.

Teichmann, S. A. & Veitia, R. A. (2004) Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics,* 167**,** 2121-5.

Thomas, B. C., Pedersen, B. & Freeling, M. (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res,* 16**,** 934-46.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* 22**,** 4673-4680.

Thomson, J. M., Gaucher, E. A., Burgan, M. F., De Kee, D. W., Li, T., Aris, J. P. & Benner, S. A. (2005) Resurrecting ancestral alcohol dehydrogenases from yeast. *Nature Genetics,* 37**,** 630-635.

Town, C. D., Cheung, F., Maiti, R., Crabtree, J., Haas, B. J., Wortman, J. R., Hine, E. E., Althoff, R., Arbogast, T. S., Tallon, L. J., Vigouroux, M., Trick, M. & Bancroft, I. (2006) Comparative genomics of Brassica oleracea and Arabidopsis thaliana reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell,* 18**,** 1348-59.

Tsong, A. E., Miller, M. G., Raisner, R. M. & Johnson, A. D. (2003) Evolution of a combinatorial transcriptional circuit: a case study in yeasts. *Cell,* 115**,** 389-99.

Tsong, A. E., Tuch, B. B., Li, H. & Johnson, A. D. (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature,* 443**,** 415-20.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G. L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J. C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C. J., Uberbacher, E., Unneberg, P., et al. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science,* 313**,** 1596-604.

Tzung, K. W., Williams, R. M., Scherer, S., Federspiel, N., Jones, T., Hansen, N., Bivolarevic, V., Huizar, L., Komp, C., Surzycki, R., Tamse, R., Davis, R. W. & Agabian, N. (2001) Genomic evidence for a complete sexual cycle in *Candida albicans*. *Proc Natl Acad Sci U S A,* 98**,** 3249-53.

Van De Peer, Y., Taylor, J. S., Braasch, I. & Meyer, A. (2001) The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol,* 53**,** 436-46.

Van Der Walt, J. P. (1956) *Kluyveromyces* – A new yeast genus of the *Endomycetales*. *Antonie van Leeuwenhoek,* 22**,** 265-272.

Van Hoof, A. (2005) Conserved functions of yeast genes support the Duplication, Degeneration and Complementation model for gene duplication. *Genetics,* 171**,** 1455-1461.

Veitia, R. A. (2005) Paralogs in polyploids: one for all and all for one? *Plant Cell,* 17**,** 4-11.

Wang, T. & Stormo, G. D. (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proc Natl Acad Sci U S A,* 102**,** 17400-5.

Wang, W., Yu, H. & Long, M. (2004) Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. *Nat Genet,* 36**,** 523-7.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature,* 420**,** 520-62.

Werth, C. R. & Windham, M. D. (1991) A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *Am Nat,* 137**,** 515-526.

Willingham, S., Outeiro, T. F., Devit, M. J., Lindquist, S. L. & Muchowski, P. J. (2003) Yeast genes that enhance the toxicity of a mutant huntingtin fragment or alpha-synuclein. *Science,* 302**,** 1769-72.

Wittbrodt, J., Adam, D., Malitschek, B., Maueler, W., Raulf, F., Telling, A., Robertson, S. M. & Schartl, M. (1989) Novel putative receptor tyrosine kinase encoded by the melanoma-inducing Tu locus in Xiphophorus. *Nature,* 341**,** 415-21.

Wolfe, K. (2004) Evolutionary genomics: Yeasts accelerate beyond BLAST. *Curr Biol,* 14**,** R392-R394.

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

Wolfe, K. H. (2006) Comparative genomics and genome evolution in yeasts. *Philos Trans R Soc Lond B Biol Sci,* 361**,** 403-12.

Wolfe, K. H., Morden, C. W. & Palmer, J. D. (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc Natl Acad Sci U S A,* 89**,** 10648-52.

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

Wolfe, K. H. & Shields, D. C. (1997b) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature,* 387**,** 708-713.

Wong, S., Butler, G. & Wolfe, K. H. (2002) Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc Natl Acad Sci U S A,* 99**,** 9272-7.

Wong, S., Fares, M. A., Zimmermann, W., Butler, G. & Wolfe, K. H. (2003) Evidence from comparative genomics for a complete sexual cycle in the "asexual" pathogenic yeast *Candida glabrata*. *Genome Biol,* 4**,** R10.

Wong, S. & Wolfe, K. H. (2005) Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nature Genetics,* 37**,** 777-82.

Wong, S. & Wolfe, K. H. (2006) Duplication of genes and genomes in yeasts. IN SUNNERHAGEN, P. & PISKUR, J. (Eds.) *Comparative genomics.* Heidelberg, Springer-Verlag.

Wood, V., Gwilliam, R., Rajandream, M. A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E. J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., Mcdonald, S., Mclean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'neil, S., Pearson, D., Quail, M. A., Rabbinowitsch, E., Rutherford, K., Rutter, S., Saunders, D., Seeger, K., Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R. G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schafer, M., Muller-Auer, S., Gabel, C., Fuchs, M., Dusterhoft, A., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T. M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dreano, S., Gloux, S., et al. (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature,* 415**,** 871-80.

Woolfit, M., Rozpedowska, E., Piskur, J. & Wolfe, K. H. (2006) Genome survey sequencing of the wine spoilage yeast Dekkera (Brettanomyces) bruxellens. *Eukaryotic Cell,* In Press.

Wu, C. I. & Ting, C. T. (2004) Genes and speciation. *Nat Rev Genet,* 5**,** 114-22.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., Li, R., Xu, Z., Li, S., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., Geng, J., Li, G., Shi, J., Liu, J., Lv, H., Li, J., Wang, J., Deng, Y., Ran, L., Shi, X., Wang, X., Wu, Q., Li, C., Ren, X., Wang, J., Wang, X., Li, D., Liu, D., Zhang, X., Ji, Z., Zhao, W., Sun, Y., Zhang, Z., Bao, J., Han, Y., Dong, L., Ji, J., Chen, P., Wu, S., Liu, J., Xiao, Y., Bu, D., Tan, J., Yang, L., Ye, C., Zhang, J., Xu, J., Zhou, Y., Yu, Y., Zhang, B., Zhuang, S., Wei, H., Liu, B., Lei, M., Yu, H., Li, Y., Xu, H., Wei, S., He, X., Fang, L., Zhang, Z., Zhang, Y., Huang, X., Su, Z., Tong, W., Li, J., Tong, Z., Li, S., Ye, J., Wang, L., Fang, L., Lei, T., Chen, C., Chen, H., Xu, Z., Li, H., Huang, H., Zhang, F., Xu, H., Li, N., Zhao, C., Li, S., Dong, L., Huang, Y., Li, L., Xi, Y., Qi, Q., Li, W., Zhang, B., Hu, W., et al. (2005) The Genomes of Oryza sativa: a history of duplications. *PLoS Biol,* 3**,** e38.

Zhang, J., Dean, A. M., Brunet, F. & Long, M. (2004) Evolving protein functional diversity in new genes of Drosophila. *Proc Natl Acad Sci U S A,* 101**,** 16246-50.

Zhang, P., Gu, Z. & Li, W. H. (2003) Different evolutionary patterns between young duplicate genes in the human genome. *Genome Biol,* 4**,** R56.