



Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

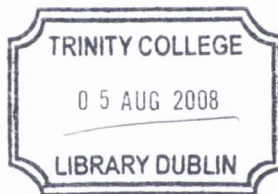
Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.



THOSIS

8580

Genomic Evolution of Polyploid Yeast Species

by Jonathan Gordon

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

Department of Genetics
University of Dublin
Trinity College

October 2007



Declaration

This thesis is submitted by the undersigned for the degree of Doctor of Philosophy from Trinity College Dublin. Unless otherwise stated, the work in this thesis was undertaken by the author alone. I agree that this thesis may be lent or copied by the Trinity College Dublin Library.

Signed.....

Jonathan Gordon

31st October, 2007

Summary

In Chapter 2 of this thesis I study how the genome structure and content of yeast has evolved since a Whole Genome Duplication (WGD) occurred in the lineage approximately 100 million years ago. By using genome data from several extant species whose lineages arose both before and after the WGD, I parsimoniously reconstructed the genome order and content for the ancestor that existed immediately prior to the WGD. This ancestor contained at least 4703 loci, and there are 66 reciprocal translocations, 73 inversions, 5 Robertsonian translocations, 9 single-gene transpositions and 2 multigene transpositions found between the genome duplicated ancestor and the extant *S. cerevisiae* genome. Some of these rearrangements are phylogenetically informative because they are shared by one or more species to the exclusion of others. The phylogeny inferred by these shared rearrangements differs from most sequence-based phylogenies. There have been 153 genes gained in *S. cerevisiae* since the ancestor, and there are 144 loci in the ancestor that have been deleted in double copy in *S. cerevisiae*. Breakpoints created by rearrangements and the locations of gene gains are enriched for tRNA genes in *S. cerevisiae*. There is evidence of breakpoint reuse in rearrangements, and there is an average use of 1.22 rearrangements per breakpoint.

In Chapter 3, I examine the evolutionary dynamics of rearrangements in the region of the *MAT* locus in the post-WGD yeast clade by comparing the gene orders around the *MAT* locus in extant species. Part of the *MAT* chromosome in *Z. rouxii* CBS732 was sequenced and used in conjunction with the inferred ancestral *MAT* chromosome to infer changes in the structure of the *MAT* chromosome in the lineage of *S. cerevisiae*. Relative to the pre-WGD ancestor, multiple blocks of genes have been either deleted or moved from beside the *MAT* locus to other genomic locations in many of the hemiascomycete species examined. Some of these relocations occur on shared branches of the phylogeny and are common between certain species to the exclusion of outgroup species. The genomic region beside the *MAT* locus probably represents a hot-spot for genomic rearrangement and deletion by the gradual erosion of the chromosome. The relocation process is especially predominant in post-WGD species that contain the Ho endonuclease, which creates a double strand

break (DSB) at the *MAT* locus during mating type switching. In addition, the mating-cassette X and Z regions in hemiascomycetes often incorporates parts of genes that reside beside the *MAT* locus, which may provide evidence for the mechanism of the deletion and relocation process.

Finally, in Chapter 4 I examine the likely origin of a strain of *Z. rouxii* (ATCC 42981) that is known to contain duplicate copies of the *SOD2* and *HOG2* genes relative to the *Z. rouxii* type-strain (strain CBS 732). Sequencing of random plasmid end reads reveals that *Z. rouxii* ATCC 42981 has two 26S rRNA genes, which closely resemble those from the *Z. rouxii* type-strain and that of another closely related species, *Z. pseudorouxii*. In addition, comparing contigs from ATCC42981 with contigs from *Z. rouxii* CBS732 reveals that the ATCC 42981 contigs map in a two to one fashion to the CBS 732 contigs. One of these regions is practically identical to the type-strain region while the other is 80-90% similar at the nucleotide sequence level. Other strains designated as *Z. rouxii* or *Z. pseudorouxii* show a similar pattern with available sequence data, and along with ATCC 42981 they probably represent a new hybrid species clade that was formed by the hybridisation of *Z. rouxii* and *Z. pseudorouxii*.

Acknowledgements

Firstly, I would like to thank my parents Hugh and Margot, and my sister Candice for their love and support over the years, without which this work would not have been possible.

I would like to thank all members of the Wolfe lab, both past and present who I have had the pleasure of working with: Mario, Greg, Simon, Lukasz, Ailis, Andrew, John, Matt, Brian, Devin, Meg, Marie, Kevin, Gavin, Nora, Jeff, Carolin, Seán and Nadia. Your advice has been invaluable and I thoroughly enjoyed many of the fine discussions over pints in Kennedy's.

Finally, I would like to thank Ken Wolfe for being a great supervisor. Your knowledge of the field has never ceased to amaze me and I am in your debt for the many directions you have helped me to travel in my research.

Table of Contents

CHAPTER 1 INTRODUCTION1

1.1 Comparative Genomics1

 1.1.1 Inferring Gene Functions3

 1.1.2 Genome Content and Structure Evolution4

1.2 Hemiascomycete Yeasts6

 1.2.1 Species Studied6

 1.2.2 Genome Structure7

1.3 The *MAT* Locus.8

 1.3.1 Silent Cassette Structure11

 1.3.2 Mechanism of Mating-Type Switching13

 1.3.3 Evolution of the *MAT* Locus System15

1.4 Whole Genome Duplication17

 1.4.1 Ohno Hypothesis17

 1.4.2 WGD in hemiascomycetes18

 1.4.3 Fate of WGD duplicate genes19

1.5 Gene content evolution of genomes24

 1.5.1 Gene Losses24

 1.5.2 Gene gains26

 1.5.2.1 Duplication26

 1.5.2.2 HGT28

 1.5.2.3 *de novo* Creation of Genes29

1.6 Structural evolution of genomes30

 1.6.1 Models of chromosome rearrangement34

1.7 Synteny conservation and gene clustering34

1.8 Ancestral gene order reconstruction37

CHAPTER 2 GENES ADDED, GENES LOST, GENOMES REARRANGED: THE EVOLUTIONARY ROUTE FROM A RECONSTRUCTED ANCESTOR TO THE MODERN *Saccharomyces cerevisiae* GENOME41

2.1 Abstract	41
2.2 Introduction	42
2.3 Results and Discussion	44
2.3.1 Ancestral Genome Reconstruction	44
2.3.2 Rearrangement route to the current <i>S. cerevisiae</i> genome	49
2.3.3 Breakpoint re-use	53
2.3.4 Coincidence of sites of gene gain, evolutionary breakpoints, tRNA genes and origins of replication	54
2.3.5 Categories of gained genes	57
2.3.6 Functions of genes gained in the <i>S. cerevisiae</i> lineage since WGD	64
2.3.6.1 Nitrogen metabolism and transport	64
2.3.6.2 Sterol transporters	67
2.3.6.3 Thiamin uptake and biosynthesis genes	67
2.3.6.4 Ethanol production and consumption	68
2.3.6.5 Amplified membrane protein gene families	69
2.3.6.6 GATA family transcription factors	70
2.4 Conclusion	71
2.5 Methods	72
2.5.1 Parsimonious inference of ancestral gene content and order	72
2.5.2 Evolution of ancestral organisation into <i>S. cerevisiae</i> gene order	73
2.5.3 Inference and analysis of probable gene gains in the <i>S. cerevisiae</i> lineage since the WGD	74
2.5.4 Computer simulations	74

CHAPTER 3 EVOLUTIONARY EROSION AND TRANSPOSITION OF GENES FLANKING THE *MAT* LOCUS IN YEAST SPECIES76

3.1 Abstract	76
3.2 Introduction	77
3.3 Results	82
3.3.1 <i>Z. rouxii</i> MAT chromosome	82
3.3.2 Gene content in the ancestral MAT chromosome	83
3.3.3 Erosion and Transposition (EAT) mechanism	86
3.3.4 Asymmetric pattern of gene deletions between MAT and HML	91
3.3.5 Fate of the Duplicated MAT locus	92
3.3.6 Truncation of genes and loss of intergenic DNA in the X and Z regions	93
3.4 Discussion	95
3.4.1 <i>Z. rouxii</i> MAT chromosome and MAT locus	95
3.4.2 Fate of the duplicated MAT locus	97
3.4.3 Evolution of the MAT locus and bordering genes	98
3.4.4 Feedback in the triplicated MAT-HML-HMR system	100
3.5 Methods	101

CHAPTER 4 THE *Zygosaccharomyces rouxii* STRAIN ATCC 42981 IS AN ALLOPOLYPLOID HYBRID BETWEEN TWO CLOSELY RELATED *Zygosaccharomyces* SPECIES103

4.2 Introduction	103
4.3 Results	106
4.3.1 <i>Z. rouxii</i> ATCC 42981 is a recent hybrid between two <i>Zygosaccharomyces</i> species	106
4.3.2 Comparison between <i>Z. rouxii</i> ATCC 42981 and other <i>Zygosaccharomyces</i> hybrids	108
4.4 Discussion	108
4.5 Methods	116

REFERENCES118

TABLE OF FIGURES

Figure 1.15

Figure 1.212

Figure 1.314

Figure 1.416

Figure 1.519

Figure 1.632

Figure 2.145

Figure 2.248

Figure 2.350

Figure 2.451

Figure 2.553

Figure 2.655

Figure 2.759

Figure 2.866

Figure 3.178-80

Figure 3.283

Figure 3.384-85

Figure 3.487

Figure 3.588

Figure 4.1109

Figure 4.2110

Figure 4.3113

Figure 4.4114

TABLE OF TABLES

Table 1.16

Table 2.154

Table 2.2		58
Table 2.3		60-64
Table 2.4		65
Table 3.1		94
Table 3.2		95
Table 4.1		107
Table 4.2		111

ABBREVIATIONS

WGD – Whole Genome Duplication

asgs – **a**-specific genes

αsgs – α-specific genes

hsgs – haploid specific genes

ORC – Origin Recognition Complex

DCS – Double-Conserved Syteny

RGL – Reciprocal Gene Loss

HGT – Horizontal Gene Transfer

RT – Reciprocal Translocation

YGOB – Yeast Gene Order Browser

ABC – ATP-Binding Cassette

GAPDH – glyceraldehydes 3-phosphate dehydrogenase

DSB – Double Strand Break

EAT – Erosion And Transposition

APPENDIX I LIST OF FILES ON SUPPLEMENTARY CD

Figure S2.1		FigureS2-1.pdf
Table S2.1		TableS2-1.xls
Table S2.2		TableS2-2.xls
Table S3.1		TableS3-1.xls
Table S3.2		TableS3-2.xls
Z. rouxii CBS 732 plasmid end reads		Zr_CBS732_reads.fasta

Z. rouxii CBS 732 sequenced contigs	Zr_CBS732_contigs.fasta
Z. rouxii ATCC 42981 plasmid end reads	Zr_ATCC42981_reads.fasta
Z. rouxii ATCC 42981 contigs	Zr_ATCC42981_contigs.fasta

Genomic Evolution of Polyploid Yeast Species

Chapter1

Introduction

In this chapter I give an overview of the field of comparative genomics, specifically in the context of the hemiascomycete yeast species. I also briefly provide information about these species (mostly concentrating on the model organism *S. cerevisiae*), and I discuss the mechanism and evolution of mating-type switching in yeast. In addition, I consider the different evolutionary changes in terms of both genome content (gain, duplication and deletion) and arrangement that can occur in genomes, (including the whole genome duplication (WGD) that occurred in the hemiascomycetes), and their evolutionary consequences. Finally I review ancestral gene order reconstruction by comparisons between extant species.

1.1 Comparative Genomics

The relatively new field of comparative genomics involves the study of the genetic relationships between and within organisms by comparing shared and unique features of their genomes. The underlying concepts of comparative genomics are that functional regions of genomes tend to be maintained by purifying selection and evolve much slower than non-functional regions, and that genetic similarities including gene order shared by different species are generally inherited from a common ancestor. The power of comparative genomics generally increases with the number of species or sequences that are compared (Thomas and Touchman 2002; Kellis et al. 2004). In some cases, genomic regions may evolve too quickly or too slowly to give useful information in certain comparisons between different species, which may limit comparative power depending on the phylogenetic depth and distribution of the species involved. Comparative genomics between organisms has proven useful in identifying genes in unannotated genomes and new

genes in already annotated genomes, as well as regulatory regions upstream of genes, which also evolve at a slower rate than the non-functional DNA surrounding them (Cliften et al. 2003; Kellis et al. 2003; Dietrich et al. 2004). Sequences in different species that are related to each other by a speciation event are called orthologues, while sequences which are related to each other by duplication of an ancestral sequence are called paralogues. Sequence conservation and identifiable changes between species can be used for inferences about the function and evolution of genes since the common ancestor of those sequences. It is also possible to make intragenomic comparisons between genes based on measures such as sequence similarity and the expression profiles of the genes (Li, Yang, and Gu 2005).

The first comparative genomics studies involved bacterial comparisons as these were the first genomes to be fully sequenced (Fleischmann et al. 1995; Koonin, Tatusov, and Rudd 1995; Mushegian and Koonin 1996; Tatusov et al. 1996; Blattner et al. 1997; de Rosa and Labedan 1998). Bacteria were also ideal as model organisms due to their small size, compact genomes, fast replication times and ease of culturing and genetic manipulation in laboratory conditions, as well as for medical and industrial reasons. The result was a wealth of functional and metabolic pathway data available at the time of genome sequencing (Koonin, Tatusov, and Rudd 1995). These data were important for further functional prediction both within and between organisms by observing similarities between gene sequences. The first fully sequenced eukaryotic genome was that of baker's yeast *Saccharomyces cerevisiae* in 1996, which was chosen for sequencing for many of the same reasons as bacteria; it is a single celled organism with a relatively compact genome containing few introns and repeat sequences, it is easy to grow in lab conditions, it is easy to manipulate and it has also been used as a model organism of economic importance prior to genomic sequencing (Goffeau et al. 1996; Dujon et al. 2004). At present the hemiascomycete yeasts represent one of the best sets of organisms for comparative genomics as there are a number of fully sequenced species that span a range of phylogenetic distances. In addition there is a wealth of functional data regarding the model organism *S. cerevisiae* including thousands of microarray experiments and knockout phenotypes for every identified gene (Winzeler et al. 1999; Wolfe 2006). This data in *S.*

cerevisiae allows inferences about the physiology of the other hemiascomycetes through genomic comparisons, as well as about the evolutionary processes affecting all of the species. In turn data from the other hemiascomycete yeasts can be used to better understand the *S. cerevisiae* genome, for example in the identification of new previously unannotated genes and the detection of gene regulatory functional elements (Cliften et al. 2001; Cliften et al. 2003; Kellis et al. 2003).

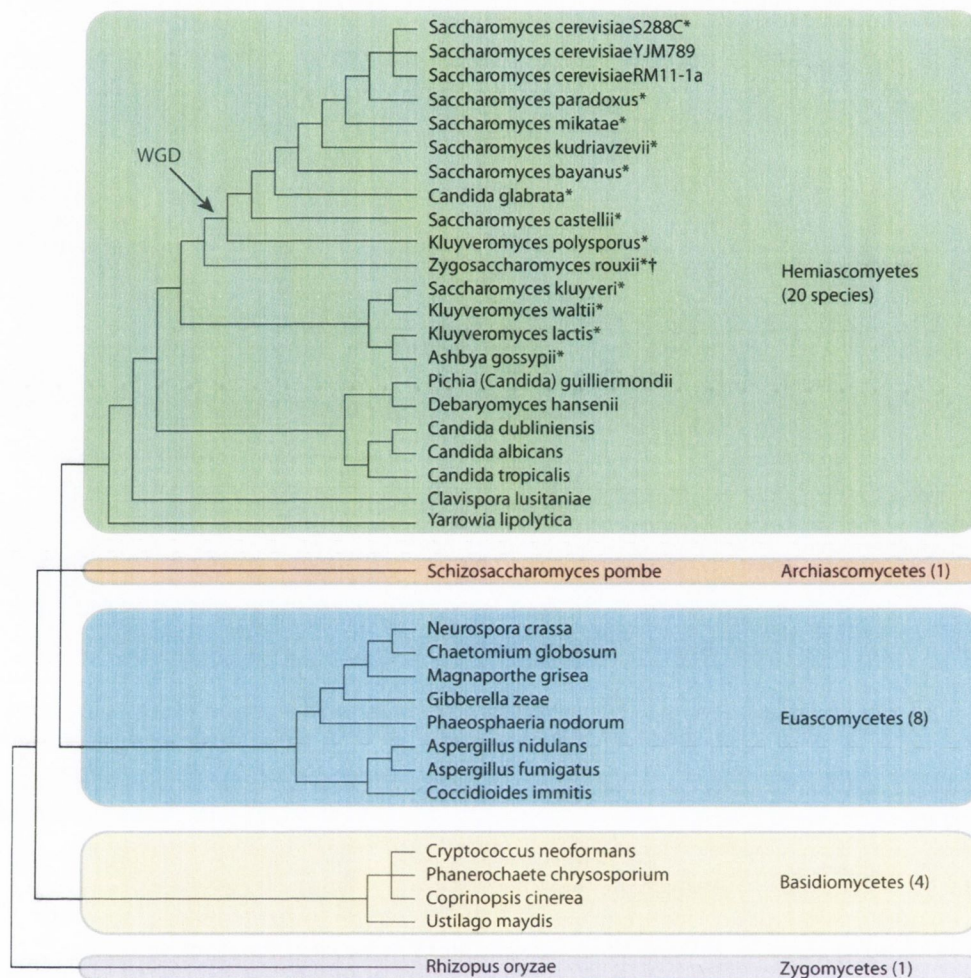
As this thesis deals with yeast comparative genomics and genome evolution I will discuss comparative genomics using the yeasts *Saccharomyces cerevisiae*, *Candida glabrata*, *Saccharomyces castellii*, *Kluyveromyces polysporus*, *Kluyveromyces lactis*, *Kluyveromyces waltii* and *Ashbya gossypii* (Figure 1.1) (Goffeau et al. 1996; Dietrich et al. 2004; Dujon et al. 2004; Kellis, Birren, and Lander 2004; Scannell et al. 2007).

1.1.1 Inferring Gene Function

By recognising the similarity of protein coding gene sequences between species it is possible to assign putative protein functions to genes. This can be accomplished in an organism that does not have functional protein data by finding orthologues of that gene in a model organism for which functional data is available, *e.g.* inferring the putative functions of genes in a newly sequenced yeast species from the functions of their homologues in *S. cerevisiae* (Dietrich et al. 2004; Dujon et al. 2004). These extrapolations of gene function across species can be inaccurate however, and incorrect annotations can be further propagated to other genes with similar sequences, a process which is called “annotation transfer” (Bork and Bairoch 1996). Different gene families tend to be present and conserved at different levels across the tree of life. Some families are vital to basic cellular processes can be highly conserved from higher eukaryotes to bacteria, while other gene families may be species or genus specific. Inferences about protein function can also be made by examining gene expression, as genes that function in the same cellular process are expected to be co-expressed (Eisen et al. 1998).

1.1.2 *Genome Content and Structure Evolution*

In comparing the whole genome sequences of relatively closely related organisms like the hemiascomycete yeasts, orthologue assignment between species is relatively easy due to a high degree of protein sequence and gene synteny conservation among many of the species (Wong, Butler, and Wolfe 2002; Dietrich et al. 2004; Dujon et al. 2004; Kellis, Birren, and Lander 2004). In cases where full genome data is available, it is possible to track the position of most orthologues in the genomes by their syntenic context in addition to their protein sequence identity. In some cases, orthologous relationships between genes of different species can be inferred where the genes have diverged to a degree where they are not identified as orthologues by sequence based methods (Byrne and Wolfe 2005). In these cases, it is often a combination of shared syntenic location, similar gene lengths and weak BLAST hits or a chain of BLAST hits between several genes at the same syntenic location that allow an inference of orthology (Byrne and Wolfe 2005). It is not possible to track genes which lie at the telomeres as telomeric gene orders and contents change too rapidly for comparative approaches of this phylogenetic depth (Kellis et al. 2003). Knowledge of the positions of all non-telomeric genes in a genome as well as shared syntenic regions across species allows for inferences about the structural changes of a genome in different lineages. If there is a disruption in gene order between two species, then it can be inferred that a rearrangement occurred at that position in one of their lineages. If the comparison of the same region is made between multiple species then it may become possible to infer which lineage the rearrangement occurred in, and the ancestral order for the species compared. This depends on the presence or absence of the rearrangement in the different species and the relationship of the species to each other in the phylogeny. One major advantage of using comparative genomics in hemiascomycete yeasts to track the structural evolution of genomes is that the phylogeny contains a whole genome duplication (WGD; Section 1.4) (Wolfe and Shields 1997). This created two sets of identical chromosomes which evolved independently through rearrangements, essentially increasing the comparative power both within and between species in terms of gene order context. Unfortunately the large amount of gene loss that followed the WGD also obscured much of the orthology created in the WGD.



† Partial genome sequence data for *Z. rouxii* was generated for this study.

* Species used for genomic comparisons in this work.

Figure 1.1 Approximate phylogenetic relationships of the fungal species with publicly available sequenced genomes of more than 3X coverage. Partial sequencing of *Zygosaccharomyces rouxii* (marked in red) was conducted for Chapter 3 and 4 of this thesis. Modified from (Wolfe 2006) to include species sequenced since the original publication, and to separate the *Candida* clade from the other pre-WGD species (Rokas et al. 2003). A whole genome duplication (WGD) event is marked with an arrow. The original figure was a composite of several fungal phylogenies, is not drawn to scale and may contain some errors.

Table 1.1 Sequenced hemiascomycete species genomes

	Organism	Size (Mb)	Coverage	Chromosomes	Annotated Genes	Isolation Source	Reference
Post-WGD	<i>Saccharomyces cerevisiae</i> S288C	12.2	Finished	16	5616	rotten fig	Goffeau et al. 1996
	<i>Saccharomyces cerevisiae</i> YJM789	11.8	10X	16	5904	Lungs of an AIDS patient	Wei et al. 2007
	<i>Saccharomyces cerevisiae</i> RM11-1a	~12	10X	16	~6000	Vinyard	Mortimer 2000
	<i>Saccharomyces paradoxus</i>	11.6	7.7X	16	4788	Tree exudate/Soil	Kellis et al. 2003
	<i>Saccharomyces mikatae</i>	11.2	5.9X	16	4525	Soil*	Kellis et al. 2003
	<i>Saccharomyces kudriavzevii</i>	10.6	3.4X	16	3813	Decayed leaf*	Kellis et al. 2003
	<i>Saccharomyces bayanus</i>	11.3	6.4X	16	5857	<i>Mesophylax adopersus</i> (insect)*	Kellis et al. 2003
	<i>Candida glabrata</i>	12.3	Finished	13	5181	Human faeces*	Dujon et al. 2004
	<i>Saccharomyces castellii</i>	11.1	3.9X	9	5596	Soil*	Cliften et al. 2003
	<i>Kluyveromyces polysporus</i>	14.7	7.8X	13	5652	Soil*	Scannell et al. 2007
Pre-WGD	<i>Zygosaccharomyces rouxii</i>	12.8	-	7	331	Concentrated black-grape must*	Chapter 4
	<i>Saccharomyces kluyveri</i>	10.2	3.6X	7	2970	<i>Drosophila pinicola</i> *	Cliften et al. 2003
	<i>Kluyveromyces waltii</i>	10.7	8X	8	5230	Holly tree exudate*	Kellis et al. 2004
	<i>Kluyveromyces lactis</i>	10.6	Finished	6	5327	Creamery*	Dujon et al. 2004
	<i>Ashbya gossypii</i>	9.2	Finished	7	4716	Cotton	Dietrich et al. 2004

* Information from CBS website (<http://www.cbs.knaw.nl/yeast/biologics.aspx>)

1.2 Hemiascomycete Yeasts

1.2.1 Species Studied

Yeast is a generic term that is used to describe species of fungi that maintain a single-celled lifestyle (Dujon 2006). Several branches of the fungal phylogenetic tree contain these single celled fungi. The hemiascomycetes or Saccharomycotina are a subgroup of ascomycete fungi that contain most of the identified yeasts to date. The hemiascomycetes are also known as the “true yeasts” or “budding yeasts” because they maintain a single celled lifestyle and generally reproduce through budding, as opposed to the majority of other fungi which are filamentous (Dujon 2006). At the time of writing at least 19 publicly available hemiascomycete genomes have been sequenced to 3X coverage or greater (Figure 1.1) (Wolfe 2006; Scannell et al. 2007). Several of these species are relevant to the work presented in this thesis, and I will restrict the information presented here to those species. The research presented here uses the whole genome duplication (WGD) in yeasts as a central point from which to elucidate the evolution of *S. cerevisiae* and the other post whole genome duplication yeasts, so I used yeasts whose lineages lie

phylogenetically close to or arose from the WGD event for genomic comparisons (Figure 1.1, species marked with “*”). The lineages studied that diverged from the *S. cerevisiae* lineage before the WGD are *Kluyveromyces waltii* (Kellis, Birren, and Lander 2004), *Kluyveromyces lactis* (Dujon et al. 2004) and *Ashbya gossypii* (Dietrich et al. 2004), with some partial genomic sequence data from *Saccharomyces kluyveri* (Cliften et al. 2003) and *Zygosaccharomyces rouxii* (this study). The post-WGD species used are *Saccharomyces cerevisiae* (Goffeau et al. 1996), *Candida glabrata* (Dujon et al. 2004), *Saccharomyces castellii* (Cliften et al. 2003), *Kluyveromyces polysporus* (Scannell et al. 2007). Some comparisons also used the *Saccharomyces sensu stricto* species *Saccharomyces paradoxus* (Cliften et al. 2003; Kellis et al. 2003), *Saccharomyces mikatae* (Cliften et al. 2003; Kellis et al. 2003), *Saccharomyces kudriavzevii* (Cliften et al. 2003) and *Saccharomyces bayanus* (Cliften et al. 2003; Kellis et al. 2003) to elucidate some of the more recent genomic evolution between these species and *Saccharomyces cerevisiae*.

The phylogeny of the sequenced hemiascomycete yeasts is presented in Figure 1.1. The range of habitats and physiologies presented by the hemiascomycete yeasts studied is varied (Table 1.1), and the evolutionary distances spanned are comparable to that of the phylum of chordates (Dujon et al. 2004; Dujon 2005; Dujon 2006). Even the *Saccharomyces sensu stricto* species represent an evolutionary divergence that is estimated to be comparable to that found in the order of mammals (Kellis et al. 2003; Dujon et al. 2004; Dujon 2005; Dujon 2006). Due to the large range of divergence and physiologies presented by the hemiascomycete yeasts, comparative studies promise to uncover the mechanisms of both the genomic differences and similarities between the species which affect their individual physiologies and ecologies.

1.2.2 Genome Structure

As mentioned above, one of the main reasons that *Saccharomyces cerevisiae* was the first eukaryotic organism to have its genome fully sequenced was its relative lack of complexity and compact genome when compared to other eukaryotic species. The *Saccharomyces cerevisiae* genome contains slightly more than 14Mb (including an unsequenced rDNA

repeat) of DNA on 16 chromosomes, and an estimated gene complement of approximately 5500 genes (Goffeau et al. 1996; Kellis et al. 2003). Approximately 72% of the genome is protein coding, and the average intergenic region size is less than 1kb (Goffeau et al. 1996; Byrnes, Morris, and Li 2006; Wolfe 2006). There are approximately 255 spliceosomal introns in 250 (4.5%) protein coding genes, which is a very low number for a eukaryotic organism (Goffeau et al. 1996; Lopez and Seraphin 2000). There is also a very low number of transposable elements in the genome, with an estimated 51 Ty retrotransposons and 280 solo Ty LTRs, in total making up 3.1% of the yeast genome (Kim et al. 1998). In total 85% of the yeast genome is transcribed in rich media, the additional 13% to the protein coding genes being RNA genes such as tRNAs, snRNAs and rRNAs (David et al. 2006). The remainder of the structural genome is composed of telomeres, centromeres, ARS elements (origins of replication) and regions involved in mating-type switching (see below).

The other hemiascomycetes species studied have similar genome sizes and structures to *S. cerevisiae* (Table 1.1). The pre-WGD species tend to have smaller genomes and gene counts than the post-WGD species. The reason for this is the whole genome duplication itself, which will be discussed at length later in Section 1.4.

1.3 The *MAT* Locus

The primary method of reproduction in the hemiascomycetes is by mitotic cell division, which is achieved by budding. In this process, a protrusion appears on the mother cell, into which an identical copy of the genome is placed, and which eventually separates to become a new identical daughter cell. Most if not all of the hemiascomycetes also are capable of sexually mating, although parts of the mating process may only occur during times of cell stress such as nutrient starvation (Kunz, Taylor, and Haynes 1985; Herskowitz 1988). In yeasts there are three cell types that are defined by the sequence that is at the *MAT* locus (Hicks 1977; Hicks, Strathern, and Klar 1979; Haber 1998; Haber 2002). Two of these cell types are the **a**- and α -type haploid cells. The variants at the *MAT* locus, *MATa* and *MAT α* , are referred to as idiomorphs, because unlike alleles they differ in size, sequence and gene-

content (Metzenberg and Glass 1990). It is only possible for **a**-type haploid cells to mate with α -type haploid cells and vice versa. When two cells mate, they form the third cell type; an **a/a** diploid cell which can undergo meiosis and spore formation, something that cannot occur in haploid cells. *Saccharomyces cerevisiae* is generally a homothallic yeast (although heterothallic strains have been isolated), which means one cell can give rise to diploid progeny (by budding and mating between the mother and daughter cells) that are capable of undergoing meiosis (Herskowitz 1988). Mother-daughter mating is possible because mothers can switch mating-type. There is only one genetic difference between homothallic and heterothallic strains of *S. cerevisiae*, which is that homothallic strains carry a working copy of a gene called HO, while heterothallic strains carry a defective version of the gene (Herskowitz 1988). The Ho protein plays an important role in mating-type switching, which is why heterothallic strains cannot switch their mating-type, and thus require contact with a haploid cell of opposite mating-type for mating to occur.

There are three mating cassettes, all located on chromosome III of *S. cerevisiae*: the *MAT* locus which lies on the right arm of chromosome III, and one silent cassette for each mating type at the left (*HML α*) and right (*HMR α*) arms at either telomere of chromosome III (Hicks 1977; Hicks, Strathern, and Klar 1979; Haber 1998; Haber 2002). The silent cassette boundaries are defined by homology to sequence at the *MAT* locus. There is also a small locus control region for mating-type switching which affects recombination during mating-type switching along the left chromosome arm called the Recombination Enhancer (RE) (Bressan, Vazquez, and Haber 2004). Haploid cells can undergo mating-type switching by replacing the sequence at the *MAT* locus with the sequence of the opposite mating-type idiomorph from one of the silent loci. This model of mating-type switching is called the cassette model and was proposed by Hicks *et al.* in 1977. Only cells that have budded at least once can switch mating type, which means that a mother cell will always be a nearby daughter cell with the opposite mating type with which it can mate (Herskowitz 1988; Haber 1998; Bressan, Vazquez, and Haber 2004). Once mating has occurred, the diploid cell formed can reproduce mitotically or undergo meiosis to produce haploid spores.

Each cassette in *S. cerevisiae* contains genes which act as transcription factors for other genes expressed specifically by each mating-type when they are located at the *MAT* locus. The *MATa* idiomorph contains a gene called *MATa1*, and the *MAT α* idiomorph contains two genes, *MAT α 1* and *MAT α 2*. Silent copies of the α idiomorph sequence are normally present at the HML, and silent copies of the **a** idiomorph at the HMR. These genes perform different functions depending on whether the cell is in a haploid or diploid state. In the diploid cell no mating occurs, and all of the mating-type genes are in a silenced state (Tsong et al. 2006). This silencing in diploids occurs via the interaction of the gene products of the **a1** gene and the α 2 gene, which are both homeodomain proteins, to form a heterodimer with an increased DNA binding affinity relative to either protein in monomeric or homodimeric form. This heterodimer binds to an upstream operator of haploid specific genes (*hsgs*) and represses them (Goutte and Johnson 1993; Goutte and Johnson 1994; Haber 1998).

In haploid *S. cerevisiae* cells the **a**-type genes are constitutively expressed (Haber 1998; Haber 2002). The α 2 gene is required for the repression of constitutive **a**-specific genes (*asgs*) and its gene product functions by forming a homodimer homeodomain protein that binds to an operator upstream of *asgs*. The α 1 gene is required for the expression of α -specific genes (*asgs*) in a *MAT α* cell and its gene product is a transcription co-activator that binds cooperatively with the Mcm1 transcription factor to transcribe α -specific genes (Hagen et al. 1993; Goutte and Johnson 1994; Haber 1998). The **a1** gene appears to play no role in haploid gene expression or repression in *S. cerevisiae*. In the pre-WGD species the **a** idiomorph contains a second gene called *MATa2* which codes for a HMG domain protein and has been lost in the post-WGD species (Dranginis 1989; Butler et al. 2004). It is thought that **a2** gene was lost around the same time as the WGD, and occurred through a rewiring of the control of the mating-type specific gene expression network (as discussed below) (Butler et al. 2004; Tsong et al. 2006).

Recently an evolutionary study into the roles of **a**- and α -type genes has provided a likely explanation for the loss of the **a2** gene in post-WGD species (Tsong et al. 2006). In an experimentally validated multiple comparative analysis between hemiascomycete yeast

species that lie between *S. cerevisiae* and *C. albicans* in the phylogenetic tree, Tsong *et al.* reported strong evidence that the regulation of **a**- and α -type genes has changed significantly in the *S. cerevisiae* lineage (Tsong *et al.* 2006). In a pre-WGD ancestor of *S. cerevisiae* and *C. albicans*, the roles of $\alpha 1$ in haploid cells and the **a1**- $\alpha 2$ heterodimer in diploid cells were the same as in *S. cerevisiae*. However, the **a**-type genes were not constitutively expressed and the **a2** gene was required to positively co-regulate *asgs* as a heterodimer with Mcm1. In addition to this, the $\alpha 2$ gene did not repress *asgs* like in *S. cerevisiae*. Over evolutionary time two basic steps were required to transition to the regulatory functions observed in *S. cerevisiae*; firstly, the expression of the *asgs* genes became independent of the ancestral **a2** activator, and secondly, the $\alpha 2$ gene became a repressor of *asgs*. These transitions were made possible by slightly modifying the *cis*- and *trans*-elements of the *asgs* to constitutively bind the ubiquitous transcription factor Mcm1 along with concomitant substitutions of some of the residues of the Mcm1 protein to alter its heterodimerisation affinity from **a2** to $\alpha 2$ (Tsong *et al.* 2006). Interestingly it appears that the *K. lactis* lineage may represent an intermediate state in the evolution of this system where *asgs* are positively regulated by **a2**-Mcm1 heterodimers and repressed by $\alpha 2$ -MCM1 heterodimers (Tsong *et al.* 2006).

1.3.1 Silent Cassette structure

The structure of the silent cassettes in *S. cerevisiae* is described in terms of 5 regions; W, X, Y, Z1, Z2 (Figure 1.2) (Haber 1998; Haber 2002). These regions are defined through comparisons between the **a**- and α -cassettes. Each cassette type is nearly identical in the X and Z1 regions, and they are primarily differentiated by the Y region (**Ya** and **Y α**). The W and Z2 regions are only found in the α -cassettes (Figure 1.2, grey regions), and have been defined by comparing the HML α and the *MAT* locus which contains the α -cassette in the sequenced type-strain of *S. cerevisiae* (Haber 1998; Haber 2002). The $\alpha 1$ and $\alpha 2$ genes are transcribed divergently out of the **Ya** region present only in the α idiomorph. These genes overlap the Z1 and X regions respectively. The **a1** gene is located completely within the

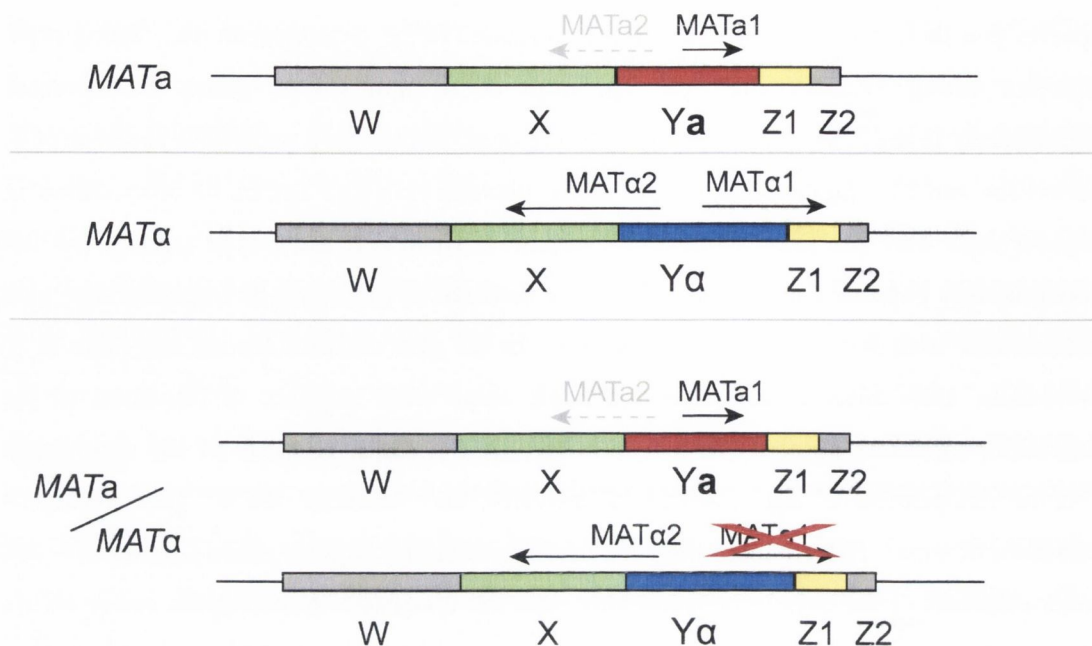


Figure 1.2 Schematic diagram of the structure of the *MAT* locus idiormorphs, redrawn from (Haber 2002). The *MATa* and *MATα* idiormorphs are distinguished by their *Ya* (650 bp) and *Ya* (750 bp) regions and the expression of *MAT* genes in different cell types (red and blue regions respectively). The X (green) and Z1 (yellow) regions are identical between the cassettes. The W and Z2 regions (grey) are defined by homology between the *MAT* locus and the *HMLα* silent cassette, and are not present in the *HMRa* silent cassette. In a haploid *α*-type cell, *MATα1* gene encodes a coactivator of *α*-specific genes that acts in conjunction with Mcm1, while *MATa2* encodes a corepressor of *a*-specific genes that also acts with Mcm1. In a *MATa/MATα* diploid, *MATa1* and *MATα2* encode proteins that act as a corepressor of haploid specific genes, including *MATα1* (red X). The *MATa2* gene (represented by a grey dotted arrow) is present in some pre-WGD species, but has been deleted from all of the post-WGD species.

Ya region which is present only in the *a* idiormorph. In the pre-WGD species the *a1* and *a2* genes are transcribed divergently like the *α* idiormorph genes.

The silent cassettes at the HML and HMR loci are rendered transcriptionally silent by the deacetylation of nucleosomes positioned along the lengths of the silent cassettes (Haber 1998; Haber 2002). This heterochromatin configuration is maintained through binding between 2 *cis*-elements (I and E) that lie on either side of both silent cassettes with a large number of *trans*-acting factors. These *trans*-acting factors include the transcription factors Rap1 and Abf2, histone proteins, the replication Origin Recognition Complex (ORC), the

Silent Information Regulator Proteins (Sir1-Sir4), deacetylases, transacetylases and chromatin assembly factors (Sherman and Pillus 1997; Haber 1998; Haber 2002).

1.3.2 Mechanism of mating-type switching

Mating-type switching in *S. cerevisiae* has become a paradigm for models of mitotic recombination and the study of double strand break (DSB) repair (Paques and Haber 1999; Haber et al. 2004; Haber 2005). Mating-type switching is essentially a directional gene conversion event whereby the sequence existing at the *MAT* locus is replaced by sequence information from the silent version of the other type of cassette at the appropriate HM-locus. In *S. cerevisiae* this process is initiated by a double strand break at the *MAT* locus caused by the haploid specific endonuclease Ho (Herskowitz 1988; Haber 1998). Ho is an endonuclease that binds to a 24 bp recognition sequence that spans the Y/Z1 boundary and cleaves the DNA in the Z region adjacent to the Y/Z1 boundary (Step A, Figure 1.3) (Nickoloff, Chen, and Heffron 1986). The silent chromatin state at the HM-loci means that the recognition sequence in them cannot be cleaved by the HO endonuclease, thus ensuring the directionality of the gene conversion event from silent HM cassette to *MAT* locus. The HO endonuclease creates a 4 bp 3' overhang, which is extended to long 3' tails on either side of the DSB by a 5' to 3' exonuclease consisting of the Rad50, Mre11 and Xrs2 proteins (Haber 1998; Haber 2002). These 3' tails interact with several recombination proteins that initiate strand invasion at homologous sequence (in this case the homologous Z1 region of the silent cassette) (Haber 1998; Haber 2002). The 3' end of the invading strand acts as a primer to initiate the DNA synthesis to copy the Y region of the donor cassette, which will replace the Y region of the *MAT* locus.

In the classic DSB repair model of Szostak *et al.*, one donor strand invades and becomes the template for new DNA synthesis and is retained at the recipient locus while the other strand of the donor acts as a template for new DNA synthesis to recover the double stranded donor locus (Paques and Haber 1999; Haber et al. 2004). In this model, both donor and recipient loci are made up of one original donor strand and one newly synthesised strand, thus the synthesis of DNA is symmetrical. It has been experimentally

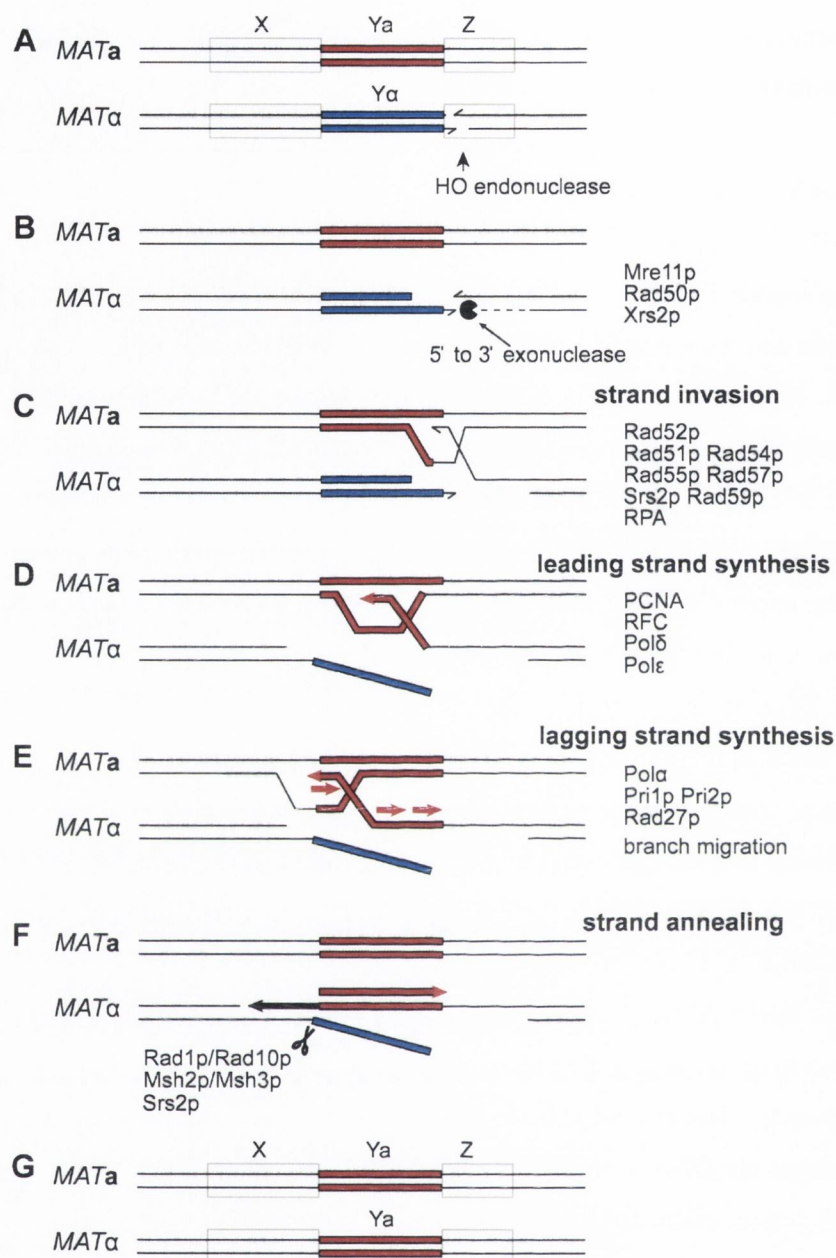


Figure 1.3 A synthesis-dependent strand-annealing model for *MAT* switching, redrawn from (Haber 2002). Proteins involved at different stages of the process are shown. In this mechanism, a DSB is converted into single stranded 3' tails. *RAD52* strand invasion allows the invading 3' end to initiate new DNA synthesis, requiring RFC, PCNA, and either Polδ or Polε. At the same time, a second strand is synthesised by lagging-strand replication, requiring the Polα and primase. Unlike normal semiconservative DNA replication, it is proposed that the newly synthesised strands are displaced from the template by branch migration. Removal of the original Y-region is catalysed by a flap endonuclease (Rad1/Rad10, Msh2/Msh3 and Srs2).

verified that the donor sequences remain unchanged during *MAT* switching and the gene conversion at the *MAT* locus involves the conservative inheritance of both newly synthesised DNA strands (Hicks 1977; Haber 1998). After the strand synthesis of the new cassette at the *MAT* locus, strand annealing of the new sequence must occur. At this point, there is still a single stranded copy of the old Y region attached to the chromosome which must be removed before strand annealing. The removal of this single stranded region is thought to occur through the action of a flap endonuclease made up of several protein components including Rad1, Rad10, Msh2, Msh3 and Srs2 (Haber 1998). Once this is removed, the annealing of the newly synthesised Y region of the *MAT* cassette can occur and mating-type switching is complete.

In the most recent models (known as Synthesis Dependent Strand Annealing models) of mating-type switching, the strand invasion creates a replication fork where there is both leading and lagging strand synthesis (Steps D and E, Figure 1.3) (Haber et al. 2004). In this model there is an asymmetric synthesis of new DNA, where the newly synthesised strands associate and anneal together at the *MAT* locus (all of the DNA replaced at the *MAT* locus is newly synthesised DNA) while the donor strands reassociate and remain totally unchanged by the process (Haber et al. 2004). It is hypothesised that the association of the new strands together occurs by the migration of the Holliday junction that follows the replication fork. Interestingly, this mechanism precludes the formation of two stable Holliday junctions that can be resolved by crossing over, and this lack of crossing over during mating-type switching is the reason for synthesis of new DNA only at the recipient locus (Haber 1998; Haber et al. 2004).

1.3.3 Evolution of the *MAT* Locus System

The hemiascomycete yeasts show variation in the basic mechanism of mating-type switching as well as in the control of the expression of mating-type specific genes as in the example of the evolution of the control of *asg* expression between *C. albicans* and *S. cerevisiae* above. Single locus, two-allele homothallic mating systems are found as distantly from the hemiascomycetes as the archaeascomycete *Schizosaccharomyces*

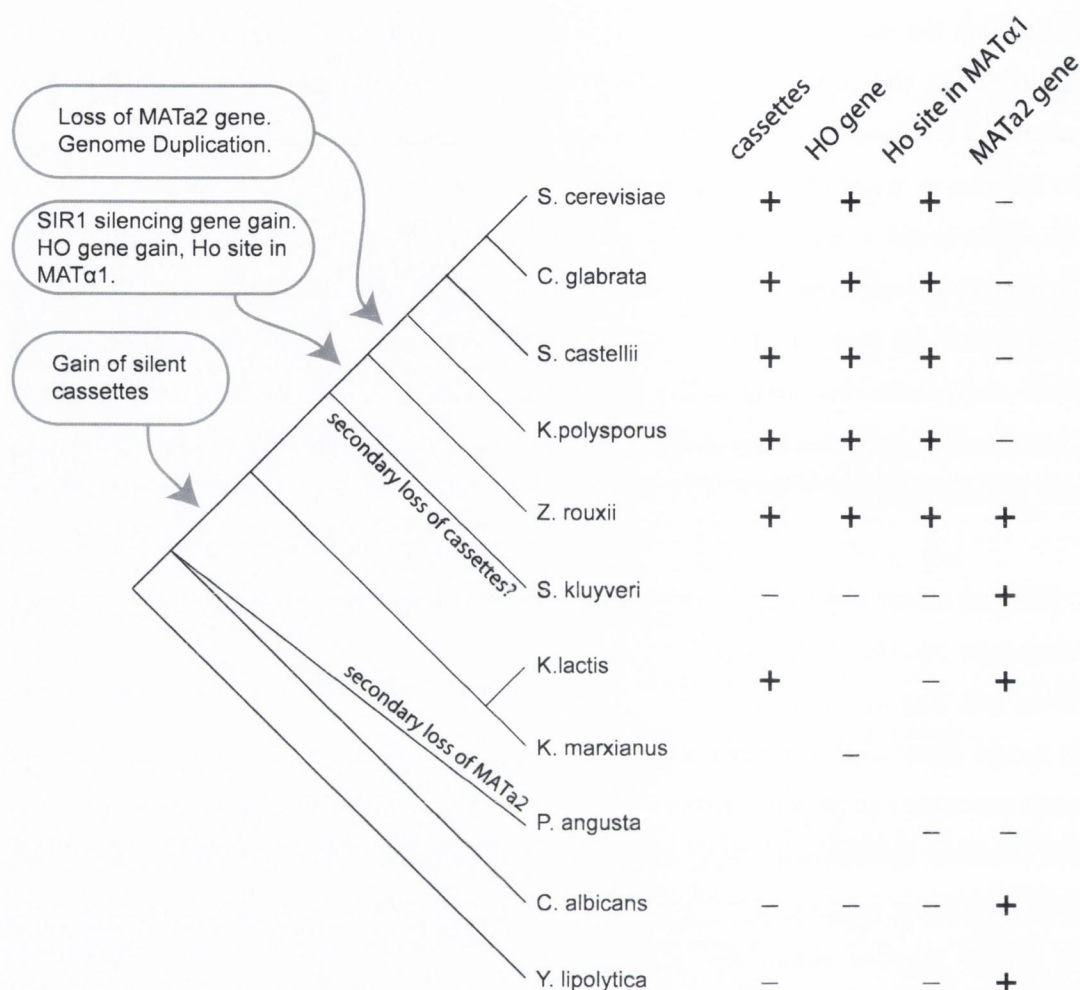


Figure 1.4 Summary of events in *MAT* and *HO* evolution. The tree on the left is schematic and is not drawn to scale. Figure is modified from (Butler et al. 2004) to include *K. polysporus* and to correct the topology of the tree.

pombe (Yamada-Inagawa, Klar, and Dalgaard 2007). Certain features of the mating-type switching found in *S. cerevisiae* have evolved recently within the hemiascomycete lineage.

The silent cassettes are found in species that phylogenetically lie between *S. cerevisiae* and *K. lactis*, *A. gossypii* and *K. waltii*, but not in more distantly related hemiascomycetes (Figure 1.4) (Butler et al. 2004; Fabre et al. 2005). The *HO* endonuclease gene was gained prior to the speciation of the *Z. rouxii* lineage before the whole genome duplication (and putatively as early as the *Kluyveromyces thermotolerans* ancestor (Fabre et al. 2005)). All

the species that contain the HO gene also have the 24 bp recognition site at the *MAT* locus (Butler et al. 2004). Interestingly it appears that silencing at the HM-cassettes was also modified at around the same time, as SIR1 (one of the genes involved in HM-silencing) was gained prior to the *Z. rouxii* lineage speciation (Figure 1.4; Chapter 4) (Haber 2005).

1.4 Whole Genome Duplication

1.4.1 Ohno Hypothesis

In 1970 a seminal book was published by the late Susumu Ohno, in which he proposed that gene and genome duplication were the primary forces that drove the evolutionary development of complexity in organisms (Ohno 1970). Rather than evolution occurring purely through mutations in gene sequences, he theorised that genetic novelty was primarily a consequence of gene duplication. Through duplication the redundancy created would allow one copy to retain the biological function of the gene while the other could accumulate an array of mutations that could modify the function of the gene. In this way, biological complexity could increase without necessarily altering vital cellular pathways in a detrimental way for the cell. While the most frequent outcome of this process would be the non-functionalisation of the gene as proposed by Haldane (1933), rare cases would occur where beneficial mutations might arise that selection could act upon. This model of gene duplication and asymmetric divergence is currently known as *neofunctionalisation*. In his book Ohno hypothesised that genome duplication had played a large role in the evolution of vertebrates in particular. The most popular form of this hypothesis known as the 2R hypothesis is that there were two rounds of duplication during early vertebrate evolution. While the 2R hypothesis has been a contested issue in molecular evolution, the increasing number of sequenced vertebrate genomes has recently provided strong evidence for the 2R hypothesis (Dehal and Boore 2005; Kasahara 2007; Nakatani et al. 2007). Furthermore it is beyond doubt that additional genome duplications have occurred during the evolution of some vertebrate groups, such as teleost fish (Jaillon et al. 2004).

1.4.2 WGD in hemiascomycetes

In 1987, it was first hypothesised that the *S. cerevisiae* genome might have undergone a whole genome duplication (Figure 1.5) during its evolution based on the pattern of duplicated histone genes, all of which are arranged in sets of duplicated adjacent pairs (HTA1-HTB1, HTA2-HTB2, HHT1-HHF1, HHT2-HHF2) (Smith 1987). Soon after the sequencing of the *S. cerevisiae* genome it was found that there were many duplicated genes in the genome. In 1997 Wolfe and Shields produced evidence that the extensive duplicates in the yeast genome were created by a whole genome duplication (Wolfe and Shields 1997). They found that the genome could be described as containing blocks of duplicate genes that covered at least 50% of the yeast genome. Due to a large amount of gene deletion, many of the genes located within the duplicate blocks were not actually in a duplicate gene pair, which accounted for only 25% of the genes in the duplicated blocks. It is now thought that only 8% of genes in *S. cerevisiae* are maintained in duplicate (Scannell et al. 2006; Scannell et al. 2007). Phylogenetic trees showed that the duplication event had occurred at some point after the speciation of the *Saccharomyces* genus from the *Kluyveromyces* genus, and dated the duplication crudely to approximately 100 million years ago. They also suggested that the duplication may have been instrumental in the evolution of an anaerobic phase in the yeast physiology and may have occurred at around the same time as the angiosperms arose and proliferated.

Full validation of the WGD in yeasts came with the publishing of the genome sequence of the pre-WGD yeast *Kluyveromyces waltii* (Kellis, Birren, and Lander 2004). The genome of *Saccharomyces cerevisiae* maps in a 2:1 fashion in 253 double-conserved synteny (DCS) blocks when compared to *K. waltii*. These blocks cover 88% of the *S. cerevisiae* genome, providing strong evidence that the paralogous genes in *S. cerevisiae* were all derived from one duplication event. It was also found that extensive gene deletions among the duplicates created a situation where the duplicated regions each contain a subset of the genes found in the orthologous region of *K. waltii* with only a minority of gene pairs are retained in duplicate. Similar patterns of “double-conserved synteny” were reported in the comparison of *S. cerevisiae* with other pre-WGD yeasts such as *Ashbya gossypii* (Dietrich et al. 2004)

and *Kluyveromyces lactis* (Dujon et al. 2004) which were sequenced just after *K. waltii*, and indeed the same pattern emerges when comparing any of the other post-WGD yeast genomes to a pre-WGD genome (Byrne and Wolfe 2005).

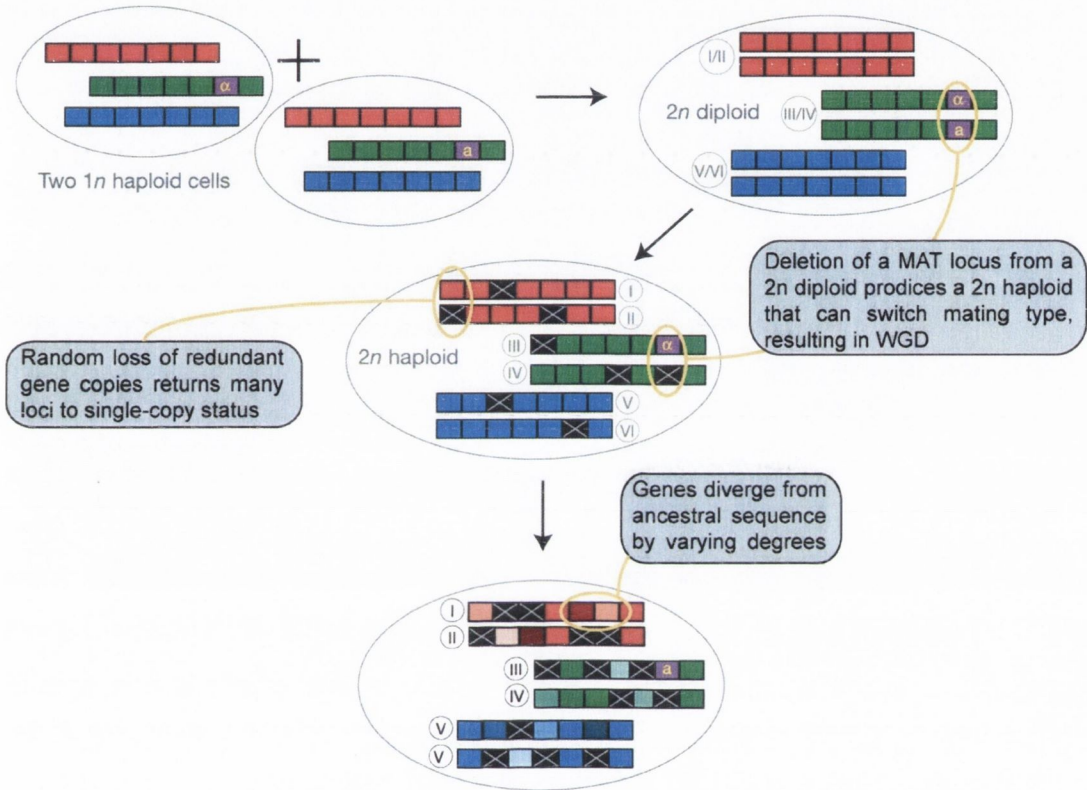


Figure 1.5 Schematic diagram of the likely events that occurred in the yeast lineage that underwent a WGD. Loss of a *MAT* locus from a 2n diploid created a 2n haploid capable of mating-type switching. Subsequent gene loss of most redundant genes gave rise to the interleaved pattern of duplicate genes across sister chromosomes seen in extant post-WGD species. This figure is modified from (Scannell et al. 2006).

1.4.3 Fate of WGD duplicate genes

Evolutionary theory considers that a duplicated gene can have one of four fates. The WGD ancestor presented the organism with an enormous evolutionary opportunity in the form of

an extra copy of every gene. The most common outcome of gene duplication is the subsequent *deletion* of one of the duplicates. Extensive loss is observed in *S. cerevisiae* where only approximately 8% of genes have been retained in duplicate from the WGD, while 91% were retained in 1 copy and 1% were lost in double copy (Wolfe and Shields 1997; Scannell et al. 2006; Scannell et al. 2007). The other three gene fates all involve the retention of both genes. For this to occur, it is thought that both genes must be functionally important or else one of them would have been deleted in the buffering presence of the other.

Firstly, it is possible that duplicates are retained for dosage effects. This can occur where ancestral genes which interact in a stoichiometric fashion are duplicated and are retained in duplicate for maintenance of the ratios between the gene products, a process known as the balance hypothesis (Papp, Pal, and Hurst 2003). Selection for increased flux in relevant pathways during times of cell stress or nutrient limitation can also result in dosage selection to amplify the copy number of individual genes (Kondrashov and Kondrashov 2006; Conant and Wolfe 2007). There are well known cases of drug resistance in yeast and other organisms occurring through the amplification of drug target and transporter genes (Leppert et al. 1990; Nourani et al. 1997). Ecological adaptations to environmental stresses through the duplication of environmental response genes such as metallothioneins in the adaptation to heavy metals can also occur through the selection for increased dosage (Fogel and Welch 1982; Karin et al. 1984; Kondrashov and Kondrashov 2006).

A recent study by Conant and Wolfe in yeast shows compelling evidence of the retention of some ohnologues (paralogues created by a WGD) for dosage effects in the glycolytic pathway (Conant and Wolfe 2007). They hypothesised that the large-scale deletion of redundant duplicates in the wake of the WGD afforded a relative increase in concentration of rate limiting enzymes in certain metabolic pathways in *S. cerevisiae* such as glucose metabolism. They found that there was an excess of retained WGD duplicates involved in glycolysis compared to the genomic average. By using models of *S. cerevisiae* metabolism they demonstrated that many of these retained duplicates are indeed at rate limiting steps in glycolysis (such as hexose transporters), and increased concentrations of these enzymes

increases the glycolytic flux. They also found evidence that the higher levels of pyruvate created by glycolysis increase flux through fermentation, especially in high glucose growth conditions. This adaptation may have been vital to the evolution of the Crabtree effect in *S. cerevisiae*, which allows a high growth rate when glucose is in excess by fermenting additional pyruvate when the respiratory metabolism is saturated (Piskur et al. 2006). If this dosage effect on glycolysis occurred soon after the WGD (as expected by the high rate of duplicate gene loss; (Scannell et al. 2006), then it may have allowed the WGD species to out-compete other pre-WGD yeasts in conditions of high glucose. Thus it seems plausible that the retention of some duplicates may indeed be the result of selection for increased dosage of certain proteins.

The other two major gene fates are *neofunctionalisation* and the more recent model of *subfunctionalisation* (Hughes 1994; Lynch and Force 2000a). Neofunctionalisation (as mentioned in the description of Ohno's work above) theoretically occurs when the redundancy created between duplicated genes allows one of the pair to become mutationally labile while the other one is constrained to maintain the ancestral function of the genes. For neofunctionalisation to occur the diverging copy must accumulate adaptive mutations which allow it to perform a new function. In subfunctionalisation, both duplicates theoretically accumulate detrimental mutations but in a manner that they can complement each other. In this manner they can partition their functions to perform the ancestral function (Force et al. 1999; Lynch and Force 2000a). It is considered that neofunctionalisation might be less frequent than subfunctionalisation because it relies on positive selection acting on rare gain-of-function mutations while subfunctionalisation can occur in a selectively neutral manner (Force et al. 1999).

Putative support for both neofunctionalisation and subfunctionalisation has been observed in genomic studies, although concrete examples of either are rare due to a lack of functional evidence, particularly about the ancestral states of genes. It is still under debate whether either of these models is applicable to real cases of duplicate gene retention, and if so, at what frequency. One of the major problems with differentiating between these models is that there is an overlap between their expected indicators. One of the key predictions for

genes that undergo neofunctionalisation is a rate asymmetry in their evolution (Kellis, Birren, and Lander 2004; Lynch and Katju 2004; Byrne and Wolfe 2007). It has been shown that this is indeed the case in many of the ohnologues created by the WGD in yeast. Byrne and Wolfe have recently shown that 89% of retained WGD duplicates shared between *S. cerevisiae* and *S. castellii* show at least moderate rate asymmetry, and 27% show strong asymmetry (Byrne and Wolfe 2007). This is similar to previous analyses of duplicate asymmetry which show that 20-30% of duplicates in various organisms show asymmetric sequence evolution (Conant and Wagner 2003; Kellis, Birren, and Lander 2004). Byrne and Wolfe also showed that this asymmetry is consistent across species, *i.e.* that the same duplicate showed rate asymmetry in different post-WGD species, and also that in cases where the duplicates were returned to single copy, the gene that was lost tended to be the accelerated copy (Byrne and Wolfe 2007). This implies that the asymmetric evolutionary fate of each duplicate was defined soon after they were duplicated. In accordance with these findings, the accelerated duplicates also tend to be less essential and less well characterised than the slower evolving presumably ancestral type duplicate, as would be expected in cases of neofunctionalisation (Byrne and Wolfe 2007).

The lack of mutual exclusivity between the models of neofunctionalisation and subfunctionalisation has led to a lack of examples of clear cut neofunctionalisation. One good candidate is the duplication of ADH1 which gave rise to ADH2 (Thomson et al. 2005). The duplication of ADH1 occurred in the ancestor of the *Saccharomyces sensu stricto* species. The ADH1 gene generates ethanol from acetaldehyde (which is produced by the loss of CO₂ from pyruvate), a reaction that also recycles NADH to NAD⁺ in the cell. The ADH2 gene on the other hand reconverts ethanol to acetaldehyde, which gets shunted into the citric acid cycle. This detour from the path of pyruvate entering directly into the citric requires one molecule of ATP (Thomson et al. 2005). Thompson *et al.* (2005) reconstructed the ancestral protein sequence for the ADH1 and ADH2 genes, and synthesised this protein to test its enzymatic dynamics. They showed that this reconstructed ancestral protein performs the same function as ADH1. This suggests that prior to the duplication of ADH1, ethanol produced by the ancestral yeast was released into

the environment and not reused in cellular energy production. The novel function of ADH2 ethanol recycling undoubtedly played a large role in the increase in efficiency of energy usage by ethanol-producing yeast species and appears to be a clear case of neofunctionalisation.

It has been shown experimentally that several asymmetric ohnologues from the WGD in yeast (that cannot complement each other's deletion) can be complemented by a single orthologous gene from a pre-WGD species when they are both knocked out in *S. cerevisiae* (including SIR3/ORC1, HBS1/SKI7 and RNR2/RNR4) (van Hoof 2005). This observation fits the model of subfunctionalisation because it would be expected that the ancestral gene would be able to complement the deletion of only one of the ohnologues in a neofunctionalisation scenario. Asymmetry is thus not unique to neofunctionalisation. In a subfunctionalisation scenario, genes could theoretically inherit different numbers of ancestral functions which could also lead to rate asymmetry (He and Zhang 2005). Furthermore, some of the reported examples of rate asymmetry of duplicated yeast genes may not be statistically significant. A statistical analysis of yeast ohnologues identified by Kellis *et al.* (2004) showed that under an assumption of equal evolutionary rates between duplicates, there was a 32% chance of obtaining a higher deviation between them than the average deviation seen in the real data (Lynch and Katju 2004). This result shows that the observed levels of sequence asymmetry may not be compelling evidence for neofunctionalisation as the primary explanation for duplicate gene retention after a WGD.

Asymmetry can also be observed in properties other than evolutionary rate. Analyses of the protein interaction network and large-scale microarray analyses of *S. cerevisiae* have revealed that levels of redundancy (*i.e.* shared interactions) of gene duplicates decrease relatively rapidly after the duplication and that functional divergence between the duplicates is often asymmetrical. (Wagner 2001; Wagner 2002; Wagner 2003). Most of the interaction divergence between duplicates is due to degenerative mutations which cause the loss of interactions, while new interactions form at a rate estimated to be three orders of magnitude smaller than the rate of loss (Wagner 2001; Wagner 2003). This data may indicate evidence for a greater amount of *subfunctionalisation* than *neofunctionalisation* in

duplicates. However, the proportion of proteins for which protein interactions have been documented is only about 16% of the proteins in *S. cerevisiae*, so it is unclear whether this pattern of evolution is applicable to all proteins in an organism (Wagner 2001).

Another type of duplicate retention mechanism that has been observed operates by the divergence of expression profiles between duplicate genes (Lynch and Force 2000a; Kwast et al. 2002; Duarte et al. 2006). This process is sometimes referred to as expression subfunctionalisation/neofunctionalisation. In a study of the expression of different *S. cerevisiae* genes under aerobic and anaerobic conditions, at least 117 gene pairs created by the WGD showed differential expression with respect to oxygen (Kwast et al. 2002). Thus, because they are expressed under different conditions these results imply that the functions of these gene pairs may have also diverged. It is likely that expression divergence in this manner may also coincide with sequence divergence between the duplicates.

As the above examples suggest, there are still many questions regarding the actual modes of retention of duplicate genes in genomes. It is possible that neither subfunctionalisation or neofunctionalisation completely models the fates of duplicate genes, and several variants and mixtures of the different models have also been proposed (He and Zhang 2005).

1.5 Gene content evolution of genomes

1.5.1 Gene Losses

The gene content of the genomes of closely related species is usually very similar or even identical because they developed from a recent common ancestor, while it can differ greatly between distantly related species because of the extended evolutionary time between them. The gene content of organisms evolves through time through the processes of the loss and gain of genes. Gene loss is thought to account for most of the lineage specific gene repertoire differences between eukaryotes (Krylov et al. 2003). Gene loss may be extensive and occur rapidly under strong selective pressures or population genetic effects as is seen in the streamlining of pathogen genomes (Katinka et al. 2001; Dujon et al. 2004;

Keeling 2004). Gene loss most likely occurs through the process of non-functionalisation of the gene by degenerative mutations in the coding or promoter sequences of genes. If the non-functionalising mutation is not filtered out by purifying selection, the whole gene is likely to become lost from the population. Whole genome comparisons of the gene organisation of various hemiascomycete species has shown that genome size reduction (deletion) after WGD proceeded through single gene inactivations and deletions rather than large deletions of multiple adjacent genes (Byrnes, Morris, and Li 2006). The deletion of WGD duplicates also appears to have occurred in a randomly distributed manner across the genome (Byrnes, Morris, and Li 2006). The frequency of gene loss among duplicates is far greater than for single copy genes in the genome because the redundancy between duplicates allows one gene to be deleted while the other can buffer its loss (see above).

In species that underwent a WGD like in the hemiascomycete lineage, there has been extensive gene deletion, with the majority of genes returning to single copy in the genome. This is the default fate for most WGD duplicates, and much of the deletion occurred soon after the WGD event (Scannell et al. 2006; Scannell et al. 2007). The current model for the yeast WGD suggests that an ancestor with approximately 5000 genes in its genome underwent WGD to transiently have 10000 genes, followed by extensive losses to leave around 5500 genes in the current *S. cerevisiae* genome (including 500 ohnologue pairs) (Scannell et al. 2006; Scannell et al. 2007). Mass gene loss of this variety can lead to qualitative as well as quantitative differences in the gene content of organisms that speciated while the process was ongoing, such as in the post-WGD clade of the hemiascomycetes. Initially the duplicates are all theoretically identical to their ohnologue and both have an equal probability of loss. For many genes that were lost independently in different species after their divergence, a different copy was retained in each species. This should theoretically have no quantitative effect on the organisms and is likely to have been selectively neutral, but it means that the chromosomal context of the genes is different between species. It has been suggested that these reciprocal gene losses (RGLs) could lead to reproductive isolation between species as at any given reciprocal loss location there is a 0.25 chance that a spore with no copy of the gene will be formed after mating between two lineages (Scannell et al. 2006). The chance of forming viable spores decreases rapidly as

the numbers of these reciprocal loss events increases because it is proportional to $(0.75)^n$, where n is the number of loci where RGL has happened. These reciprocal loss events may have even played a role in reproductive isolation and speciation of the post-WGD hemiascomycete species (Scannell et al. 2006).

1.5.2 *Gene gains*

1.5.2.1 Duplication

Most new genes in genomes are created by the duplication of already existing genes, and it is estimated that the rate of duplication is of a similar order of magnitude as the rate of nucleotide mutation (1% per gene per million years; (Lynch and Conery 2000). Evolutionary DNA duplications can encompass regions of genes in segmental chromosomal duplications, as well as on a gene by gene basis or whole genome basis (Parry, Sharp, and Parry 1979; Wolfe and Shields 1997; Lynch and Conery 2000; Koszul et al. 2004). There are at least two qualitatively different types of single gene duplication: tandem duplication, which forms a new gene copy lying directly adjacent to the originator gene, and dispersed duplications where the duplicate gene is not linked to the originator gene. Dispersed duplications are thought to be generated either by a retrotransposition of a mRNA transcript of the precursor gene (Schacherer et al. 2004), or by DNA-based duplication (Schacherer et al. 2005), but the relative frequencies at which these two processes occur in yeast is yet unknown.

Segmental duplications originate through aneuploidy, which occurs by nondisjunction of sister chromatids during anaphase of meiosis or mitosis (Parry, Sharp, and Parry 1979). This involves the unequal partitioning of chromosomes or segments of chromosomes into the daughter cells and results in duplications of the regions involved in one daughter cell and deletions in the other cell. Extensive aneuploidy has been observed in deletion strains of *S. cerevisiae* and may be explained in some cases by possible selection for increased dosage of genes within the duplicated region (Hughes et al. 2000; Koszul et al. 2004). This may represent an easy way for a species to duplicate important genes, but many

aneuploidies are associated with genetic defects such as tumours in metazoan organisms. The duplication of many genes along a chromosomal region may lead to imbalances in the protein abundance in the cell and many aneuploid phenotypes are indicative of protein degradation and folding distress indicating that the additional proteins produced may be disruptive to cellular physiology (Torres et al. 2007). In general, aneuploidies cause a proliferative disadvantage to yeast cells which increases with the amount of extra DNA involved, and some aneuploidies are lethal to cells (Torres et al. 2007). It appears that WGD is not as deleterious in general as aneuploidy, because the uniform increase in all genes and gene products maintains their relative dosages in the cell. It is possible that the mechanism of WGD is complete nondisjunction of chromosomes during cell division, but this is uncertain. One requisite that is known about in yeast is the haploidisation of the MAT locus by deletion of one of the alleles in a diploid (Figure 1.5) (Greig et al. 2002a; Greig et al. 2002b). This allows mating type switching to occur and sexual reproduction to recommence.

Tandem duplicates are formed by unequal crossing over between sister chromosomes during prophase I of meiosis. Depending on the position of the crossover, the duplicated region may contain part of a gene, a whole gene or several genes (Zhang 2003). Tandem gene duplicates reside next to their originating sequence and maintain the same orientation and presumably the same or very similar genomic context. Duplication by retrotransposition occurs by the reverse transcription of mature mRNA to cDNA followed by random insertion into the genome (Long et al. 2003a; Zhang 2003). It has been shown that the expression of Ty elements are required for the reverse transcription and integration of retrotransposed reporter genes (Derr, Strathern, and Garfinkel 1991; Derr and Strathern 1993). This indicates that the reverse transcriptase and integrase proteins encoded on the Ty elements are likely to be involved in retrotransposition (Wilhelm, Wilhelm, and Gabriel 2005). Quantitatively these two types of duplication are generally the same, but the new genomic context of a retrotransposed duplicate gene may create qualitative differences between tandem and retrotransposition duplication (Cusack and Wolfe 2007). Retrotranspositions do not have a promoter region as they come from only the coding sequence of a gene, and as such require the acquisition of regulatory sequences

in their new location or they will quickly become a pseudogene (Long et al. 2003a). A new or absent regulatory sequence will likely cause differential expression between the parent and retrotransposed genes. Also, different parts of the genome contain chromatin domains which can differ in their conformation (Batada and Hurst 2007; Batada, Urrutia, and Hurst 2007), and movement to a new chromatin domain could also qualitatively alter the expression of the relocated duplicate. Apart from the WGD event, these types of duplication are probably the foremost creators of redundancy and genetic novelty in yeast. It is unclear whether non-tandem duplications of single genes in yeast can be formed by mechanisms other than retrotransposition.

1.5.2.2 HGT

Another possible source of new genes is the acquisition of DNA from a source other than the parent cell. These events are known as horizontal gene transfer (HGT). This type of genetic transfer is common in bacteria and protists with phagotrophic lifestyles, and has also been found to a lesser extent in certain unicellular and multicellular eukaryotes (Andersson 2005). The contributing effect of this form of genetic transfer to the evolution of yeast species appears to have been small and only a handful of candidates for HGT have been found (Gojkovic et al. 2004; Hall, Brachat, and Dietrich 2005; Woolfit et al. 2007). Tracking these events is inherently difficult because many confounding factors may interfere with their identification. These include sequence divergence, DNA contamination, lack of phylogenetic coverage to identify donor species, gene loss or the inability to determine the direction of transfer (Hall, Brachat, and Dietrich 2005).

Many of the eukaryotic organisms that exhibit HGT have come into close contact with the organisms that DNA was transferred from, for example protists that engulf their food by endocytosis (Doolittle 1998). Cases of metazoan HGT involve infection of the germ line cells, as in the case of endosymbiont bacteria *Wolbachia pipientis* (Hotopp et al. 2007). *Wolbachia* are known to infect the egg cells of at least 20% of insect species as well as filarial nematodes and are inherited maternally. Horizontal gene transfers of less than 500bp to greater than 1Mb to these host genomes has been documented (Hotopp et al.

2007). Other cellular infection agents that can contribute genes to a host are viruses (such as retroviruses), transposable elements and Alu elements that insert into the host genome (Long et al. 2003a). While elements like these are commonly known as selfish elements, they have been known to pass on new genes to their hosts such as reverse transcriptase and integrase genes, both of which may be required for the successful retrotransposition of genes in a genome (Derr, Strathern, and Garfinkel 1991; Derr and Strathern 1993).

In *S. cerevisiae* and other yeast species, small fragments of mitochondrial sequence have been transferred to the nuclear genome and are thought to integrate at sites of double strand break repair (Ricchetti, Fairhead, and Dujon 1999; Blanchard and Lynch 2000). This is also a type of HGT, albeit not from another organism, but the integration at double strand break repair sites may be a general feature of HGT or retrotransposition insertion events. One further possible source of genetic exchange in yeast are natural plasmids and viruses that many of the species harbour (Futcher 1988; Wickner 1996; Schmitt and Breinig 2002).

1.5.2.3 *de novo* Creation of Genes

A further possible route of gene acquisition is the *de novo* creation of coding genes from previously non-coding sequence through neutral mutations (Long et al. 2003a). As well as a coding ORF a gene created in this manner would also need to acquire upstream regulatory sequences for its expression. The gene would likely need to acquire a function almost immediately after its creation; otherwise there would be no purifying selection to avoid non-functionalisation by random mutations. On these grounds *de novo* gene creation is probably a rare occurrence, although at least five potential cases have been identified in the *Drosophila* genus (Levine et al. 2006) and *de novo* creation of part of a gene has also been observed elsewhere (Chen, DeVries, and Cheng 1997; Cheng and Chen 1999; Long et al. 2003a). It is not yet clear whether *de novo* gene creation has been a significant factor in the evolution of yeast genome content, but in Chapter 3 I identify more than 50 genes that have been gained in *S. cerevisiae* since the WGD that do not show evidence of duplication or horizontal gene transfer, and which may have formed *de novo*.

New genes may also arise by stitching together two existing genes into a mosaic gene, which often involves duplication of one or both of the parent genes. Evidence for this process has been observed in many metazoan species, but once again this process appears to be rare in yeast (Long et al. 2003b; Conant and Wagner 2005; Jones and Begun 2005; Wang et al. 2006). One example of a known chimeric gene in yeast is the STA1 glucoamylase gene, which exists in *S. diasticus* (often classified as a variety of *S. cerevisiae* rather than a separate species) but not in laboratory strains of *S. cerevisiae* (Yamashita et al. 1985; Yamashita, Nakamura, and Fukui 1987; Dranginis 1989). STA1 codes for an amylase enzyme that is secreted out of the cell. The gene was formed by duplication of an intracellular amylase gene, with one copy becoming fused to the N-terminal secretion signal of a flocculin gene. One of the main mechanisms of chimeric gene creation is exon shuffling which involves recombination between introns of different genes (Long et al. 2003a). This is probably rare in yeast due to the paucity of introns in its genes. Some other possible mechanisms for the generation of chimeric genes are illegitimate recombination between two ORFs, retrotransposition of a gene into another gene, or fusion of two adjacent genes (Jardine et al. 2002). Retrogenes in rice often recruit exons from regions flanking the insertion site, creating a large number of chimeric genes in this species (Wang et al. 2006). At least 15 other gene fusion candidates have been found in *S. cerevisiae* based on the comparison of domain structures of enzymes performing homologous functions between *S. cerevisiae* and *E. coli* (for example, the *FAS3* locus in yeast corresponds to the *accA*, *accB*, *accC*, and *accD* genes which are structured in an operon in *E. coli*), but these events are probably ancient and the exact mechanism of their formation is not known (Jardine et al. 2002).

1.6 Structural evolution of genomes

Gene order along chromosomes evolves through the physical rearrangement of chromosomal segments or genes. The movement of chromosomal segments can involve intrachromosomal rearrangements by inversion of sections of the chromosome, or interchromosomal rearrangements (by either reciprocal translocation (RT) or non-

reciprocal translocation. Single gene rearrangement events are transpositions of genes and can be both intra- and interchromosomal) (Figure 1.6).

Inversion rearrangement events are rearrangements whereby a section of chromosome between two breakpoints is reversed in orientation. Reciprocal translocations on the other hand involve the swapping of chromosomal segments between two breakpoints located on different chromosomes. Reciprocal translocations are only permissible if they produce daughter chromosomes that each retains a centromere. A classic example of a reciprocal translocation among yeast species was described by (Ryu, Murooka, and Kaneko 1998) who showed that homologous recombination between duplicated RPL2 ribosomal protein genes on chromosomes II and IV resulted in a RT between *S. cerevisiae* and *S. bayanus*. Further RTs among *sensu stricto* species and even lab strains of *S. cerevisiae* have also been documented (Casaregola et al. 1998; Fischer et al. 2000; Kellis et al. 2003).

Rearrangement events are caused by recombination between two chromosomal sites, which may be facilitated by sequence similarity. Highly similar regions of DNA such as transposable elements (Ty elements in *S. cerevisiae*), ribosomal proteins, tRNAs or duplicated regions may play a key role in these sorts of rearrangement. It has been shown that recombination between two nearly identical tRNA genes increases when both of them are actively transcribed (Pratt-Hyatt et al. 2006). Rearrangement between chromosomes in *S. cerevisiae* is found to increase with decreased levels of the replicative DNA polymerase (Pol α), and under these conditions the recombination is found to occur between homologous regions of Ty elements in *S. cerevisiae* (Lemoine et al. 2005). A recent comparative analysis between the complete genomes of the laboratory strain (S288C) and a clinical isolate (YJM789) of *S. cerevisiae* showed that the only genomic

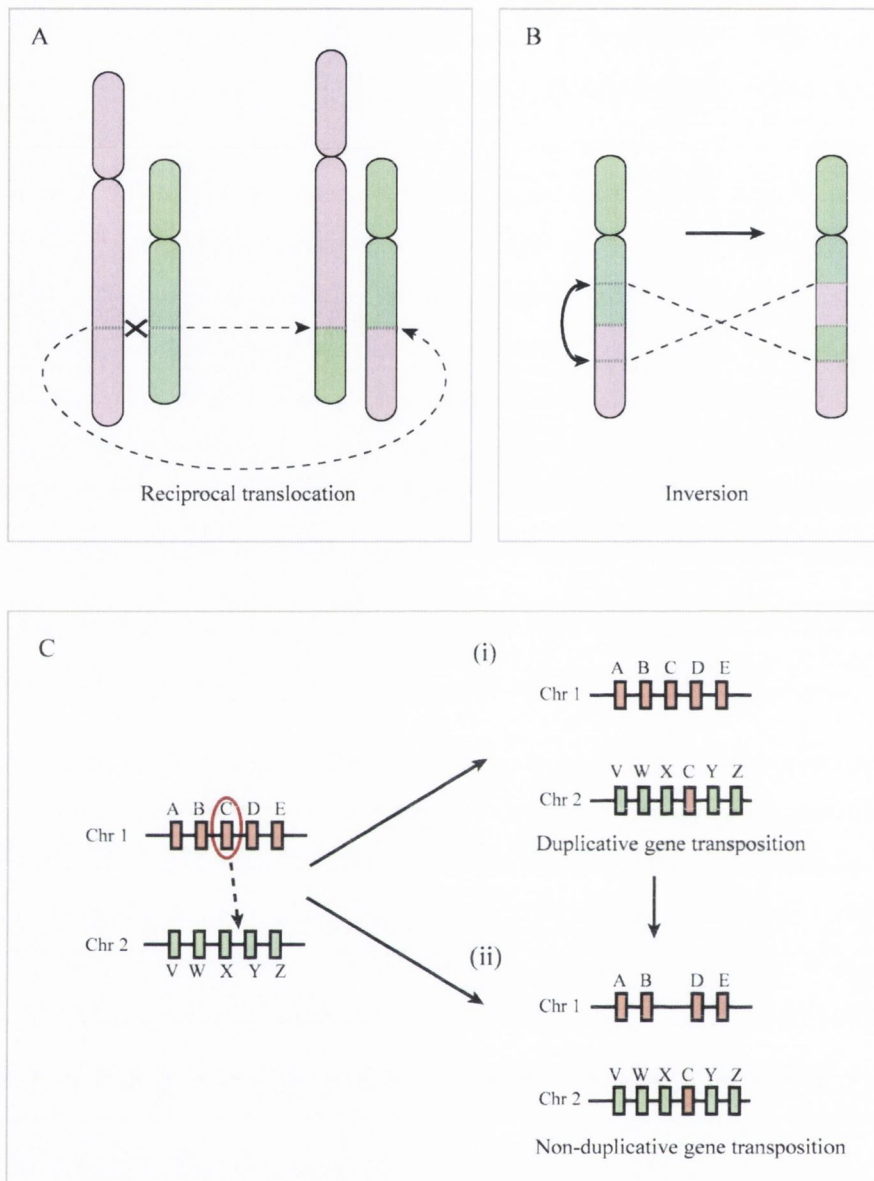


Figure 1.6 Schematic diagrams of different types of physical genome rearrangement. A) Reciprocal translocations involve the reciprocal swapping of the chromosome arms between two chromosomes at breakpoints on each chromosome. B) Inversions involve the reciprocal swapping of the ends of chromosomal segments between breakpoints (at the arrows) found on the same chromosome. This results in the reversal of the DNA sequence bordered by the breakpoints. C) Single gene movement in a genome generally proceeds by transposition. There are two outcomes for transposed genes: both the progenitor gene and the transposed gene can be maintained (shown in (i)), or the progenitor gene is lost (shown in (ii)). It is unsure whether non-duplicative transposition can proceed without duplicative transposition first occurring.

rearrangements were an inversion and a reciprocal translocation, both of which must have been recent events (Wei et al. 2007). The inversion breakpoints were flanked by inverted repeat sequences, which suggests that the inversion was formed by recombination between these repeats. Comparisons of synteny between *S. cerevisiae* and *A. gossypii* show that many synteny breakpoints formed by rearrangements contain tRNA genes, which may have acted as substrates for homology induced recombination (Dietrich et al. 2004).

Non-reciprocal chromosomal translocations (also known as Robertsonian translocations) differ from inversions and reciprocal translocations because they only involve one breakpoint. The rearrangement involves a segment of chromosome breaking off and rejoining to the telomeric end of another chromosome, with the breakpoint becoming the site of a new telomere on the original chromosome.

Gene location rearrangements occur by a transposition, a process which involves the movement of a gene from one genomic location to another regardless of mechanism. Some cases of transposition are known to involve transposable elements or certain proteins which they encode (Derr, Strathern, and Garfinkel 1991; Derr and Strathern 1993). It is possible that this genic relocation may be the same process as that of duplicative retrotransposition except with the loss of the original gene. The movement of a gene in such a manner will have the same consequences on its expression profile as for duplicative retrotransposition (discussed above).

While these types of chromosomal and genic rearrangement events have been common throughout the evolution of eukaryotes, it is unclear what roles they may have played in the evolution of species. It has been proposed that they present a neutral mechanism for the generation of species barriers (Lynch and Force 2000b). Gene relocation may also have allowed the physical clustering of certain groups of genes to evolve within genomes (Wong and Wolfe 2005).

1.6.1 Models of chromosome rearrangement

There are two major competing models to describe the distribution of rearrangements in genomes. These are the *random breakage* model of chromosome evolution, and the *fragile breakage* model of chromosome evolution (Ohno 1973; Pevzner and Tesler 2003). The random breakage model was proposed first, by Susumu Ohno (Ohno 1973). This model implies that breakpoints occur at random sites in the genome, so breakage is unlikely to occur at the same point twice. In the decades since it was proposed, this model has been supported by different genomic studies with increasing levels of resolution (Nadeau and Taylor 1984; Copeland et al. 1993; DeBry and Seldin 1996; Burt et al. 1999; Ma et al. 2006). The fragile breakage model was proposed by Pevzner and Tesler in 2003 (Pevzner and Tesler 2003).

The fragile model proposes that there are regions in the genome where breakpoints occur more frequently than expected by chance. The existence of fragile regions would also increase the frequency of breakpoint reuse, which occurs at low levels in the random breakage model. An increasing amount of genomic data supports the fragile model in organisms including mammals (Murphy, Stanyon, and O'Brien 2001; Larkin et al. 2003; Zhao et al. 2004; Lemoine et al. 2005). This would agree with observations of recombination between homologous sequence such as tRNAs and transposable elements mediating rearrangements in yeasts. It is likely that a lack of resolution in previous genomic studies underlies much of the debate between the two models of chromosome rearrangement. In particular, it is arguable whether multiple breakpoints that lie physically very close to one another in the genome but do not coincide, constitute a single "fragile site" or not (Lee et al. 2007). The increasing resolution and phylogenetic depth of genomic data will likely end this debate in the near future.

1.7 Synteny conservation and gene clustering

Hemiascomycete yeasts exhibit a high degree of gene order conservation (synteny) when compared to groups of species with similar molecular diversity such as vertebrates. For example, *S. bayanus* and *S. cerevisiae* have a molecular divergence comparable to that

found between human and mouse, yet there are only five reciprocal translocations and three inversions between the yeast species compared to at least 245 rearrangements between human and mouse (Kellis et al. 2003; Pevzner and Tesler 2003). It is likely that both genome size and the frequency of repeat elements impact on the rate of rearrangements as there is a higher chance of recombination between repeats in a repeat rich genome, or between homologous regions which could arise more frequently by chance in a large genome. The human and mouse genomes contain at least 2.5Gb of DNA and are repeat rich (40%), while the yeast species have much smaller genomes (~14Mb) which contain low numbers of repeats (<1%). Most of the changes to gene adjacency in yeasts occurred by gene deletion (especially in the post-WGD species) rather than by chromosomal rearrangement (Keogh, Seoighe, and Wolfe 1998).

The low rates of rearrangement and high rates of gene deletion found in the hemiascomycetes may be indicative of selection acting to maintain certain clusters of genes. The effects of reciprocal translocations on the fitness of *S. cerevisiae* has been demonstrated by (Colson, Delneri, and Oliver 2004). They created different translocations in an isogenic strain and competed them in nutrient limiting chemostat conditions to gauge the fitness differences between the different translocation strains. The derived strains exhibited differential growth under a variety of conditions, indicating that rearrangements of gene order can affect the fitness of an organism. If certain gene orders are preferential for growth under certain environmental conditions, then selection may act on the gene order of a genome to maintain the adjacency of some genes into clusters. These clusters will tend to exhibit low recombination rates as breaking them apart by rearrangement would decrease the fitness of the organism.

The clustering of certain genes in a genome is extremely common in bacteria, and many genes exist in operons which contain genes involved in a similar function and are often co-regulated by the same promoter sequences. By contrast functional clusters of genes in eukaryotes appear to be generally rare, although they are common in some species. For instance, in *C. elegans* around 15% of genes are found in bacteria-like operons which are coregulated and cotranscribed (Lercher, Blumenthal, and Hurst 2003). One of the most

striking examples of a eukaryotic metabolic gene cluster is the Aflatoxin cluster found in some *Aspergillus* species, which contains 25 genes involved in the biochemical pathway of the production of aflatoxin, a highly carcinogenic mycotoxin (Yu et al. 2004).

In *S. cerevisiae* large clusters of functionally related genes are rare. The best example is the clustering of six genes involved in the usage of allantoin as an alternative nitrogen source are located in a cluster in a subtelomeric region in *S. cerevisiae* (Wong and Wolfe 2005). This cluster has been assembled recently in the history of *S. cerevisiae* and is located in a region of low recombination despite being in a subtelomeric location. While large-scale clustering of functional genes in yeast may be rare, there is evidence of non-random gene order and clustering of genes. Adjacent genes in the *S. cerevisiae* genome are co-expressed far more often than expected by chance and often belong to the same functional category (Cohen et al. 2000; Hurst, Williams, and Pal 2002). Yeast genes that function in the same metabolic pathway or complex are generally located more closely than expected by chance although they may not be adjacent (Lee and Sonnhammer 2003; Teichmann and Veitia 2004). Essential genes are also clustered in *S. cerevisiae* in a manner independent of the clustering of co-expressed genes, and tend to exist in regions that exhibit low rates of recombination (Pal and Hurst 2003). Recent studies have shown that gene order in yeast may be influenced by selection for the clustering of genes into chromosomal domains with different chromatin formations (Batada and Hurst 2007; Batada, Urrutia, and Hurst 2007). These domains have different nucleosome occupation which is associated with different levels of transcriptional noise. Transcriptional noise can be highly detrimental for certain genes (such as dosage sensitive essential genes) while having little effect on others. The WGD in yeast may have been pivotal for the formation of these transcriptional domains and the reduction of transcriptional interference between genes through the process of duplicate deletion. Duplication followed by deletion of many genes may have allowed the separation of genes with similar transcriptional noise sensitivity into clusters on the different sister regions, which would decrease the transcriptional interference in the genome. In addition, the WGD followed by deletions increased the average intergenic distance between adjacent genes, and reduced the bias of divergently/convergently transcribed genes seen in pre-WGD species (Byrnes, Morris, and

Li 2006). These side-effects may also have reduced the transcriptional interference between neighbouring genes.

1.8 Ancestral gene order reconstruction

The reconstruction of ancestral gene orders from the comparisons of whole genome sequences of extant species can provide information about the physical evolution of genome structure and content for a known phylogeny. Once an ancestral gene order has been inferred at a node in a known phylogeny, it becomes possible to track evolutionary events such as chromosomal rearrangements, transpositions, duplications, gene deletions and gene gains in the lineages derived from this node (Chapter 2 and Chapter 3). By examining these events it may be possible to obtain information about their mechanisms and evolutionary dynamics. It may also be possible to link events like gene gains and losses to functional differences between different species. This information combined with functional data from extant species could potentially provide an insight into the past ecologies and phenotypes of lineages, and the evolutionary pressures that may have been exerted upon them.

In cases where the branching of a phylogeny is uncertain, but a reconstruction of an ancestral node has been successful, it may be possible to obtain the true topology of the tree by tracing the evolutionary events along the lineages leading to the extant species. If any two lineages share one or more rearrangement events with the exclusion of another species that maintains an ancestral state, then it can be assumed that the species with the shared event clade together in the phylogeny with the ancestral-type species as an outgroup. In this instance, the event likely occurred on a shared branch of two of the species after the outgroup lineage had speciated. In the case of the hemiascomycete phylogeny, most sequence based phylogenetic trees clade *S. cerevisiae* and *S. castellii* with *C. glabrata* as an outgroup. However, there are numerous cases of shared events in *C. glabrata* and *S. cerevisiae* where *S. castellii* maintains the ancestral state, and no discovered cases of *S. cerevisiae* and *S. castellii* sharing events absent from *C. glabrata*. This indicates that *S. cerevisiae* and *C. glabrata* are more closely related than previously indicated by sequence based phylogenies (Chapter 2; Scannell *et al.* 2006, Supplementary Information).

The general method that has previously been used to infer ancestral gene orders is to use a parsimonious approach. These studies have generally centred around the anthropocentric ancestral reconstructions of vertebrates, eutherian mammals, placental mammals, primates, carnivorous mammals although ancestral reconstructions have also been elucidated for rodentia, mitochondria and chloroplasts (Chowdhary et al. 1998; Blanchette, Kunisawa, and Sankoff 1999; Muller et al. 1999; Murphy, Stanyon, and O'Brien 2001; Bowers et al. 2003; Larkin et al. 2003; Richard, Lombard, and Dutrillaux 2003; Yang et al. 2003; Jaillon et al. 2004; Zhao et al. 2004; Froenicke 2005; Murphy et al. 2005; Kohn et al. 2006). Knowledge of the phylogeny of the species being compared is required and more than two species are required in order to allow the assignment of rearrangements to specific branches of the tree. The basic concept of parsimonious ancestral reconstruction is that if two species share the same gene order of orthologous genes, then they inherited that arrangement from their common ancestor and it was uninterrupted over time. With time, the passage of chromosomal rearrangements will tend to disrupt the gene order relationships shared by related species. It is generally assumed that each rearrangement is a unique event. Even allowing for breakpoint reuse each chromosomal rearrangement creates two breakpoints in the genome, and it is unlikely that both breakpoints will be used in the same rearrangement independently in different species.

By comparing a rearranged species with one that diverged before a rearrangement event it is possible to deduce the nature of the rearrangement. In the case of a reciprocal translocation, the region around each breakpoint formed by the RT will be represented by two different genomic regions in each species relative to the other. For inversions there will be a reversal of gene order between two points on the same chromosome between the species. Pairwise comparisons like these are non-directional, so in both of these cases a third ancestral ordered outgroup species is required to pinpoint the rearrangement to one of the species. If all rearrangements between species can be assigned to specific branches on the phylogeny, it is possible to trace them back to the ancestral orders.

The ability to identify orthologous and paralogous regions between and within species is of high importance for ancestral reconstruction, as is the phylogenetic depth and range of the

species used in the analysis. To identify orthologous regions of genes between species, a combination of sequence similarity and synteny conservation of groups of genes are generally used. If the species are separated by too great a phylogenetic depth, then signals of orthology like sequence similarity may be obscured and multiple rearrangements along both branches may reduce synteny conservation to a low level making it difficult to determine true orthology between the genomes. Very closely related species will not make a large contribution towards inferring a deeper ancestral order because both species will likely differ little from their own common ancestor and will only reveal species specific differences between them.

Initial studies to reconstruct the ancestral gene order in vertebrate species used karyotypic comparisons between species with a known phylogeny to piece together ancestral karyotypes. Initially, many of the genomes compared had large genomes and little or no sequence data, which is why physical mapping, cytogenetic mapping, genetic linkage, chromosome banding and hybridisation methods such as ZOO-FISH chromosome painting were generally used for reconstruction purposes (Chowdhary et al. 1998; Blanchette, Kunisawa, and Sankoff 1999; Muller et al. 1999; Murphy, Stanyon, and O'Brien 2001; Bowers et al. 2003; Larkin et al. 2003; Richard, Lombard, and Dutrillaux 2003; Yang et al. 2003; Jaillon et al. 2004; Zhao et al. 2004; Froenicke 2005; Murphy et al. 2005; Kohn et al. 2006). These methods have a relatively low resolution with mean breakpoint sizes of over 1 Mb and are unable to identify intrachromosomal rearrangements between species (Froenicke et al. 2006; Ma et al. 2006; Robinson, Ruiz-Herrera, and Froenicke 2006). They are also not suitable for long range comparisons between distantly related species as they rely on DNA-DNA hybridisation. These methods do have the advantage of being able to obtain data easily from multiple species which allows an increase in resolution and data quality (Froenicke 2005). As genome sequences become more abundant, *in silico* approaches are being used alone or in conjunction with chromosome painting methods to reconstruct ancestral gene orders, allowing a higher resolution picture of orthology between organisms and the quality of ancestral reconstructions (Murphy et al. 2005; Bourque, Tesler, and Pevzner 2006; Kohn et al. 2006).

Most of the reconstructions of ancestral gene order in eukaryotic organisms have been focused on ancestral vertebrates and mammals. Recently Sankoff *et al.* applied rearrangement median and genome halving algorithms to the genome of the yeast *S. cerevisiae* to acquire a putative pre-polyploid gene order (Sankoff, Zheng, and Zhu 2007). A comparison between this order and my parsimony-based reconstruction of the pre-polyploid yeast ancestor will be discussed in Chapter 2.

Genomic Evolution of Polyploid Yeast Species

Chapter 2.

Genes added, genes lost, genomes rearranged: The evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome

2.1 Abstract

Comparative genomics among multiple species can be used to infer the history of genomic rearrangements that occurred during their evolution. Here we aligned synteny blocks from eight yeast species to infer the most parsimonious gene content and gene order that existed in the genome of an extinct ancestral yeast about 100 Mya, immediately before it underwent whole-genome duplication (WGD). The reconstructed ancestral genome contains 4703 loci in order on eight chromosomes. We then inferred the series of rearrangement steps that led from this ancestor to the current *Saccharomyces cerevisiae* genome: relative to the ancestral genome we observe 66 reciprocal translocations, 73 inversions, 5 Robertsonian translocations, 9 single-gene transpositions and 2 multigene transpositions. We identified 153 genes that have been gained by *S. cerevisiae* in the time since the WGD, and 144 loci at which *S. cerevisiae* did not retain either of the gene copies that were formed by WGD. Analysis of the reciprocal translocations and inversions required to transform the ancestral gene order into that of *S. cerevisiae* reveals fragile chromosomal sites that were reused as breakpoints multiple times. We find that the insertion sites of recently gained genes tend to be associated with tRNA genes and origins of replication. Many of the gained genes in *S. cerevisiae* have functions in anaerobiosis, cell wall morphology, or the metabolism and transport of alternative nutrient sources.

2.2 Introduction

Inferring the genome organization and gene content of an extinct species has the potential to provide detailed information about the recent evolution of species descended from it. If we know what was present in the genome of an ancestor, we can deduce how a current-day descendant differs from it and we can ask questions about why it differs. The most recent changes in a genome are often the most interesting ones, because they reflect the most recent (or even current) evolutionary pressures acting on that genome.

Yeast species offer a promising target for ancestral genome reconstruction because many genomes have been sequenced and they show extensive colinearity of gene order among species (Keogh, Seoighe, and Wolfe 1998; Llorente et al. 2000; Dietrich et al. 2004; Byrne and Wolfe 2005). As the number of sequenced genomes from related species rises, so too does the precision with which we can reconstruct their history. In this study we compare the genomes of a group of species in the subphylum Saccharomycotina, spanning an evolutionary time-depth that is comparable to that of the vertebrates (Dujon 2006). A whole-genome duplication (WGD) event occurred during the evolution of this subphylum (Wolfe and Shields 1997), and we can compare the genomes of several species (including *Saccharomycetes cerevisiae*) that are descended from this event to the genome sequences of several species that branched off before the WGD occurred. In this study we focus on an ancestor that existed approximately 100 Mya, at the point immediately before the WGD occurred. The evolutionary route from this ancestor to *S. cerevisiae* is of particular interest from both genome evolution and organismal evolution standpoints, because it corresponds to a period of time during which the *S. cerevisiae* lineage became increasingly adapted to rapid fermentative growth (Thomson et al. 2005; Piskur et al. 2006), and during which extensive rearrangement of the genome occurred (including the deletion of thousands of redundant copies of duplicated genes; (Scannell et al. 2007).

Inferring the set of genes that existed in the immediate pre-WGD ancestor, and the ancestral gene *order* of those genes enables us to trace all of the inter- and intra-chromosomal rearrangements that occurred on the route from this ancestor to the current

S. cerevisiae genome, which is informative about the molecular mechanisms of evolutionary genome rearrangement. It also allows us to infer the phylogenetic relationships among the post-WGD species based on the presence of some shared chromosomal rearrangements among species, which provides a phylogenetic signal that is independent of the signal from DNA or protein sequences.

Knowing the gene *content* of the ancestral genome allows us to identify genes that have been added to, or lost from, the *S. cerevisiae* genome during the past 100 Myr. Changes in gene content can provide a strong indication of the changing evolutionary circumstances of a species. These changes can be either passive or adaptive. Genes for unnecessary functions will not be maintained. There will be no purifying selection against mutations in them, so they will be lost passively from the genome. Previous studies have identified passive losses of some genes whose functions became unnecessary in particular yeast species. For example *Candida glabrata* appears to have switched from a free-living to a pathogenic lifestyle relatively recently and has lost whole sets of genes coding for the biochemical pathways for catabolism of galactose (*GAL* genes) (Hittinger, Rokas, and Carroll 2004) and allantoin (*DAL* genes) (Dujon et al. 2004), presumably because it never encounters these molecules in its new environment on mammalian hosts.

Conversely, gene gains can be indicative of the adaptive evolution of a species. In this study we identify gene gains in *S. cerevisiae* based on their appearance as insertions relative to the syntenic region in the inferred ancestral genome, so our set of gained genes potentially includes any gained by horizontal gene transfer (Hall, Brachat, and Dietrich 2005) or even *de novo* invention (Levine et al. 2006) as well as duplication. Only a tiny fraction of adaptive mutations are likely to be gene gains, but it is possible that many of the events of gene gain were adaptive in yeast species. A dramatic example is the gain (by duplication) of the *ADH2* gene in *S. cerevisiae*, which enabled this species to respire ethanol efficiently instead of leaving it to evaporate (Thomson et al. 2005). A second example is the gain (by horizontal transfer) of a bacterial *URA1* gene for anaerobic uracil biosynthesis in an ancestor of the *Saccharomyces-Kluyveromyces* clade; the gain of this gene resulted in the subsequent loss of its pre-existing eukaryotic paralogue *URA9* from the

S. cerevisiae genome (Gojkovic et al. 2004; Hall, Brachat, and Dietrich 2005). Not all gene gains are expected to be adaptive, because new genes gained by duplication can be maintained by the non-adaptive process of subfunctionalization (Force et al. 1999). The extent to which newly-formed duplicate genes are preserved in yeast genomes by subfunctionalization as opposed to adaptive mechanisms (such as selection for increased dosage) is currently unclear: several examples of subfunctionalised gene pairs have been reported (Hittinger, Rokas, and Carroll 2004; van Hoof 2005) even though the large population sizes of yeast species should make subfunctionalization improbable (Lynch and Force 2000a). Even though it may not be possible to conclude that any particular gene gain was adaptive, the clear links between the functions of the gained genes *ADH2* and *URA1* and the adaptation of *S. cerevisiae* to a fermentative lifestyle suggested to us that a systematic search for all the genes that were gained by this genome since WGD would be worthwhile.

2.3 Results and Discussion

2.3.1 Ancestral Genome Reconstruction

We reconstructed the gene order and gene content of the yeast ancestor that existed immediately prior to WGD, using a manual parsimony approach (Figure 2.1). We used the Yeast Gene Order Browser (YGOB; Byrne and Wolfe 2005) to browse systematically through alignments of local gene orders from four post-WGD species (*S. cerevisiae*, *C. glabrata*, *S. castellii* and *K. polysporus*) and three pre-WGD species (*A. gossypii*, *K. lactis* and *K. waltii*).

The gene content of the ancestor was inferred as shown in Figure 2.1B. Any gene that is present at a conserved location in at least one pre-WGD species and one post-WGD species is inferred to have been present in the ancestral genome. Any locus that was retained in two copies in sister regions of a post-WGD species is also inferred to have been

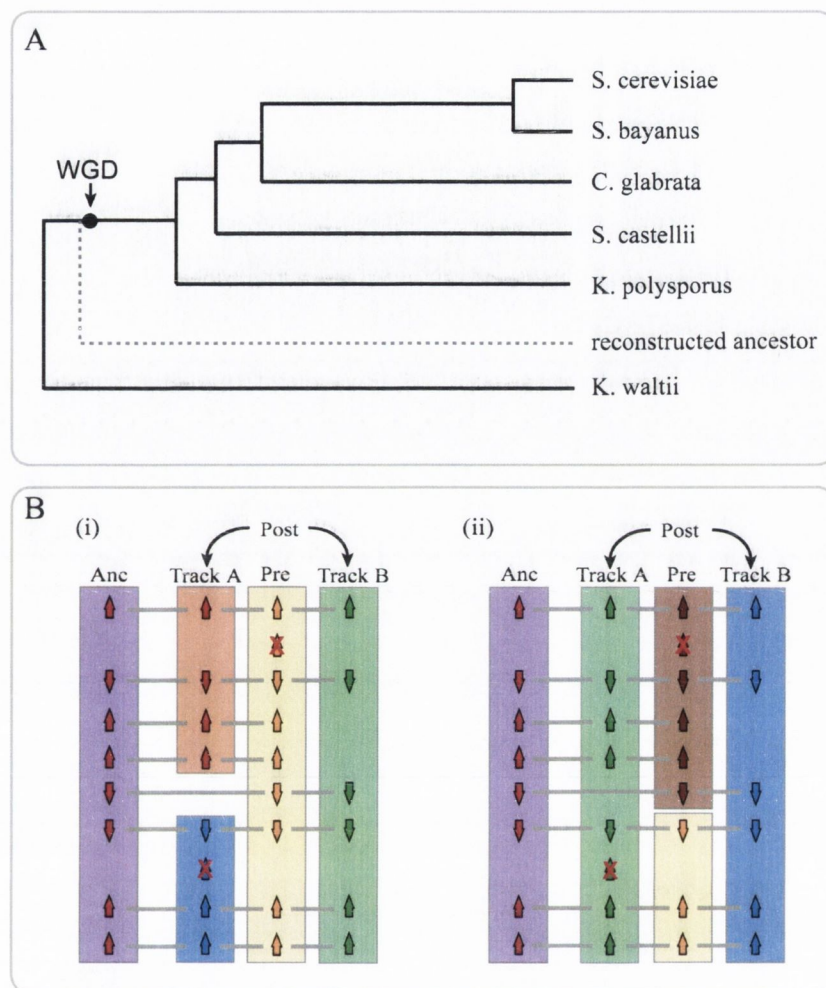


Figure 2.1 A) Relationship between the post-WGD yeast species in this study relative to an outgroup pre-WGD species (in this case *K. waltii*). A whole genome duplication (black dot) occurred in the common lineage of *S. cerevisiae*, *S. bayanus*, *C. glabrata*, *S. castellii* and *K. polysporus*. The location of the reconstructed ancestral gene order is signified by the dashed grey line joining the tree immediately prior to the whole genome duplication event. This tree is not drawn to scale. B) Reconstruction of gene orders in a common ancestor by parsimonious comparison of the gene orders in extant species. There are two types of scenario for crossing chromosomal breakpoints in our data. (i) The conservation of gene order between a pre-WGD species and one of the post-WGD tracks with a breakpoint occurring in the other post-WGD track. The inferred ancestral track is at the left of the picture. (ii) A breakpoint occurring in a pre-WGD track with syntenic continuity across the broken region in both of the post-WGD tracks. Ancestral gene content is inferred for each locus where a gene is orthologous between a pre-WGD species and at least one post-WGD track or is paralogous between post-WGD sister tracks. Genes only found in one post-WGD track or one pre-WGD track with no post-WGD support cannot be ancestrally traced and are marked with red crosses in the diagram.

present in the ancestral genome, even if it is not present at the syntenic location in the pre-WGD genomes considered. Similarly, because YGOB identifies orthology and paralogy relationships among the two sister genomic tracks from each post-WGD species, we can infer that a locus was present in the ancestor if it was retained on one sister region (say Track A) in one post-WGD species and on the paralogous region (Track B) in another post-WGD species, even if it was not retained in duplicate in either species.

The gene order of the ancestor was also inferred by parsimony. After establishing the content of the ancestral genome, we examined the adjacency relationships among these genes. Within any post-WGD species such as *S. cerevisiae*, most of the genome can be sorted into pairs of sister regions that have a double-conserved synteny (DCS) relationship with any pre-WGD species such as *K. waltii* (Dietrich et al. 2004; Kellis, Birren, and Lander 2004). Breaks in the DCS pattern correspond to two types of event: (i) a chromosomal rearrangement that occurred in the post-WGD species causes one of the two tracks from this species to have a discontinuity relative to the pre-WGD species and the other track from the post-WGD species, and (ii) a chromosomal rearrangement that occurred in the pre-WGD species after their divergence from the lineage leading to the post-WGD species results in discontinuities on both tracks from the post-WGD species relative to the rearranged region in the pre-WGD species. These two types of event were called single-breaks and double-breaks of synteny, respectively, by Dietrich *et al.* (Dietrich et al. 2004). For each single-break of synteny, because there are genome sequences from multiple post-WGD species, and because the endpoints of the chromosomal rearrangements in each species generally do not coincide in all species, it is possible to infer the species and chromosomal track on which the break happened. This inference also tells us the ancestral gene order across the site of breakage: in general the ancestral order has been disrupted in one track in one post-WGD species, but it is still conserved in the same track from the other post-WGD species, in the sister track from all the post-WGD species, and in the pre-WGD species. Similarly for each double-break of synteny, because there are multiple genome sequences from pre-WGD species we can in general identify the break as having occurred in a particular pre-WGD species. A small number of double-breaks of synteny are caused by situations where all the pre-WGD species show a different gene order compared to both

of the post-WGD species tracks, which share the same order. These breaks correspond to rearrangements that occurred either on the branch between the root of the tree and the common ancestor of the pre-WGD species, or on the branch between the root and the common ancestor of the post-WGD species (before the WGD happened), but we cannot tell which. We do not include these breaks in our analysis because we are only interested in events that occurred after the WGD, and in these cases we infer the pre-WGD ancestral order from the two post-WGD species tracks.

Reconstruction by this method resulted in an inferred ancestral genome with eight chromosomes, containing 4703 genes. The content of this genome can be viewed using YGOB (<http://wolfe.gen.tcd.ie/ygob>) and is listed in Table S2.1 (on the CD provided). The gene set represents the intersections of orthologous genes between pre-WGD and post-WGD species and between ohnologues (paralogues formed by WGD) across the post-WGD species. The ancestral gene set accounts for 5107 genes currently present in *S. cerevisiae* (1102 ohnologues and 4005 single copy genes), which covers all genomic regions in *S. cerevisiae* except for variable subtelomeric regions (discussed below). The *S. cerevisiae* genome can be mapped onto the inferred ancestral genome in 244 synteny blocks (Figure 2.2, Table S2.2 on the CD provided), that "tile" together in an unambiguous 2:1 fashion across the ancestral genome (Figure 2.2). The other post-WGD species can be mapped similarly onto the ancestral genome: there are an estimated 594 blocks for *C. glabrata* and 631 for *K. polysporus* (the *S. castellii* genome annotation is highly fragmented into lots of contigs, so a realistic synteny block number estimation is not possible). The pre-WGD genomes map onto the inferred ancestral pre-WGD genome in 365 (*A. gossypii*), 427 (*K. lactis*) and 225 (*K. waltii*) blocks, respectively. Note that these numbers are dependent on the degree of fragmentation of the genome assembly of these species, and are probably overestimated in some cases.

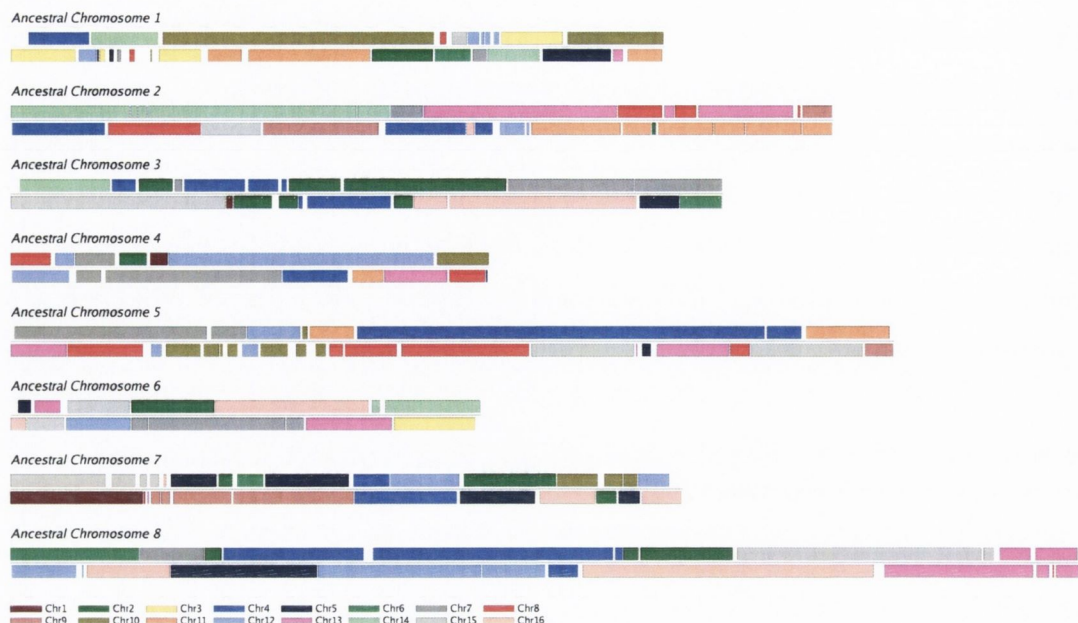


Figure 2.2 Synteny block map of the pre-WGD species ancestral gene order in terms of *S. cerevisiae* gene order. Each coloured block represents a region in *S. cerevisiae* that is collinear with a region from the ancestral order. There are a total of 244 synteny blocks in *S. cerevisiae* which map onto the ancestor in an unambiguous 2:1 pattern. The colour corresponding to each current *S. cerevisiae* chromosome is shown at the bottom. Figure generated by Kevin Byrne.

Our inferred ancestral gene set is incomplete in some respects. First, the gene content in subtelomeric regions is different in every species and the dynamic nature of these regions makes it impossible to trace evolutionary events over long timescales. The reconstructed gene order and content covers only the internal regions of chromosomes, extending out to the approximate borders of the subtelomeric regions where synteny conservation stops. In *S. cerevisiae*, for example, a total of 295 genes are located in subtelomeric regions extending beyond the ends of the ancestral reconstruction. These genes make up the majority of the *S. cerevisiae* genes that do not have counterparts in the ancestral genome. Second, we cannot detect genes that may have been present in the pre-WGD ancestor but were subsequently lost, in both copies, by all the post-WGD species considered. Third, there is some ambiguity concerning the timing of inferred gene gains. For example, the *HO* endonuclease gene is not present in the pre-WGD genomes but is present as a single-copy orthologous gene in all four post-WGD genomes. Considering only these species,

this gain could have occurred in the pre-WGD ancestor or it could have occurred in the time interval after WGD but before the divergence of the four post-WGD species. In fact, we know from additional information that *HO* was gained prior to WGD, because it is present at a syntenic location in another pre-WGD species, *Zygosaccharomyces rouxii* (Butler et al. 2004). In cases where we did not have any additional information of this kind, we assumed that the gene was present in the ancestor.

Sankoff and colleagues (Sankoff, Zheng, and Zhu 2007) recently computed a structure for an unduplicated ("diploid") ancestor of *S. cerevisiae*, by comparison to *A. gossypii* and by using a combination of two computer science algorithms (the rearrangement median and genome halving algorithms). The ancestral structure that they propose is substantially different to ours (Figure 2.3), but they did not search exhaustively for an optimal solution so it is likely that better solutions remained undiscovered in their study (D. Sankoff, personal communication). In any case, the results are not expected to be identical because their input data was of lower resolution than ours since they only considered two species. Their ancestral structure has nine chromosomes, so at least one of them has no centromere; they did not enforce a requirement that there should be one centromere per chromosome at all stages of the inferred rearrangement process.

2.3.2 Rearrangement route to the current *S. cerevisiae* genome

Using the breakpoints between *S. cerevisiae* syntenic blocks in the ancestral genome, we inferred the large scale chromosomal rearrangements that have occurred in the *S. cerevisiae* lineage since the WGD. Most rearrangement events could be classified as either inversions or reciprocal translocations. Inversions were defined as events where the two endpoints of the rearrangement were on the same ancestral chromosome. Note that if a region that has been inverted also contains an endpoint of a reciprocal translocation, it is sometimes impossible to tell the difference between this scenario and a scenario of two reciprocal translocations; thus our method probably misclassifies some inversions as reciprocal translocations.

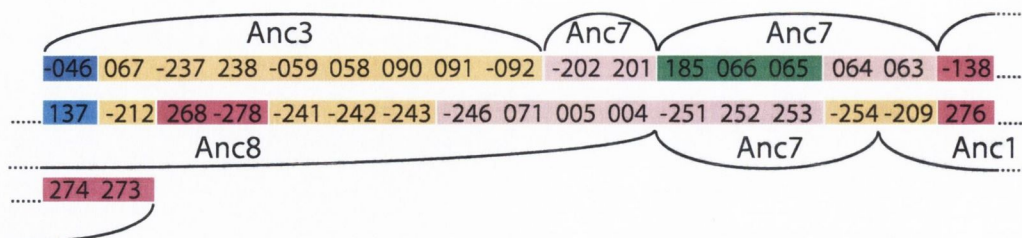


Figure 2.3 Comparison between an ancestral chromosome from Sankoff *et. al* (2007) and the parsimoniously generated ancestor from this study. The coloured blocks each represent a block of genes in *S. cerevisiae*, The brackets indicate different syntenic regions in the parsimonious ancestor covered by the blocks. Manually checking the syntenic boundaries between the ancestral reconstructions confirms that the parsimonious reconstruction is correct in these cases.

In total we inferred 73 inversion events and 66 reciprocal translocations on the evolutionary path from the ancestor to *S. cerevisiae*. We found that five of the inversions occurred at telomeres, and there were also five Robertsonian translocation events (exchanges between a telomere and an internal region of another chromosome; one of these events occurs at a shared inversion/translocation breakpoint). The data indicate that some intergenic regions were re-used as breakpoints in more than one rearrangement event. We classified the rearrangements as consisting of 34 simple inversion events (not overlapping other inversions or reusing breakpoints), 39 complex inversion events (overlapping other rearrangements and/or reusing breakpoints), 44 simple reciprocal translocation events, and 22 reciprocal translocation events involving breakpoint reuse (Figure 2.4).

If some post-WGD species share a rearrangement relative to the ancestor but others retain the ancestral gene order, the rearrangement event is a phylogenetically informative character. We searched for rearrangements shared by any pair of post-WGD species. As described below, we found many that support the branching order of the post-WGD species shown in Figure 2.5. We did not find any shared rearrangements supporting alternative

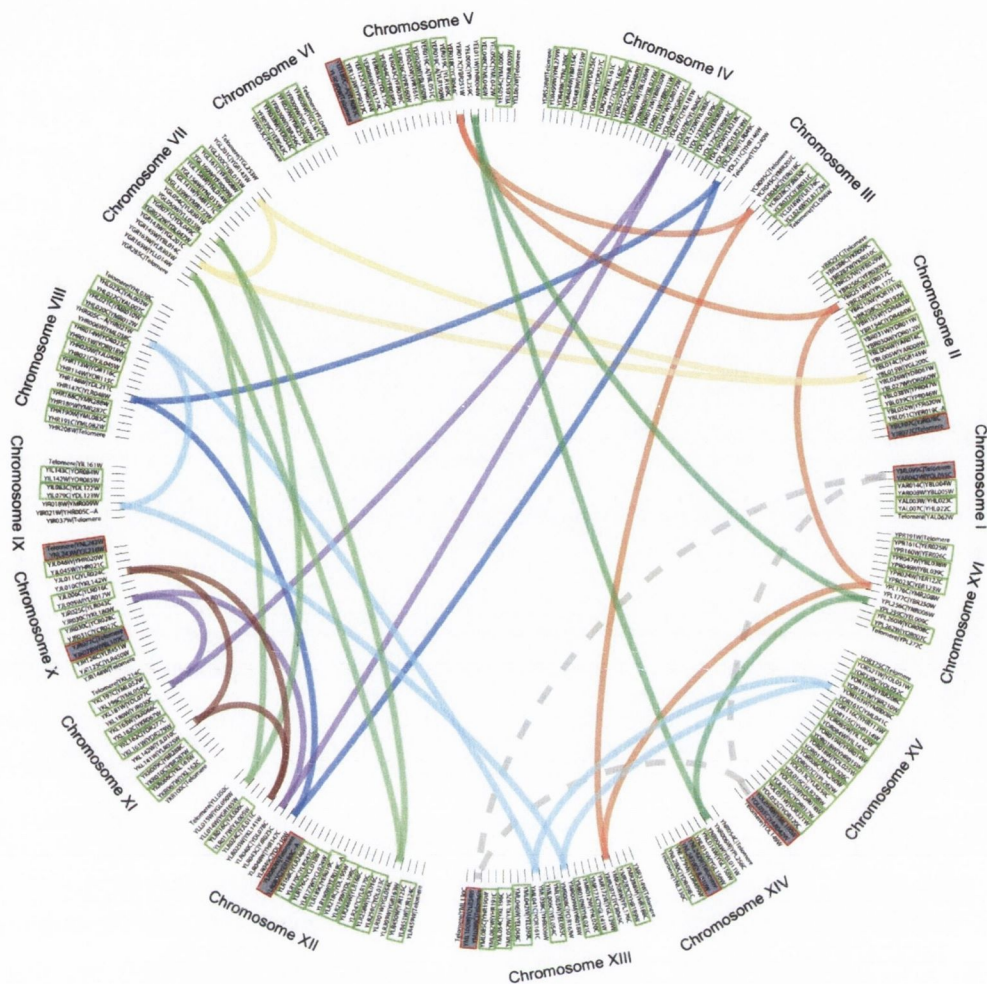


Figure 2.4 Breakpoints between *S. cerevisiae* synteny blocks relative to the ancestral genome and arranged by their order along each extant *S. cerevisiae* chromosome. Each pair of genes separated with a “|” represent adjacent genes in the ancestral genome that were separated by a reciprocal translocation in the *S. cerevisiae* lineage. Each breakpoint is drawn twice because each rearrangement created breakpoints at two different genomic locations in *S. cerevisiae*. A green box has been drawn around each pair of breakpoints that were formed by one reciprocal translocation event with no breakpoint reuse. In these cases, the genes bordering the same side of each breakpoint (as ordered) are adjacent in the current *S. cerevisiae* genome (e.g. YPR161CIYER025W (from ANC7B) and YPR160WIYER026C (from ANC3B). In cases of breakpoint reuse, the genes on one side of the pair of breakpoints are adjacent in *S. cerevisiae*, while the genes on the other side are not. By iteratively linking each of the non-matching genes to the gene that is adjacent to them in the *S. cerevisiae* genome, we can describe reciprocal translocation cycles (coloured lines) that link breakpoints which have been used more than once. For a cycle with n lines there are n breakpoints locations and $n-1$ reciprocal translocation events. The breakpoint reuse for each cycle is $n/(n-1)$. We observed 1 event where a Robertsonian translocation breakpoint was reused (dashed grey line). Telomeric Robertsonian translocation events are depicted as grey shaded red boxes. Figure generated in part by Gavin Conant. A high resolution version of this figure is available on the Supplementary CD, and is called Figure S2.1.

topologies. This result supports our previous conclusion, based on shared patterns of gene losses, that *S. castellii* is an outgroup to a clade containing *C. glabrata* and *S. cerevisiae* (Scannell et al. 2006; Scannell et al. 2007). In contrast, phylogenies based on sequence analysis tend to place *C. glabrata* outside *S. castellii* and *S. cerevisiae* (Kurtzman and Robnett 2003; Hittinger, Rokas, and Carroll 2004; Jeffroy et al. 2006), a result that we suspect is an artifact.

Given this phylogeny, the four post-WGD species define four temporal intervals for rearrangements: (i) No rearrangements are shared by all four post-WGD species on the branch between the pre-WGD ancestor and the point of speciation of *K. polysporus*; (ii) 8 rearrangement events are shared by *S. castellii*, *C. glabrata* and *S. cerevisiae* (5 inversions, 1 reciprocal translocation, 1 Robertsonian translocation, 1 gene transposition and 1 transposition of a block of neighboring genes); (iii) 21 rearrangements are shared only by *C. glabrata* and *S. cerevisiae*, with *S. castellii* and *K. polysporus* retaining the ancestral organization (14 inversions, 6 reciprocal translocations, 1 gene transposition and 4 gene block transpositions); and (iv) 132 rearrangements are unique to *S. cerevisiae* (54 inversions, 59 reciprocal translocations, 4 chromosomal translocations, 11 single-gene transpositions and 4 gene block transpositions). Most of the rearrangements that are specific to *S. cerevisiae* are temporally ambiguous relative to each other, except in some cases where they overlap in such a way that they could only have occurred in a particular order (data not shown).

The lack of rearrangements in the first interval is notable because it agrees with sequence comparisons and gene loss data indicating that *K. polysporus* separated from the other lineages soon (approximately 10 million years) after the WGD (Scannell et al. 2007). No rearrangements could be found that were shared by one genomic track of all the post-WGD species, relative to the other track and the ancestor, prior to this speciation. It also indicates that if the WGD event was an allopolyploidization rather than an autopolyploidization, then the two hybridizing genomes must have been completely colinear.

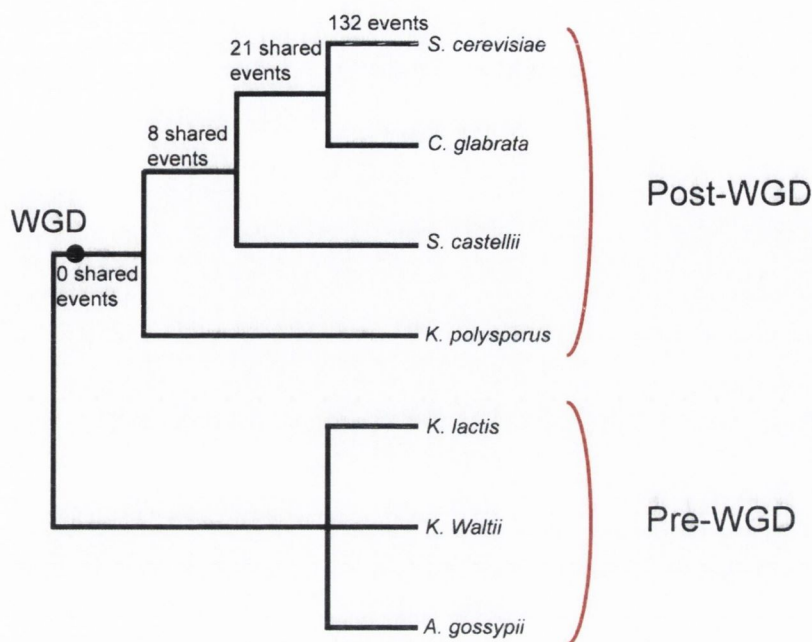


Figure 2.5 Shared genomic rearrangements since the time of the WGD support the clading of *S. cerevisiae* and *C. glabrata* with *S. castellii* as an outgroup in the post-duplication yeast species. The number of rearrangement events are shown along branches leading the *S. cerevisiae* from the WGD. Shared genomic rearrangements amongst species are phylogenetically informative. In the case of *S. cerevisiae*, *C. glabrata* and *S. castellii*, sequence based trees consistently give a topology with *C. glabrata* as an outgroup to *S. cerevisiae* and *S. castellii*. Our findings of shared rearrangements between *C. glabrata* and *S. cerevisiae* with *S. castellii* maintaining an ancestral state are very good evidence for a topology that differs from the sequence based trees, especially as no shared rearrangements between *S. cerevisiae* and *S. castellii* with the exclusion of *C. glabrata* were found.

2.3.3 Breakpoint re-use

In the manual reconstruction of extant *S. cerevisiae* chromosomes from the ancestral gene order, it was necessary to infer breakpoint reuse at the ends of some synteny blocks. Reused breakpoints appear as cycles in the map of breakpoint pairs (Figure 2.4). The re-use of breakpoints has previously been identified in studies on mammalian genome evolution.

Table 2.1 Breakpoint re-use in *S. cerevisiae*

Event type	Number of events	Number of breakpoints		Re-use ratio
		Expected	Observed	
Reciprocal translocations	66	132	118	1.12
Inversions	73	141*	126	1.16
Robertsonian translocations	5	5	5	1
All events	144	278	228†	1.22

* Five inversion events occur at telomeres, so each makes only one breakpoint.

† 21 breakpoints are shared by a reciprocal translocation and an inversion.

We find that for both reciprocal translocations and inversions, there are fewer breakpoints than expected if every event had unique ends (Table 2.1). The average number of breaks per used site is 1.12 for reciprocal translocations, and 1.16 for inversions. Some sites were used as endpoints of both an inversion and a reciprocal translocation, and if we pool these two categories there are only 228 unique breakpoints instead of the expected 278, implying an average of 1.22 breaks per site (Table 2.1).

2.3.4 Coincidence of sites of gene gain, evolutionary breakpoints, tRNA genes and origins of replication.

We identified 121 sites in the ancestral genome at which genes are inferred to have been gained subsequently in the lineage leading to *S. cerevisiae*, and found that 63 (52%) of these sites are beside tRNA genes. Furthermore, it is known that origins of replication in yeast are often located near tRNA genes (Wyrick et al. 2001), and it seems plausible that origins of replication might be fragile sites for breakage and/or integration of new DNA. To test the significance of the associations among these different types of chromosomal site, we used computer simulation. We mapped onto the ancestral genome the known locations of 246 *S. cerevisiae* tRNA genes (from SGD; Nash et al. 2007), 269 *S. cerevisiae* origins of replication (from OriDB; Nieduszynski et al. 2007), 228 evolutionary breakpoints on the *S. cerevisiae* lineage (Table S2.2 on CD), 121 sites of gene gain in *S. cerevisiae*, and 47

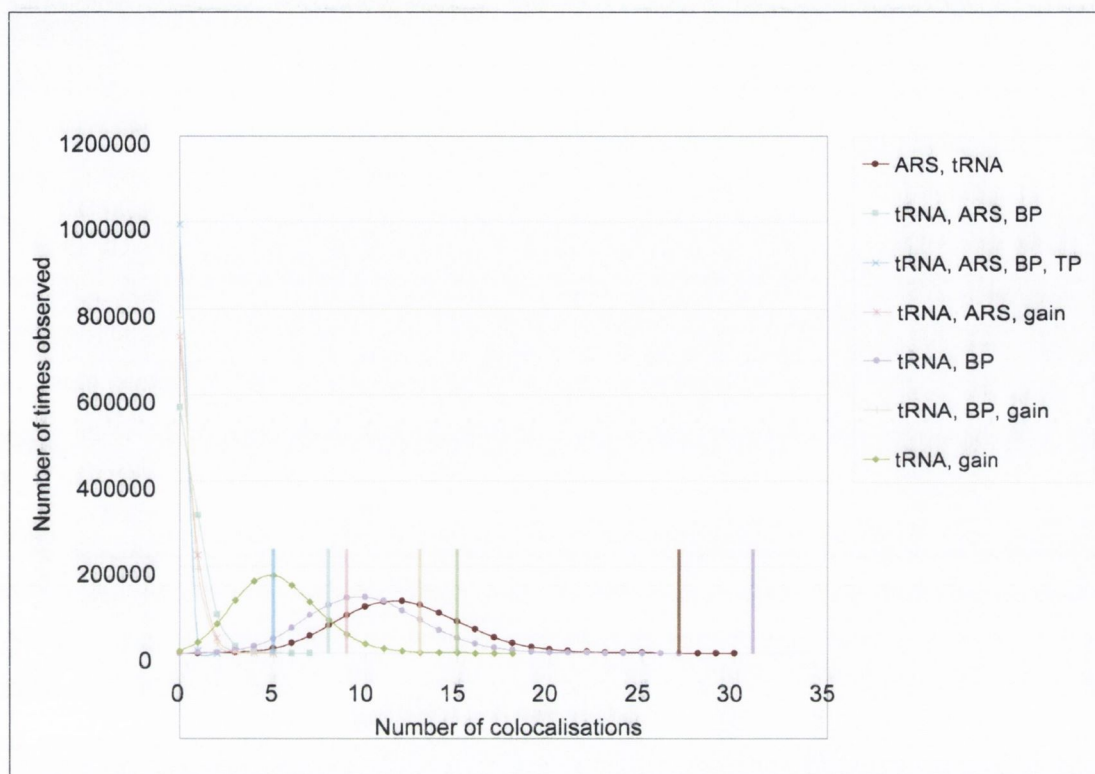


Figure 2.6 Significant results from the 1,000,000 simulations of coincidences of randomly located tRNAs, origins of replication (ARS), ancestral breakpoints (BP), gene gain locations and transpositions (TP) in a genome with 5100 intergenic regions. The vertical coloured lines are observed values in our data, and the curves are the expected distributions under the assumption of random localisation of elements in a genome. Due to cases where multiple elements of the same type occur at the same location in our real data, we counted the number of locations where elements or breakpoints were coincident and not individual elements when we chose the numbers of elements in our simulation. These locations cannot by definition occur at the same place, so we chose our simulation samples without replacement. Observed data for the different categories is shown as vertical bars placed onto the chart. These categories are un-nested so as to obtain independence from other elements in each pairwise category of elements (*i.e.* The categories with more than 2 elements do not contain combinations of pairwise coincidences of elements in the category).

sites of integration of transposed genes in *S. cerevisiae*. For each pair of types of site we calculated the degree of overlap, and compared the result to one million simulations where the same numbers of sites were assigned to randomly chosen intergenic locations in the ancestor. The most significant pairwise associations all involve the locations of tRNA genes, which coincide significantly with evolutionary breakpoints ($P_{FDR} \leq 0.00002$), origins

of replication ($P_{FDR} \leq 0.00023333$), and sites at which genes were gained ($P_{FDR} \leq 0.00051429$) (Figure 2.6, Table 2.2).

Evolutionary breakpoints and sites of gene gain both showed stronger association with sites of tRNA genes than with origins of replication, but it should be noted that the locations of essentially all the tRNA genes are known whereas it is probable that many origins have not yet been identified (Nieduszynski et al. 2007). There are several plausible mechanisms by which tRNA genes could precipitate genomic rearrangements. tRNA genes exist in multiple near-identical copies in the genome, so illegitimate recombination between these sequences could result in reciprocal translocations (Fischer et al. 2000). Ty elements tend to integrate beside tRNA genes and also provide long sections of near-identical sequence scattered around the genome that could be substrates for ectopic recombination. It is known that replication forks tend to stall near highly-expressed genes (such as tRNA genes), and stall sites are thought to be hotspots for chromosomal rearrangements (Admire et al. 2006; Labib and Hodgson 2007). tRNA genes are also found to be present in “fragile” sites which can create chromosome instability (Admire et al. 2006). Finally, it has recently been shown that recombination between two nearly identical tRNA genes in *S. cerevisiae* is increased when both genes are being actively transcribed (Pratt-Hyatt et al. 2006). These observations suggest that tRNA genes, and perhaps neighbouring Ty elements, are likely to play a key role in evolutionary rearrangements in yeast.

The association between Ty insertion hotspots and tRNA genes is dependent on RNA Polymerase (Pol) III activity at the sites, and a general trend for insertion in front of Pol III transcribed genes has been demonstrated (Devine and Boeke 1996). Other factors influencing the insertion of Ty elements at tRNA genes may be the chromatin structure directly upstream of the genes, which may be directly influenced by Pol III removing nucleosomes or altering the local DNA structure (Morse, Roth, and Simpson 1992; Devine and Boeke 1996; Bachman, Eby, and Boeke 2004). It is known that integration occurs at high frequencies in nucleosome-free DNA *in vitro* (Devine and Boeke 1996). A similar effect may be occurring for the newly gained genes in *S. cerevisiae*, whereby they are more likely to reintegrate into spots of nucleosome-free DNA. As tRNAs are very highly

transcribed, these sites may offer the most opportunities for the integration of novel genes. Another possibility is that many of these new genes have been retrotransposed by Ty elements, which would target these genes to RNA Pol III-transcribed sites such as tRNAs. It has been shown that gene conversion or retrotransposition of gene sequences can occur through a reverse transcribed cDNA intermediate that requires the expression of Ty elements, presumably for the actions of the reverse transcriptase and integrase (Derr, Strathern, and Garfinkel 1991; Derr and Strathern 1993). Insertion without the integrase can also potentially occur through homologous recombination into pre-existing Ty elements or solo LTR sequences (Derr and Strathern 1993).

2.3.5 Categories of gained genes

The 121 sites at which genes are inferred to have been gained during the evolution of *S. cerevisiae* since the WGD contain a total of 153 gained genes. The *S. cerevisiae* gene set that we used in this study consists only of genes that are conserved between *S. cerevisiae* and at least one other *Saccharomyces sensu stricto* species (dN/dS ratio < 1 in the analysis of Kellis *et al.* (2003)), or duplicates of other genes in *S. cerevisiae* so we can be confident that the gains are real genes and not annotation artifacts. Some of these genes are unique to *S. cerevisiae* and *sensu stricto* species, while others are shared by the other polyploid species, allowing some inference of the relative age of the gain (Table 2.3).

We classified the gained genes according to their apparent mechanism of gain (Figure 2.7). They consist of *tandem duplications* (where a gained gene is inserted beside its progenitor), *duplicative insertions* (where a progenitor gene that remains in a conserved syntenic context became duplicated to produce a new gene that appears as an insert relative to the ancestral order), and *orphans* (where a gene that appears as an insertion relative to the ancestral gene order has no apparent progenitor). Some orphans are members of orphan families, where all members of the family are insertions relative to the ancestor. In some cases orphan genes contain an identifiable protein domain but there is no *S. cerevisiae* gene

Table 2.2 Results of random genomic element colocalisation simulation

Colocalised elements	Observed number in genome	Proportion of simulated distribution $\geq$ Observed	Simulation mean	P_{FDR}^*
<i>Pairwise</i>				
tRNA, Breakpoint	31	$<10^{-6}$	10.1151	0.00002
ARS, tRNA	27	0.00007	12.04209	0.00023333
tRNA, Gain	15	0.00018	5.3041	0.00051429
tRNA, Transposition	5	0.0487	2.0122	0.08116667
Gain, Transposition	3	0.07266	0.96821	0.09688
ARS, Transposition	5	0.06863	2.2153	0.09804286
ARS, Gain	8	0.22524	5.82289	0.25026667
Breakpoint, Gain	7	0.21662	4.89242	0.25484706
Breakpoint, Transposition	3	0.2837	1.8535	0.29863158
Breakpoint, ARS	10	0.68202	11.12061	0.68202
<i>Multiple</i>				
tRNA, Breakpoint, Gain	13	$<10^{-6}$	0.24963	4×10^{-6}
tRNA, ARS, Gain	9	$<10^{-6}$	0.31247	5×10^{-6}
tRNA, ARS, Breakpoint, Transposition	5	$<10^{-6}$	0.00527	6.6667×10^{-6}
tRNA, ARS, Breakpoint	8	$<10^{-6}$	0.56837	0.00001
tRNA, Breakpoint, Gain, Transposition	1	0.00215	0.00215	0.005375
ARS, Breakpoint, Gain	3	0.00282	0.27476	0.00564
ARS, Breakpoint, Gain, Transposition	1	0.00256	0.00256	0.00568889
tRNA, ARS, Breakpoint, Gain	1	0.01395	0.01403	0.02536364
ARS, Gain, Transposition	1	0.05384	0.05508	0.08283077
tRNA, Breakpoint, Transposition	1	0.09029	0.09436	0.1128625

* P_{FDR} is the P -value corrected for multiple testing by the Benjamini & Hochberg method (Benjamini and Hochberg 1995).

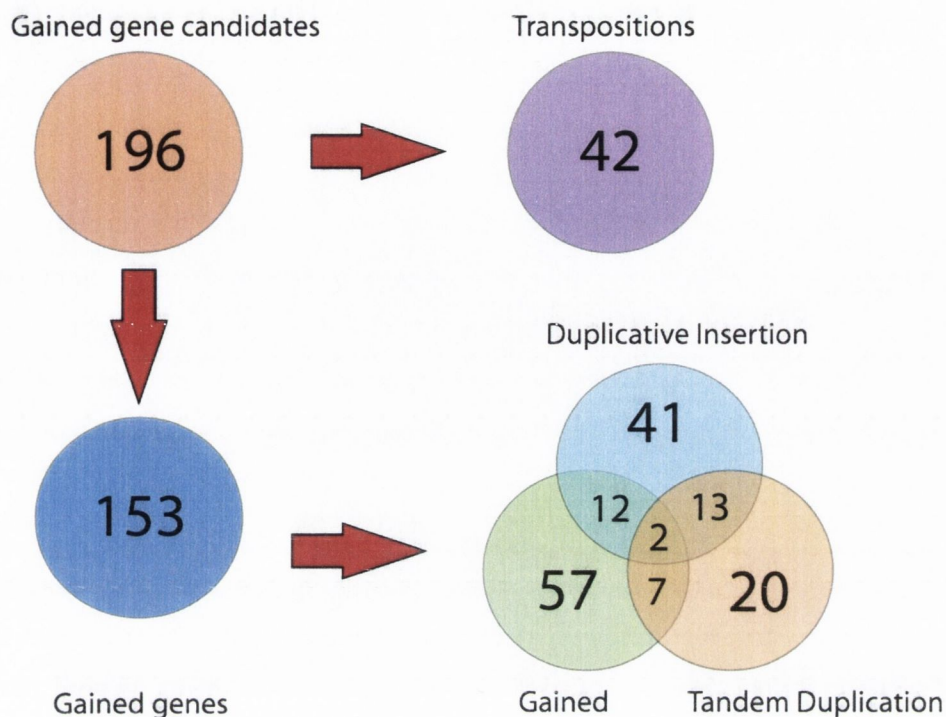


Figure 2.7 Breakdown of the filtering for potential gene gain candidates. The starting set of 196 genes the difference between the ancestral set of genes and the current set of all non-subtelomeric, non-Ty-related *S. cerevisiae* genes that are orthologous. Of the 196 starting genes, 42 have strong non-syntenic orthology to genes in at least 1 pre-duplication species, and as such are likely transpositions which are temporally ambiguous. The remaining 153 genes are potential gains, and can be placed into 3 overlapping categories. Duplicative insertions and tandem duplications are genes where a likely origin is known. In tandem duplications the parental gene is linked to the new gene unless disrupted by a chromosomal rearrangement. Duplicative insertion genes have a good candidate parental gene (that is often orthologous to an ancestral locus), but they are located at a unique location in the *S. cerevisiae* genome. The remaining set of gained genes are of unknown origin and are scattered around the genome of *S. cerevisiae*. Some of the gains can belong to more than 1 category which is shown in the overlapping sections. For example, a gene that was gained then duplicatively inserted and subsequently tandemly duplicated would belong in the intersection of all three groups.

in a conserved syntenic location that is similar enough in sequence to be regarded as a possible progenitor of the orphan. YGOB screenshots with examples of gained genes in each of these categories are shown in Figure 2.8. As well as the three major categorizations for the gained genes, there were also several cases in which genes were found to belong to

more than one of these groups, for example genes that were gained and then duplicated by either transposition or by tandem duplication.

Analysis of the functions of the gained genes should provide insight into the evolutionary pressures that have acted on *S. cerevisiae* in the period since WGD. Unfortunately, nearly half of the recent gains in the *S. cerevisiae* lineage have no functional information. Of the 121 gained genes in our dataset, all but one are non-essential. Closer examination of the single essential gene (*TEN1*) suggests that even its identification as a gene gain may be an error caused by a very rapid rate of sequence evolution, because a *TEN1* homologue was recently identified in *Schizosaccharomycyes pombe* (Rachidi et al. 2000).

The non-essentiality of gained genes is not surprising because these genes represent likely gains to an organism that was already fully functional in its environment before they were gained. Genes that were gained by duplicative insertion are likely to be the minor isoforms, with the proteins they were duplicated from retaining the major role (*i.e.*, higher expression levels and a more severe knockout phenotype; data not shown).

Table 2.3 Genes inferred to be gains in the *S. cerevisiae* lineage since the WGD.

Gene Name	Systematic Name	Functional Description	Present in other post-WGD Species*
<i>SSA1</i>	YAL005C	heat shock protein 70 (HSP70) family	Spar, Smik, Sbay
<i>PAU7</i>	YAR020C	cell wall; part of the seripauperin multigene family	Spar, Smik
	YAR023C	DUP380 family	Spar, Smik
<i>UIP3</i>	YAR027W	DUP240 family	Spar, Smik, Skud, Sbay
	YAR028W	DUP380 family	Spar, Smik? Skud
	YAR029W	DUP380 family	Spar, Smik? Skud
<i>PRM9</i>	YAR031W	DUP240 family	Spar? Smik? Skud
<i>MST28</i>	YAR033W	DUP240 family	Smik?
	YBR012C	unknown function	Scas ,Spar, Skud, Sbay
	YBR013C	unknown function	Spar, Skud, Sbay
<i>EDS1</i>	YBR033W	Zn2Cys6 transcription factor	Spar, Smik, Skud, Sbay
	YBR063C	unknown function	Spar, Smik, Skud, Sbay
<i>TIP1</i>	YBR067C	cell wall mannoprotein	Scas, Cgla ,Spar, Smik, Sbay
<i>ECM8</i>	YBR076W	cell wall	Cgla, Scas, Kpol ,Spar, Smik, Skud, Sbay
<i>PHO3</i>	YBR092C	phosphate metabolism; thiamine	Spar, Smik, Skud, Sbay

<i>PHO5</i>	YBR093C	phosphate metabolism; thiamine	Spar, Smik, Skud, Sbay
	YBR184W	alpha-galactosidase	Spar, Smik, Skud, Sbay
<i>NTC20</i>	YBR188C	spliceosome subunit	Cgla, Scas, Kpol, Spar, Smik, Sbay
	YBR285W	unknown function	Spar, Smik, Skud, Sbay
	YCL021W-A	unknown function	Spar, Smik, Skud, Sbay
	YCR007C	DUP240 family	Spar, Smik, Skud, Sbay
<i>SRD1</i>	YCR018C	GATA finger	Spar, Smik, Skud, Sbay
<i>DIA3</i>	YDL024C	Thiamin	Spar, Sbay
<i>BSC1</i>	YDL037C	cell wall	Spar, Smik, Sbay
	YDL038C	unknown function	
<i>PRM7</i>	YDL039C	Pheromone-regulated protein	Spar, Skud, Sbay
<i>PMT5</i>	YDL093W	cell wall modification	Cgla, Spar, Skud
	YDL159W-A	unknown function	Sbay
<i>ENA5</i>	YDR038C	Na pump	
<i>ENA2</i>	YDR039C	Na pump	Spar, Sbay
	YDR042C	unknown function	Spar, Smik, Skud, Sbay
<i>PHM6</i>	YDR281C	phosphate-modulated	Spar, Smik, Skud, Sbay
<i>PAU2</i>	YEL049W	part of the seripauperin multigene family	Spar? Smik, Skud? Sbay
	YEL057C	telomere maintenance	Spar, Smik, Skud, Sbay
<i>FCY21</i>	YER060W	Major Facilitator Superfamily (MFS); purine permease	Scas, Spar, Smik, Skud, Sbay
<i>FCY22</i>	YER060W-A	Major Facilitator Superfamily (MFS); purine permease	Scas, Kpol, Spar, Smik, Sbay
	YER066W	unknown function	Spar, Smik, Sbay
	YER085C	unknown function	Spar, Smik, Skud, Sbay
<i>Hexose transporter (HXT)10</i>	YFL011W	Major Facilitator Superfamily (MFS); hexose transporter	Spar, Smik, Skud, Sbay
	YFL012W	unknown function	Spar, Smik, Skud, Sbay
<i>PAU5</i>	YFL020C	part of the seripauperin multigene family	Spar, Skud, Sbay
	YFR012W-A	unknown function	Spar
	YFR026C	unknown function	Spar, Smik
	YFR035C	unknown function	Spar, Smik
	YGL015C	unknown function	Spar, Smik, Skud, Sbay
	YGL039W	GRE2 family oxidoreductase	Spar, Smik, Skud, Sbay
<i>MST27</i>	YGL051W	DUP240 family; vesicle organization and biogenesis	
<i>PRM8</i>	YGL053W	DUP240 family, pheromone regulated	Scas?
<i>GPG1</i>	YGL121C	Gprotein gamma mimic	Cgla, Scas, Kpol, Spar, Smik, Skud, Sbay
	YGL157W	GRE2 family oxidoreductase	Cgla, Spar, Smik, Skud, Sbay
	YGL230C	unknown function	Spar, Smik, Skud, Sbay
<i>PDC6</i>	YGR087C	pyruvate decarboxylase, thiamine dependent	Cgla, Spar, Sbay
	YGR153W	unknown function	Spar, Smik, Skud, Sbay
<i>GTO1</i>	YGR154C	putative cytosolic Omega class glutathione transferase	Spar, Skud
	YGR201C	unknown function	Spar, Smik, Skud? Sbay
	YHL026C	unknown function	Cgla, Kpol, Spar, Smik, Skud, Sbay
<i>DOG2</i>	YHR043C	2-deoxyglucose-6-phosphate phosphatase	Cgla, Scas, Spar, Smik, Skud, Sbay
<i>CUP1-1</i>	YHR053C	Metallothionein	
	YHR054C	unknown function	

<i>CUP1-2</i>	YHR055C	Metallothionein	Spar, Smik, Skud, Sbay
	YHR126C	unknown function	Spar, Smik, Skud, Sbay
<i>FMP22</i>	YHR198C	unknown function	Spar, Smik, Sbay
<i>FAA3</i>	YIL009W	long chain fatty acyl-CoA synthetase	Spar, Smik, Skud
<i>PDR11</i>	YIL013C	Pleiotropic drug resistance (PDR); ATP Binding Cassette (ABC) transporter	Scas, Spar, Smik, Skud, Sbay
	YIL014C-A	unknown function	Spar, Skud, Sbay
	YIL029C	unknown function	Smik
	YIL089W	unknown function	Smik, Skud, Sbay
	YIL092W	unknown function	Spar, Smik, Skud, Sbay
	YIL102C	unknown function	Spar
	YIL152W	unknown function	Scas, Kpol, Spar, Smik, Sbay
<i>GAT4</i>	YIR013C	GATA transcription factor	Spar, Smik, Skud, Sbay
<i>DAL4</i>	YIR028W	degradation of allantoin (DAL); Major Facilitator Superfamily (MFS) transporter	Scas, Spar, Smik, Skud, Sbay
<i>DAL7</i>	YIR031C	degradation of allantoin (DAL)	Scas, Spar, Smik, Sbay
	YJL027C	unknown function	Spar, Sbay
	YJL037W	unknown function	Spar, Smik, Skud, Sbay
	YJL038C	unknown function	Spar, Smik, Sbay
	YJL043W	unknown function	Spar, Smik, Skud, Sbay
	YJL052C-A	unknown function	
<i>TDH1</i>	YJL052W	GAPDH	Spar, Smik, Skud, Sbay
	YJL118W	unknown function	Spar, Smik, Skud, Sbay
	YJL144W	response to dessication	Scas, Spar, Smik, Skud, Sbay
	YJR079W	unknown function	
<i>ABM1</i>	YJR108W	Microtubule cytoskeleton organization and biogenesis	Spar, Skud, Sbay
	YJR120W	mitochondrion organization and biogenesis	Spar, Smik
	YKL106C-A	unknown function	Spar
<i>PAU17</i>	YLL025W	unknown function	Spar
<i>TEN1</i>	YLR010C	telomere length	Scas, Kpol, Spar, Smik, Skud, Sbay
<i>LOT6</i>	YLR011W	putative NADPH-dependent FMN reductase	Spar, Skud, Sbay
	YLR012C	unknown function	Agos, Klac, Scas, Spar, Smik, Skud, Sbay
<i>GAT3</i>	YLR013W	GATA transcription factor	Spar, Smik, Skud, Sbay
	YLR036C	unknown function	Spar, Smik, Sbay
<i>DAN2</i>	YLR037C	cell wall mannoprotein	Spar, Skud, Sbay
	YLR042C	unknown function	Spar, Smik, Skud, Sbay
<i>GAL2</i>	YLR081W	Major Facilitator Superfamily (MFS); Hexose transporter (HXT)	Spar, Sbay
<i>SRL2</i>	YLR082C	possible role in DNA repair	Spar, Sbay
<i>YPS3</i>	YLR121C	GPI-anchored aspartyl protease (cell wall)	Spar, Sbay
	YLR125W	unknown function	Spar, Smik, Skud, Sbay
<i>PDC5</i>	YLR134W	pyruvate decarboxylase, thiamine dependent	Spar, Skud, Sbay
<i>RNH203</i>	YLR154C	ribonuclease H	Spar, Skud, Sbay
<i>ASP3-1</i>	YLR155C	Cell-wall L-asparaginase II	
<i>ASP3-2</i>	YLR157C	Cell-wall L-asparaginase II	
<i>ASP3-3</i>	YLR158C	Cell-wall L-asparaginase II	
<i>ASP3-4</i>	YLR160C	Cell-wall L-asparaginase II	
	YLR162W	unknown function	Scas

	YLR179C	unknown function	Scas?, Spar, Smik, Skud
	YLR297W	unknown function	Spar, Smik, Skud, Sbay
<i>CDA1</i>	YLR307W	spore wall chitin deacetylase	Cgla, Spar, Smik, Skud, Sbay
	YLR415C	unknown function	Spar
	YML007C-A	unknown function	Spar, Smik, Skud, Sbay
	YML053C	unknown function	Spar, Smik, Skud, Sbay
<i>HUG1</i>	YML058W-A	mec1 checkpoint	Spar, Smik, Skud, Sbay
<i>PLB2</i>	YMR006C	Phospholipase	Cgla, Spar, Smik, Skud, Sbay
<i>SPO20</i>	YMR017W	prospore membrane	Spar, Smik, Skud, Sbay
	YMR018W	unknown function	Spar, Smik, Skud, Sbay
<i>RSF1</i>	YMR030W	respiratory growth	Spar, Smik, Skud, Sbay
<i>ARG80</i>	YMR042W	MADS box transcription factor	Cgla, Scas, Kpol, Spar, Smik, Sbay
<i>AAC1</i>	YMR056C	Respiration	Kpol, Spar, Smik, Sbay
	YMR147W	unknown function	Scas, Kpol, Spar, Smik, Skud, Sbay
<i>PA13</i>	YMR174C	proteinase inhibitor	Scas, Spar, Smik, Skud, Sbay
<i>GTO3</i>	YMR251W	putative cytosolic Omega class glutathione transferase	Smik, Sbay?
<i>ADH2</i>	YMR303C	Fermentation	Spar, Smik, Skud, Sbay
	YNL018C	unknown function	Spar, Sbay
	YNL019C	unknown function	Spar, Sbay
	YNL033W	unknown function	Spar, Smik, Skud, Sbay
	YNL034W	unknown function	Spar, Smik, Skud, Sbay
	YNL046W	unknown function	Spar, Smik, Skud, Sbay
	YNL146W	unknown function	Cgla, Scas, Kpol, Spar, Smik, Skud, Sbay
<i>SLZ1</i>	YNL196C	sporulation specific protein, Leucine Zipper Motif	Scas, Spar, Smik, Skud, Sbay
	YNL208W	unknown function	Cgla, Scas, Kpol, Spar, Smik, Skud, Sbay
	YOL014W	unknown function	Spar, Smik, Skud, Sbay
<i>IRC10</i>	YOL015W	unknown function	Spar, Smik, Skud, Sbay
	YOL083W	unknown function	Spar, Smik, Skud, Sbay
<i>ZEO1</i>	YOL109W	plasma membrane	Cgla, Spar, Smik, Skud, Sbay
<i>TIR4</i>	YOR009W	cell wall mannoprotein	Cgla, Spar, Smik, Sbay
<i>TIR2</i>	YOR010C	cell wall mannoprotein	Cgla, Scas, Spar, Smik, Skud, Sbay
<i>AUS1</i>	YOR011W	Pleiotropic drug resistance (PDR); ATP Binding Cassette (ABC) transporter	Cgla, Scas, Spar, Smik
<i>CRS5</i>	YOR031W	Metallothionein	Spar, Smik, Skud, Sbay
<i>THI71</i>	YOR071C	Major Facilitator Superfamily (MFS); transporter of thiamine or related compound;	Spar, Smik
	YOR186W	unknown function	Cgla, Spar, Smik, Skud, Sbay
<i>THI72</i>	YOR192C	Major Facilitator Superfamily (MFS); putative thiamine permease	Spar, Smik
<i>WTM1</i>	YOR230W	transcription repressor (WD40 repeats)	Cgla, Scas, Spar, Smik, Skud, Sbay
<i>HUA2</i>	YOR284W	actin patch	Spar, Smik?, Skud, Sbay
<i>PDR10</i>	YOR328W	Pleiotropic drug resistance (PDR); ATP Binding Cassette (ABC) transporter	Spar, Smik, Sbay
<i>ECM23/SRD2</i>	YPL021W	GATA transcription factor	Spar, Smik, Skud, Sbay
<i>SMA1</i>	YPL027W	prospore membrane	Spar, Smik, Skud, Sbay
<i>PMA2</i>	YPL036W	Plasma membrane H ⁺ -ATPase	Skud, Sbay
<i>OYE3</i>	YPL171C	NADPH oxidoreductase	Spar, Sbay

	YPL257W	unknown function	Spar, Smik, Sbay?
<i>THI21</i>	YPL258C	thiamine metabolism	Spar, Smik, Skud, Sbay
	YPL264C	Major Facilitator Superfamily (MFS) family - function unknown	Spar, Smik, Sbay
	YPR027C	unknown function	Spar, Smik, Skud, Sbay
	YPR071W	unknown function	Spar, Smik, Skud, Sbay
	YPR078C	Meiosis	Scas, Kpol, Spar, Smik?, Skud, Sbay
<i>THI22</i>	YPR121W	thiamine metabolism	Spar, Smik

* Uncertainties of presence in a species are marked with a “?”. The species abbreviations are: *K. polysporus* (Kpol), *S. castellii* (Scas), *C. glabrata* (Cgla), *S. bayanus* (Sbay), *S. kudriavzevii* (Skud), *S. mikatae* (Smik), *S. paradoxus* (Spar).

MIPS functional catalogue analysis revealed several functional categories that appear to be enriched ($P < 0.05$) in the set of gained genes relative to their frequencies in the whole gene set of *S. cerevisiae* (Table 2.4). As our gained genes are potentially interesting in terms of the recent evolution of *S. cerevisiae*, even a single gene may provide insights into its specific biology, so largescale functional analyses may only uncover some of the larger trends in functional evolution of *S. cerevisiae*.

2.3.6 Functions of genes gained in the *S. cerevisiae* lineage since WGD.

In the sections below we discuss some of the more interesting categories of gained genes. The gene function information in these sections is derived primarily from summaries in the SGD database (Nash et al. 2007).

2.3.6.1 Nitrogen metabolism and transport.

One of the largest functional categories in the gained set is genes that *S. cerevisiae* uses to metabolize alternative nitrogen sources. This category includes a tandem array of four copies of *ASP3*, which function in asparagine and aspartate catabolism and are induced during nitrogen starvation. We also found *DAL4* and *DAL7*, two genes that function in the uptake and utilization of the purine derivative allantoin as a source of nitrogen. *DAL4* and *DAL7* are located within a cluster of six genes that all function in allantoin catabolism, and which was assembled in a dramatic gene relocation event that coincided with biochemical

Table 2.4 Enriched MIPS functional categories in the *S. cerevisiae* lineage gene gains

MIPS functional category*	Genes	<i>P</i> -value	<i>P</i> _{FDR} [†]
degradation of asparagines	<i>ASP3-1 ASP3-2 ASP3-3 ASP3-4</i>	0.0000018	0.000153781
nutrient starvation response	<i>PHO5 ASP3-1 ASP3-2 ASP3-3 ASP3-4</i>	0.0000222	0.001246751
nucleotide/nucleoside/nucleobase transport	<i>FCY21 FCY22 AAC1 THI71 THI72</i>	0.0001040	0.00290881
Heavy metal binding (Cu, Fe, Zn)	<i>CUP1-1 CUP1-2 CRS5</i>	0.0007780	0.018689249
allantoin and allantoate transport	<i>DAL4 THI71 THI72</i>	0.0007780	0.016353093
vesicle formation	<i>PRM9 MST28 MST27</i>	0.0016100	0.030022367
alcohol fermentation	<i>PDC6 PDC5 ADH2</i>	0.0036300	0.055405946
Metal binding	<i>CUP1-1 CUP1-2 RNH203 CRS5</i>	0.0041800	0.058598145
utilization of vitamins, cofactors, and prosthetic groups	<i>PHO3 PHO5</i>	0.0086900	0.11231913
transport facilities	<i>MST28 ENA5 ENA2 MST27 PDR11 AAC1 AUS1 THI71 THI72 PDR10 PMA2</i>	0.0095800	0.115134289
nitrogen, sulfur and selenium metabolism	<i>DAL7 GAT3 ASP3-1 ASP3-2 ASP3-3 ASP3-4 ARG80</i>	0.0103000	0.115990038
inorganic chemical agent resistance (e.g. heavy metals)	<i>CUP1-1 CUP1-2</i>	0.0120000	0.125680814
Golgi	<i>MST28 MST27</i>	0.0157000	0.155151909
C-2 compound and organic acid metabolism/catabolism	<i>PDC6 ADH2</i>	0.0199000	0.185340742
ABC transporters	<i>PDR11 AUS1 PDR10</i>	0.0316000	0.265821866

* Some overlapping categories have been deleted

[†] *P*_{FDR} is the *P*-value corrected for multiple testing by the Benjamini & Hochberg method (Benjamini and Hochberg 1995).

reorganization of the upstream steps of the *DAL* pathway. The changes in this pathway have been associated with the development of a facultative anaerobic lifestyle in the *S. cerevisiae* lineage (Wong and Wolfe 2005; Merico et al. 2007).

The gained set includes five genes classified in the MIPS scheme as nucleotide / nucleoside / nucleobase transporters. These five (*THI71*, *THI72*, *FCY21*, *FCY22* and *AAC1*) are one-quarter of all such transporters annotated in the *S. cerevisiae* genome. *THI71* and *THI72* are part of the same transporter family of uracil/uridine/allantoin permeases as *DAL4* and are involved in 5-fluorocytosine sensitivity (Nelissen, De Wachter, and Goffeau 1997; Paluszynski et al. 2006). Nucleic acid bases and their derivatives can be used directly for DNA or RNA synthesis, or as a source of nitrogen atoms. The increase in the number of genes for the transport of cytosine (*FCY21*, *FCY22*) and purines (*THI71*, *THI72*, *DAL4*)

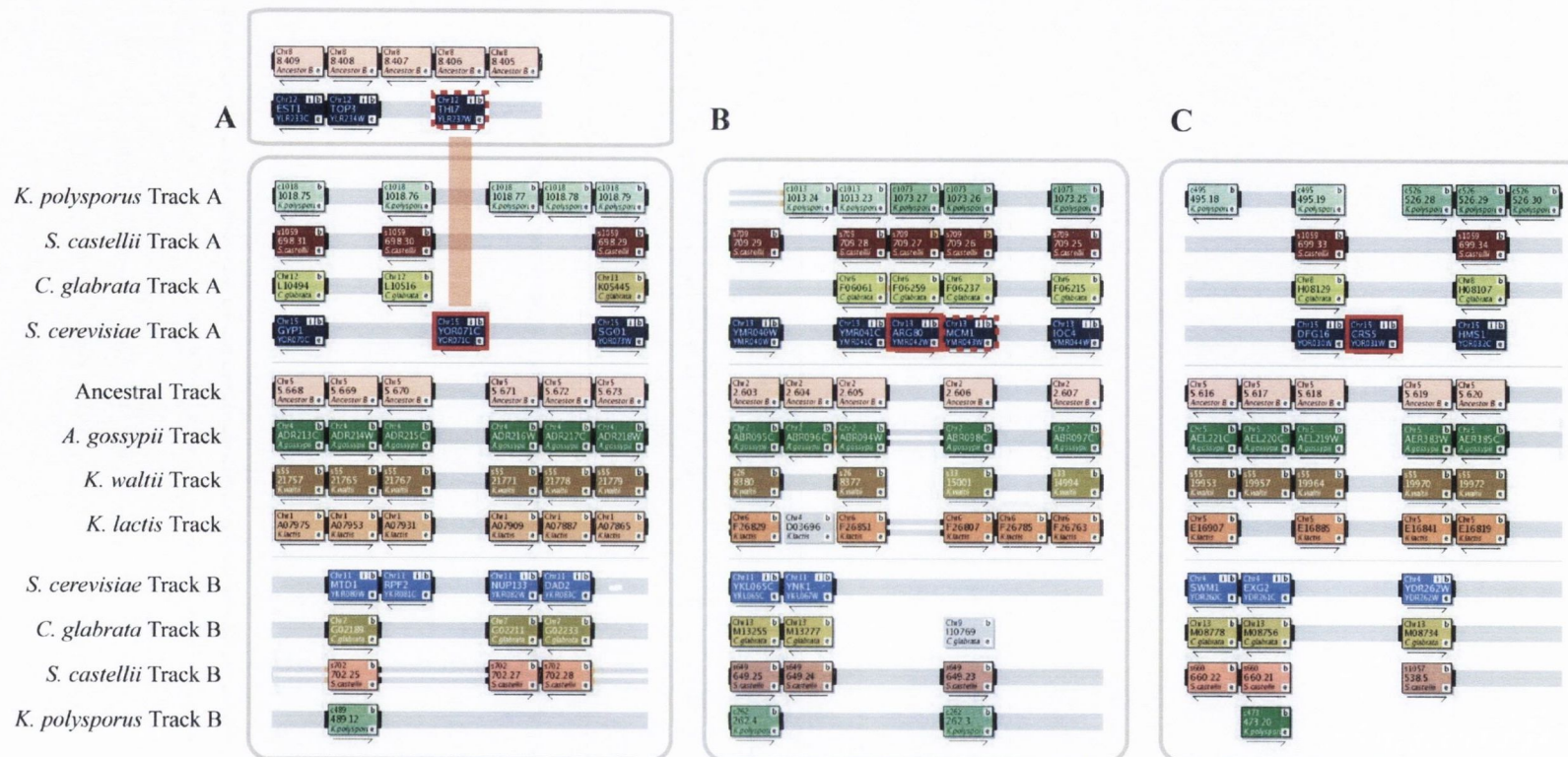


Figure 2.8 Examples of each category of novel genes in *S. cerevisiae*, visualised by YGOB (<http://wolfe.gen.tcd.ie/YGOB/>). The gene of interest is marked by a red box. A) YOR071C is a thiamine permease of the Major Facilitator Superfamily (MFS), and represents a duplicate insertion of *THI7* which is located at an ancestral locus (Anc_8.406). B) *ARG80* is a transcription factor involved in regulating arginine responsive genes and represents an old tandem duplication of *MCM1* (dashed red box). C) *CRS5* is a metallothioneine involved in copper tolerance. It is an example of a gain of

would allow an increased uptake of these molecules from the surrounding environment, and perhaps a faster growth rate or reduced expenditure of ATP on nucleobase synthesis. *AAC1* codes for a mitochondrial ATP/ADP translocator and is part of a family of proteins located in the mitochondrial membrane that are vital for anaerobic growth (Drgon et al. 1991).

2.3.6.2 Sterol transporters

Other genes in the gained set also have transporter functions. *AUS1*, *PDR11* and *YJR120W* are all putatively involved in sterol uptake (Lai et al. 2005; Jungwirth and Kuchler 2006; Raychaudhuri and Prinz 2006; Reiner et al. 2006). Sterol biosynthesis requires the presence of molecular oxygen (Raychaudhuri and Prinz 2006; Reiner et al. 2006), and the gains of genes involved in the transport of sterols from the environment would facilitate growth in hypoxic or anaerobic conditions. *AUS1* and *PDR11* are members of the large family of ABC (ATP-binding cassette) transporters and are both found as far back in the hemiascomycete phylogeny as *S. castellii*. These two genes are required for the efficient uptake of sterol in both aerobic and anaerobic conditions (Raychaudhuri and Prinz 2006). *YJR120W* is a gain of unknown origin, and has an unknown function, but is again implicated in sterol uptake (Reiner et al. 2006). All of these genes are duplicative insertions or tandem duplications of other genes that are found in the ancestral content, and thus represent the likely enhancement of already present pathways or possible modification for novel functional ability as in the case of *DAL4*.

2.3.6.3 Thiamin uptake and biosynthesis genes

While *THI71* and *THI72* are involved in the uptake of thiamin (vitamin B1), the set of gained genes also contains three others involved in the biosynthesis of thiamin (*THI21*, *THI22*, *PHO3*), and two (*PDC5* and *PDC6*) that are upregulated in the absence of external thiamin and are reliant on thiamin as a cofactor (Hohmann and Meacock 1998; Llorente, Fairhead, and Dujon 1999; Nosaka et al. 2005; Mojzita and Hohmann 2006). The Thi21 and Thi22 proteins are involved in the joining of two precursor molecules together to form

thiamin phosphate. Thiamin phosphate is not a direct precursor of thiamin diphosphate (ThDP), the biologically active form of thiamin. It is dephosphorylated to thiamin by *PHO3* before it is transported through the cell membrane by Thi71, Thi72 and their progenitor Thi7, where it is acted upon by a pyrophosphokinase to form ThDP. Interestingly, *S. cerevisiae* and the *sensu stricto* species are the only post-WGD species that are prototrophic for thiamin, which is obviously a result of these gene gains.

Thiamin diphosphate is an essential cofactor for enzymes acting in several metabolic pathways. *PDC5* and *PDC6* both code for isoforms of the pyruvate decarboxylase enzyme which utilizes ThDP as a cofactor in the conversion of pyruvate to acetyl-CoA which feeds into the TCA cycle, and are both derived from relatively recent duplication events. ThDP dependent enzymes also play prominent roles in the biosynthesis of leucine, isoleucine, valine and aromatic amino acids (Hohmann and Meacock 1998; Mojzita and Hohmann 2006). All six thiamin related genes mentioned above are relatively recent gains: they are all confined to the *sensu stricto* species except for *PDC6* which is also found in *C. glabrata*. As thiamine is essential in many metabolic pathways, the gain of additional transporters of exogenous thiamin and of biosynthesis genes may represent an important shift in the physiology of *S. cerevisiae*. One possibility, suggested by Thomson *et al.* in the context of the *PDC* and *PHO3* gene duplications, is that selection for increased ethanol production resulted in increased demand for thiamin as a cofactor for pyruvate decarboxylase.

2.3.6.4 Ethanol production and consumption

Acetaldehyde made by pyruvate decarboxylase is converted to ethanol by the alcohol dehydrogenase 1 (*ADH1*) gene product. *ADH2* is a recent duplicate of this gene that is responsible for the opposite reaction to Adh1: the consumption of ethanol present in the environment (Thomson *et al.* 2005). Resurrection of the Adh1/Adh2 ancestor protein suggests that its kinetic behavior was optimised to produce alcohol, and not to consume it (Thomson *et al.* 2005). This implies that the duplication of *ADH1* allowed the novel function of alcohol reuse to evolve in the new duplicate (*ADH2*). *ADH2* is only found in

the *sensu stricto* species, implying that the ability to use external ethanol as an alternative carbon source is a relatively recent function. Thomson *et al.* (2005) suggested that prior to this gene duplication, the ancestral yeast may have used ethanol production as a way to recycle NADH produced in glycolysis, and may also have secreted ethanol as a defensive or competitive measure against other microorganisms in its environment. It is likely that the ability to consume ethanol as a carbon source in yeast arose to limit the loss of resources to the environment. The pathway of conversion of ethanol to acetyl-CoA consumes one ATP molecule per molecule of ethanol, making it less energetically efficient than glycolysis as a path into the citric acid cycle. The advantage of this "make-accumulate-consume" strategy is that *S. cerevisiae* cells can consume glucose from the environment at a faster rate when fermenting than when respiring, so by quickly converting an environmental food resource into a form (ethanol) that cannot be used by most of its competitors (and is even toxic to them), *S. cerevisiae* grabs a larger share of the pot for itself (Thomson et al. 2005; Conant and Wolfe 2007).

Interestingly, a duplication that created the glyceraldehyde 3-phosphate dehydrogenase (GAPDH) minor isoform gene *TDH1* has the same phylogenetic profile as *ADH2*. GAPDH is the enzyme that reduces NAD^+ to NADH in the conversion of glyceraldehyde 3-phosphate to 1,3-diphosphoglycerate in glycolysis (Valadi et al. 2004; Thomson et al. 2005). *TDH1* appears to have properties not displayed by the other *TDH* isogenes, and its expression is regulated by reductive stress caused by an excess of NADH in the cell (Valadi et al. 2004). *TDH1* is also required during gluconeogenesis, which is the process by which glucose is synthesised from alternative carbon sources such as ethanol or glycerol.

2.3.6.5 Amplified membrane protein gene families

Several protein families have been amplified in *S. cerevisiae* and other post-WGD yeast species. While a few of the members are found in *S. castellii* or *C. glabrata*, the majority of the genes in these families are either specific to *S. cerevisiae*, or to the *sensu stricto* species. Most of these families also have members in subtelomeric regions which were not included in our gained gene set. The largest of these families are the *DAN/TIR*, *PAU* and

DUP240 families. There are nine members of the *DUP240* family in our gained genes dataset, six of which are in a subtelomeric tandem array on chromosome I. The *DUP240* genes are non-essential membrane spanning proteins that have been suggested to function in membrane trafficking (Despons et al. 2006). Two of the genes from the chromosome I tandem array (*PRM9* and *MST28*) appear to have been duplicated segmentally to give rise to *PRM8* and *MST27* on chromosome VII, which maintain relative orientation to each other. It has been shown however that these two regions are differentially amplified or even absent in different strains of *S. cerevisiae*, indicating that their presence in our gained set may not be biologically significant (Leh-Louis et al. 2004; Despons et al. 2006).

Several *DAN/TIR* and *PAU* (seripauperin) family genes are found in the gained gene set (Table 3). The *DAN/TIR* proteins show homology to the *PAU* proteins, but differ from them by containing extensive serine-rich domains. The gained genes *PAU2*, *PAU5*, *PAU7*, *PAU17* and *DAN2* are limited to the *sensu stricto* species, while *TIR4* is also present in *C. glabrata*, and *TIP1* and *TIR2* are also found in *C. glabrata* and *S. castellii*. These genes all code for either cell wall or integral membrane proteins and are induced by anaerobiosis (Rachidi et al. 2000; Abramova et al. 2001a; Abramova et al. 2001b; Lai et al. 2005; Murata et al. 2006). *TIP1*, *TIR2*, *TIR4* and *DAN2* all appear to function in the remodeling of the cell wall and plasma membrane in response to altered environmental conditions such as anaerobiosis or cold shock (Abramova et al. 2001a; Abramova et al. 2001b; Murata et al. 2006). The *Tip1*, *Tir1* and *Tir2* proteins are predicted to undergo O-mannosylation and GPI anchoring to the membrane (Murata et al. 2006). It is interesting to note that the *PMT5* gene in our gained set is a protein O-mannosyltransferase which is induced during anaerobiosis and is involved in cell wall modification (Lai et al. 2005).

2.3.6.6 GATA family transcription factors

Analysis of the protein domains present in the gained set revealed that four of the nine GATA zinc-finger transcription factors in *S. cerevisiae* have been gained since WGD and are shared by the *sensu stricto* species. Because they belong to the GATA family (Scazzocchio 2000), they were probably all derived by duplication of other GATA

transcription factors. The gained genes *GAT3* and *GAT4* are best reciprocal BLAST hits to each other, as are *SRD2* (also called *ECM23*) and *SRD1*, although they are generally poorly conserved except in their zinc finger domains. *SRD2* is involved in cell wall morphogenesis and may be a negative regulator of genes for pseudohyphal growth (Canizares et al. 2002). *SRD1* was identified as a suppressor of a mutation in *RRP1*, which codes for a nucleolar protein involved in rRNA maturation (Hess, Stanford, and Hopper 1994). Very little is known about the functions of *GAT3* and *GAT4*. Curiously, the functions of the other (non-gained) GATA factors in *S. cerevisiae* are much better defined; they are four regulators of nitrogen metabolism (*GLN3*, *GAT1*, *DAL80*, *DEH1*) (Cooper 2002) and the *ASH1* gene (Cosma 2004).

2.4 Conclusion

Several points need to be borne in mind when interpreting our gene gain candidate data. We must consider that any gene in our set of apparent gains may actually have been present in the ancestor at the time of the WGD. We may find a yet unsequenced species that contains the gene and allows us to infer its presence in the ancestor. Another issue is that of defining single copy orthologues and single copy paralogues. For the majority of loci in YGOB, shared gene loss patterns are informative about the relationships between homologous genes in different post-WGD species, and having accurate assignments is essential for our approach. There are however some regions where the confidence of our assignments is relatively low, particularly in *K. polysporus* as it speciated relatively soon after the WGD (Scannell et al. 2007). In view of this, our list of inferred gene gains is likely to contain some false positives but it should also contain all the genuine gains that occurred in non-subtelomeric regions of the genome.

The set of genes inferred to have been gained on the *S. cerevisiae* lineage is relatively small and their functions point squarely towards increasing adaptation to the "fermentative lifestyle" (Merico et al. 2007). They indicate increasing throughput of the glycolysis and fermentation pathways, and adaptation towards growth in conditions with little oxygen, including modifications to the cell wall and the bypass of biochemical pathways that

require molecular oxygen by importing substances from outside the cell. There are also many gained genes in our set that we have not discussed in detail here because they did not fall into larger functional groups. Further analysis of these gains on an individual basis may reveal insights into the evolution of *S. cerevisiae* and the other species in the WGD clade.

In terms of the genome rearrangement process that occurred *en route* from the pre-WGD ancestor to *S. cerevisiae*, we have found empirical evidence for the reuse of evolutionary breakpoints in *S. cerevisiae*, as well as for genomic hotspots of rearrangement. These hotspots regularly coincide with tRNA genes and origins of replication. We also found that the locations of gained genes coincide with tRNAs more frequently than would be expected by chance.

2.5 Methods

2.5.1 Parsimonious inference of ancestral gene content and order

We used text output from YGOB (Byrne and Wolfe 2005) to compare the gene order and orientation between orthologues in different species. Genes inferred to have been present in the ancestor that existed immediately prior to WGD are those that are in an orthologous relationship between a pre-WGD species and a post-WGD species or those which are paralogous between two post-WGD species. In any case where two neighboring ancestral genes have neighboring ancestral orthologues or paralogues in another species (or its sister region) for which the phylogenetic relationship spans the WGD, we compared gene order of the inferred ancestral loci between pre-WGD and post-WGD chromosomes and between the sister regions in the post-WGD species in order to infer the ancestral state at the point immediately prior to the WGD. In cases where we observe a shared syntenic context between genes in WGD-spanning species we assumed that the most parsimonious explanation is that it represents an ancestral syntenic state. Under this principle, we were able to infer the ancestral gene order of the whole genome of the common ancestor of the WGD clade (Figure 2.1 and Table S2.1 on CD). Although the common ancestor of the pre-

WGD and post-WGD species is much older than that of the duplicated chromosomal regions, if the gene order and content has been maintained in both lineages, then the same genes and order must have been present in the immediate pre-WGD ancestor.

2.5.2 Evolution of ancestral organisation into *S. cerevisiae* gene order

We described the inferred ancestral gene order in terms of synteny blocks of current *Saccharomyces cerevisiae* genes. We manually identified intrachromosomal rearrangements (inversions) between the ancestor and *S. cerevisiae* and reversed them, revising our synteny blocks, in order to more easily identify the endpoints of reciprocal translocations. For each synteny block end not at a telomere, the location is at a position in the ancestral genome that underwent a reciprocal translocation in its transition towards the current *S. cerevisiae* genome. Each synteny block end in the ancestral genome is bordered by another synteny block found elsewhere in the current *S. cerevisiae* genome. The two breakpoints at the ends of two ancestral synteny blocks now adjacent in the current *S. cerevisiae* genome were created by a reciprocal translocation event that joined them together from different ancestral locations. Concurrently the other synteny blocks that border each breakpoint in the ancestral genome were joined together by the same event. We ordered the synteny blocks in the manner in which they are found along each chromosome in *S. cerevisiae*, thus inferring all the interchromosomal rearrangements between the ancestral polyploid genome and *S. cerevisiae* (Figure 2.4). To confirm that the inferred reciprocal translocation events were correct, we found the location in *S. cerevisiae* of the other synteny block ends joined by each event. In cases where these synteny blocks are not adjacent in the *S. cerevisiae* genome, we found the ancestral breakpoint locations of the blocks that are adjacent to each of these blocks in *S. cerevisiae*, inferring another reciprocal translocation event. If the synteny blocks bordering each breakpoint were again not adjacent, this process was repeated. Once this step was completed, each of the breakpoints in synteny in the ancestral gene order relative to *S. cerevisiae* was examined at the equivalent intergenic location in both *A. gossypii* and *S. cerevisiae* genomes for genomic elements such as tRNAs and origins of replication that may have precipitated a rearrangement event.

2.5.3 Inference and analysis of probable gene gains in the *S. cerevisiae* lineage since the WGD

To obtain a set of likely gene gains since the ancestor we subtracted the set of *S. cerevisiae* genes represented in our ancestral reconstruction from the curated set of *S. cerevisiae* genes used in YGOB. YGOB contains a high confidence set of protein-coding that has been manually curated. *S. cerevisiae* genes that are present in YGOB set but absent from the inferred ancestral set are candidates for having been gained in the *S. cerevisiae* lineage after the WGD. We did not include subtelomeric genes from the YGOB set, as orthologous relationships across species break down at the telomeres (Kellis et al. 2003). This candidate set of gains was then manually checked to ensure that there were no possible non-syntenic homologues that were ancestral but missing from our ancestral genome reconstruction due to a breakdown of synteny information. Any cases where a good candidate nonsyntenic homologue was found were removed from the gained set and flagged as a likely transposition event. It is possible that the set of candidate gained genes may also contain ancestral genes that were lost in all of the pre-WGD lineages but have orthologues in more distantly related outgroups.

We tested for enrichment of functional categories in the gained gene set by using the MIPS functional catalog database from the Munich Information Centre for Protein Sequences (http://mips.gsf.de/proj/funcatDB/search_main_frame.html). We applied the Benjamini & Hochberg method to correct the *P*-values for multiple testing (Benjamini and Hochberg 1995). We used KEGG (www.genome.jp/kegg/) to examine the position of gained genes within metabolic pathways.

2.5.4 Computer simulations

To test whether certain genomic sites (evolutionary breakpoints, gene gain sites, transposition sites, tRNA genes, and origins of replication) coincide with each other than expected by chance in *S. cerevisiae*, a Perl script was written to generate simulations of the overlap between five sets randomly chosen from the same source set that was compared to

the observed data. The five sets chosen were based on the actual numbers of locations in the *S. cerevisiae* genome, at non-subtelomeric positions, for each of the five types of site. Each set was represented by choosing random numbers from 1 to 5100 (the estimated number of intergenic spaces between *S. cerevisiae* genes that are at an ancestral locus) without replacement. This operation was iterated one million times to generate a distribution to which we compared the observed data. We considered every pairwise combination of five types of site. To identify whether any pair of types of site co-localised with each other more often than expected by chance, independently of the other three types of site, we compared the random distributions for locations where there are exactly two coincident sites to the observed data, as these locations should be independent of any co-localational biases potentially introduced by other elements at that location. Correction for multiple testing was performed on the simulation *P*-values by applying the Benjamini & Hochberg method (Benjamini and Hochberg 1995).

Genomic Evolution of Polyploid Yeast Species

Chapter 3

Evolutionary erosion and transposition of genes flanking the *MAT* locus in yeast species.

3.1 Abstract

We determined the sequence of parts of the *Zygosaccharomyces rouxii* mating type chromosome, covering the region between its mating-type (*MAT*) locus and a silent cassette (*HML*), as well as *Z. rouxii* regions containing orthologues of all the genes currently on *Saccharomyces cerevisiae* chromosome III. *Z. rouxii* was chosen because it is the closest known outgroup to the whole genome duplication (WGD) clade of yeasts that includes *S. cerevisiae*. We traced the evolutionary path of the mating-type chromosome, which became duplicated in the WGD and gave rise to much of chromosome III and part of chromosome XIV of *S. cerevisiae*. We find that regions bordering the *MAT* locus are being eroded: sections of DNA flanking the *MAT* locus have progressively been deleted from the *MAT* chromosome, and in some cases have re-integrated into new sites elsewhere in the genome. The process of erosion has also eliminated noncoding DNA between the *MAT* locus and the genes that abut it, and these neighboring genes frequently encode truncated proteins. The rate of erosion has accelerated in species that underwent WGD, and we link this acceleration to the possible combined actions of the HO endonuclease and the genetic redundancy introduced by the WGD. We find that the X and Z regions that delimit the mating cassettes have boundaries that can slip to include parts of genes that were once outside the cassette, and that there is feedback of this information from the *MAT* locus to the silent cassettes.

3.2 Introduction

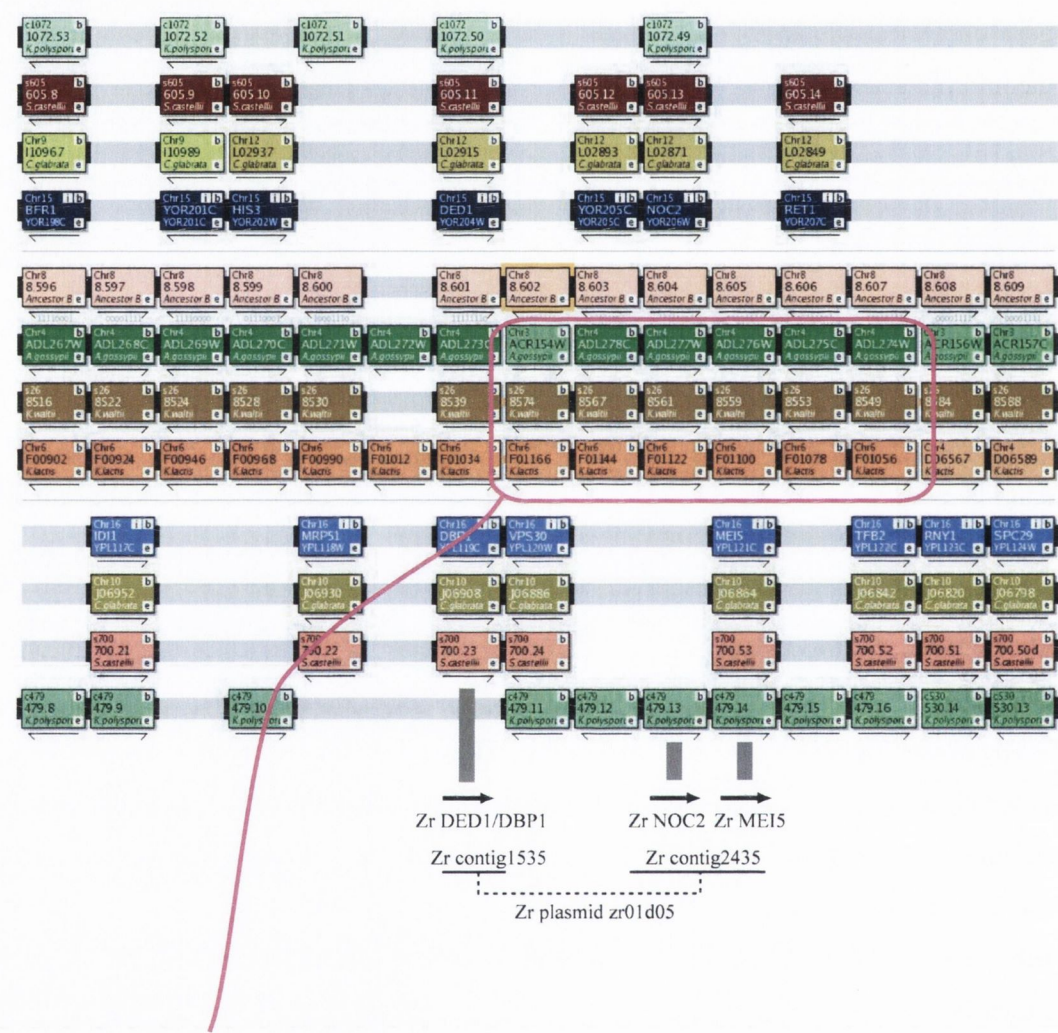
In the previous chapter I inferred the order and content of genes in the ancestral yeast species that existed immediately prior to the WGD event. In this chapter I focus on the evolution of the chromosome that contains the *MAT* locus. Our laboratory previously found that the *MAT* loci of different yeast species were located in syntenic contexts that were similar, but by no means identical, across species (Butler et al. 2004). The evolution of gene order near the *MAT* locus seemed complex, so we decided to revisit this topic using the inferred ancestral genome and additional sequence data.

To increase resolution at which we could compare evolution of the *MAT* chromosome, in addition to using public data from pre-WGD and post-WGD species (Cliften et al. 2003; Dietrich et al. 2004; Dujon et al. 2004; Kellis, Birren, and Lander 2004; Scannell et al. 2007), we undertook partial sequencing of the *Zygosaccharomyces rouxii* *MAT* chromosome and other regions homologous to the *MAT*-containing chromosome of *S. cerevisiae* (chromosome III). *Z. rouxii* is an osmotolerant yeast that is used in the commercial production of soy sauce and miso paste and is also associated with spoilage of sugar-rich foods such as jam and fruit juice (Onishi 1963; Burt et al. 1999). *Z. rouxii* contamination can cause liquid-centre chocolates to explode (Deak and Beauchat 1996).

Our interest in *Z. rouxii* stems from its phylogenetic position: it lies outside but close to the WGD event (Figure 1.4). Kurtzman and Robnett's (2003) analysis placed *Z. rouxii* in Clade 7, which is the closest known clade that is an outgroup to the clades that underwent WGD (Clades 1-6) (Scannell et al. 2007). Previous work has indicated that *Z. rouxii* did not undergo WGD (Sychrova et al. 2000; Wong, Butler, and Wolfe 2002), and we confirm that here. By analyzing random plasmid-end sequence data from the *Z. rouxii* genome (described below) we identified several genomic rearrangements that are shared between *Z. rouxii* and the post-WGD species (on both chromosomal tracks), but not by the other pre-WGD species studied (*A. gossypii*, *K. lactis* and *K. waltii*) (Figure 3.1). These shared rearrangements support Kurtzman's conclusion that *Z. rouxii* is closer to the WGD clade than are any of the other sequenced pre-WGD species.

Chromosome III of *S. cerevisiae* was the first yeast chromosome to be fully sequenced (Oliver et al. 1992), and contains the *MAT* locus which defines cell type. In a haploid yeast cell the *MAT* locus contains either **a**- or α -type genes, which are the two mating-types

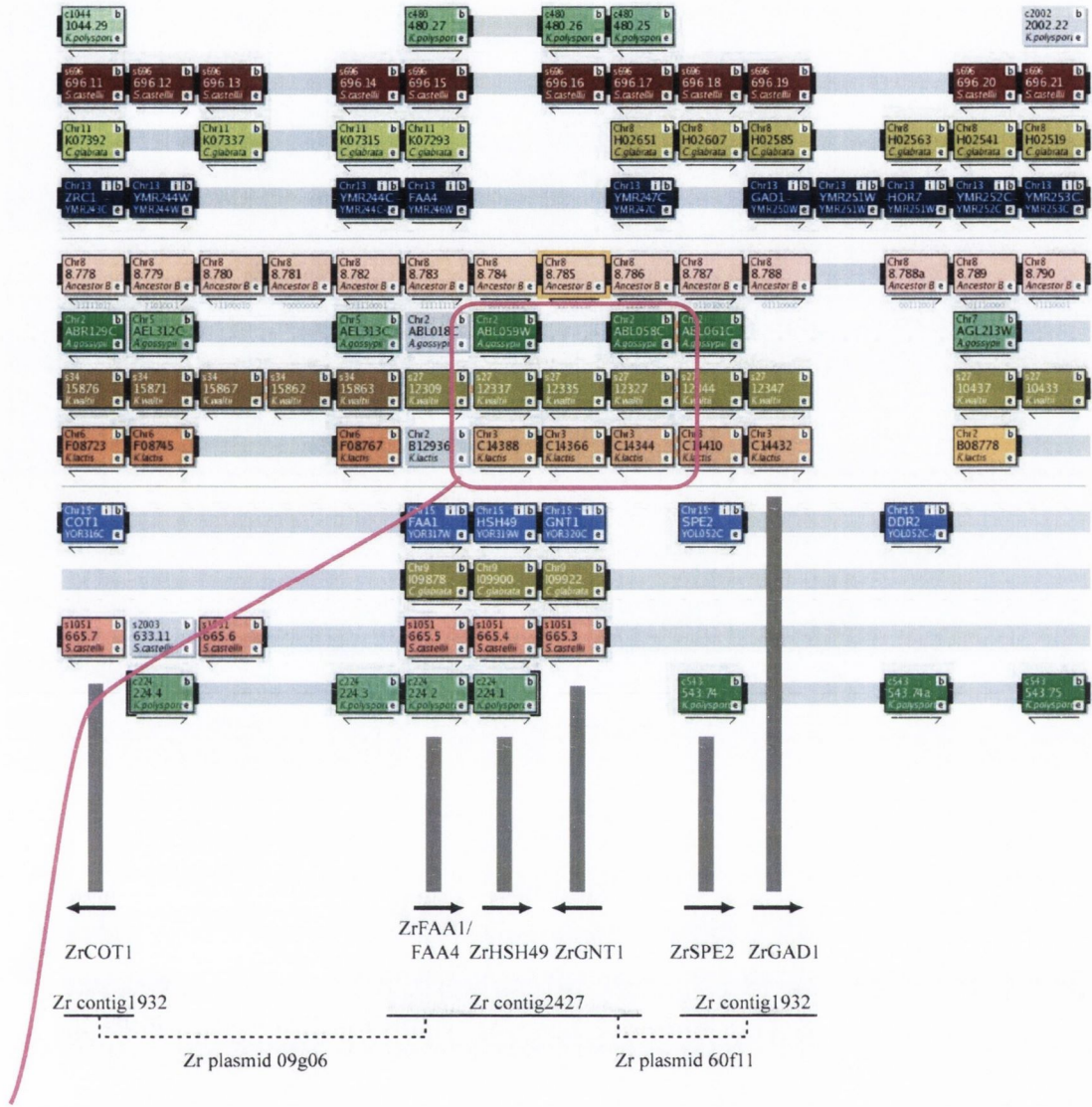
Fig. 3.1a. NOC2 region.



Region inverted in all pre-WGD tracks relative to both post-WGD tracks in all species

Z. rouxii shares the same gene order as the post-WGD species

Fig. 3.1b. GNT1 region.



Genomic map of Zr contig 2276 showing gene models and synteny across various species. The map includes genes from *K. polyspori*, *S. castellii*, *C. glabrata*, *A. gossypii*, and *K. lactis*. A red line highlights a specific genomic region, and three vertical grey bars indicate the positions of ZrYME1, ZrBNA4, and ZrBRN1.

Z. rouxii shares the post-WGD gene order.

80

Figure 3.1 Examples of regions that are collinear between *Z. rouxii* and post-WGD species that differ from the other pre-WGD gene orders. These regions reveal that *Z. rouxii* is more closely related to the post-WGD species than the other pre-WGD species are. The pictures are snapshots from the Yeast Gene Order Browser (YGOB) from the viewpoint of the inferred pre-WGD ancestral track (pink row). Each box represents a gene. Each row represents gene order in a different species/track relative to the viewpoint track. Pre-WGD species tracks are in the middle while the ohnologous post-WGD species tracks are grouped above and below. Genes separated in a row by small black rectangles are collinear, while genes separated by orange marks represent an inversion of gene order relative to the viewpoint track. *Z. rouxii* gene orders are shown below the YGOB snapshot with orthologous relationships shown with grey bars. Figures created in part by Ken Wolfe.

(Hicks 1977; Haber 1998). There are three types of yeast cell: *MAT* α haploids, *MATa* haploids and *MAT* α /*MATa* diploids. Cells with one type of *MAT* locus can only mate with cells of the other type, and diploids do not mate at all (Hicks 1977; Hicks, Strathern, and Herskowitz 1977; Haber 1998). Mating-type switching only occurs in haploid cells that have budded, changing the mating-type of the mother cell from *MATa* to *MAT* α or *vice versa*, with the result that the daughter cells directly surrounding it have the opposite mating-type to their mother cell (Strathern and Herskowitz 1979). The proportion of daughter cells which outcross after meiosis is 0.01 in *Saccharomyces paradoxus*, which is likely to be similar in the closely related *S. cerevisiae* (Tsai et al. 2008). Mother-daughter mating is likely to follow mating-type switching. The switching process occurs by the genes at the *MAT* locus being excised and replaced by an idiomorph of a silent mating cassette from *HML* or *HMR* which are located at either end of chromosome III in *S. cerevisiae*. (The word idiomorph means a version of a *MAT* or *HM* locus, analogous to an allele at other loci. The word allele is not used because the sequences are completely different.) In most strains of *S. cerevisiae*, *HML* contains a copy of the α mating-type information, and the *HMR* contains a copy of the *a* mating-type information. Switching is induced by a double strand break (DSB) at the *MAT* locus caused by the HO endonuclease, and site-specific gene conversion with the complementary silent cassette (Rudin and Haber 1988; Haber 1998).

3.3 Results

3.3.1 *Z. rouxii* MAT chromosome

We determined the sequences of 10 large contigs of genomic DNA from *Z. rouxii* CBS 732, totaling 608,894 bp and containing 331 genes (including 16 tRNA genes and 6 likely pseudogenes) (Table S3.1 on CD). The sequenced regions were chosen to include the homologues of all genes currently on *S. cerevisiae* chromosome III, and parts of the ancestral *MAT* chromosome which were chosen by comparison to *MAT* chromosomes of other pre-WGD species. Due to a chromosomal translocation in *Z. rouxii* that is described below, only four of the sequenced contigs lie on the current *Z. rouxii* *MAT* chromosome, and these span the region from the telomeric side of *HML* to eight genes past the *MAT* locus. This region contains 129 genes, nine of which do not correspond to loci in the inferred ancestral genome, and two of which are tRNA genes. Gene order along the sequenced part of the *Z. rouxii* *MAT* chromosome between *HML* and *MAT* is very similar to the inferred ancestral gene order with only six gene deletions (orthologues of Anc_1.10, Anc_1.21, Anc_1.23, Anc_1.52, Anc_1.56, and Anc_1.114; Table S3.2 on CD).

We found that the *Z. rouxii* *MAT* locus is the site of a Robertsonian chromosomal translocation event. This translocation broke an ancestral linkage by introducing a breakpoint in the intergenic region between *MAT* α 2 (Anc_1.117d) and *PEX2* (Anc_1.118). The *HML*-to-*MAT* α 2 fragment (Anc_1.1 to Anc_1.117d) from the ancestral *MAT* chromosome became fused to the whole of ancestral chromosome Anc2, by creating a new join between *MAT* α 2 and the previously telomeric gene *ATG2* (Anc_2.1) with two novel genes (which we name *ZrCHA1m* and *ZrSIR1*) sandwiched between them. (Figure 3.2). The other product of the translocation is the remainder of Anc1 (genes Anc_1.118 to Anc_1.522), which includes its centromere, with *PEX2* (Anc_1.118) now inferred to be at a telomere. Due to this translocation, all of the other regions that we sequenced from *Z. rouxii*, that are homologous to either the *S. cerevisiae* *MAT* chromosome or the ancestral *MAT* chromosome, are not actually located on the *Z. rouxii* *MAT* chromosome.

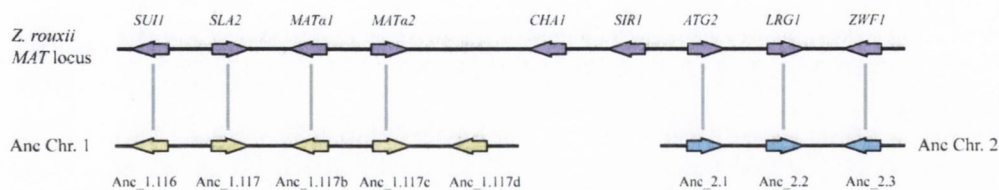


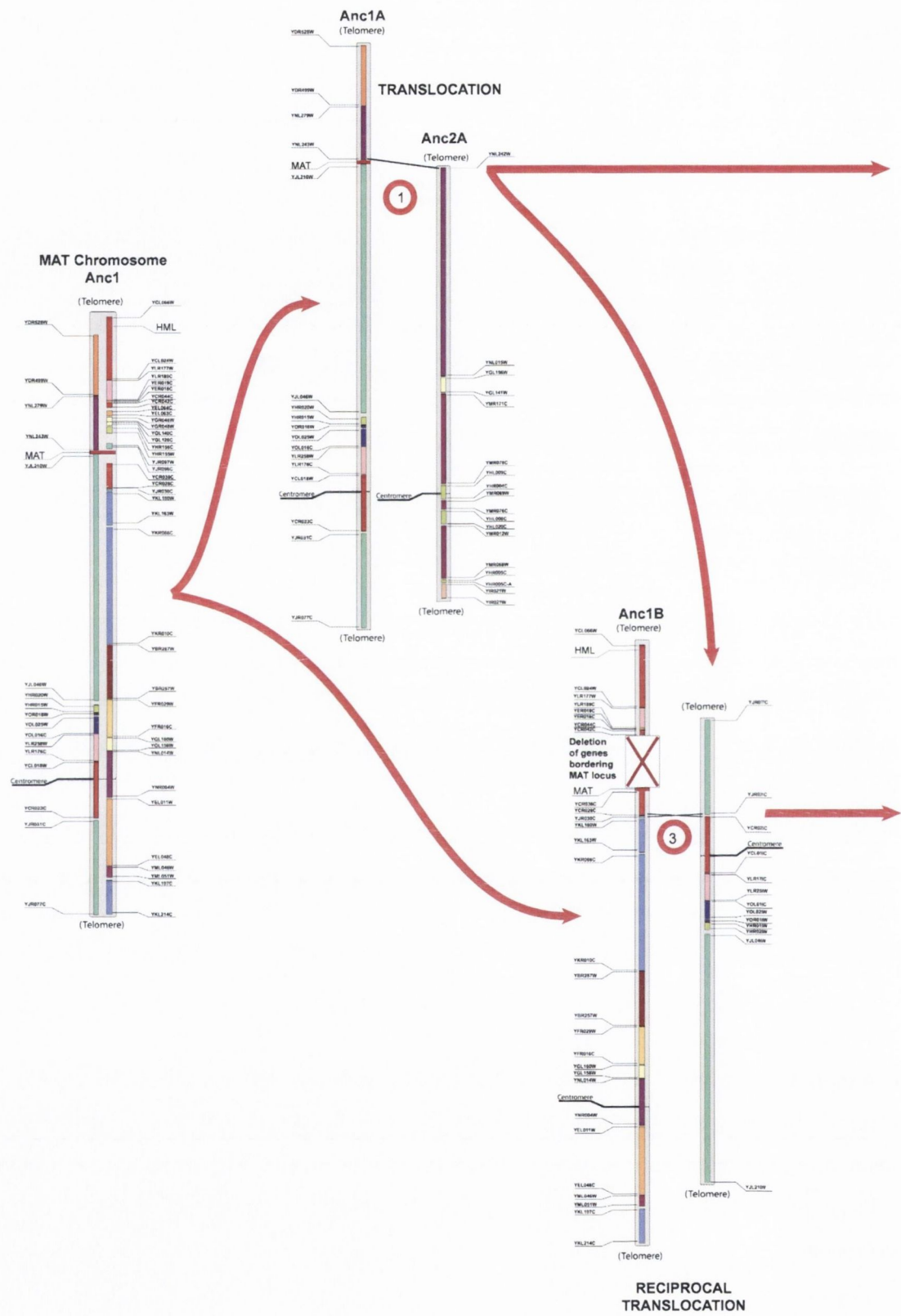
Figure 3.2 Gene order at the *Z. rouxii* *MAT* locus and ancestral orthologous regions. The genes *CHA1* and *SIR1* are unique insertions beside the *MAT* locus in *Z. rouxii*. Anc_1.117d (*DIC1*) is not present at the *Z. rouxii* *MAT* locus in this strain, although it is present at the *MAT* locus of a closely related strain, *Z. rouxii* ATCC 42981 (data not shown).

3.3.2 Gene content in the ancestral *MAT* chromosome

Our reconstruction of the *MAT* chromosome of the immediate ancestor of the post-WGD lineages contains 533 non-subtelomeric genes (Ancestral chromosome 1; Table S2.1 on CD). This number represents a lower limit because future sequencing of more genomes may reveal additional genes that must have been present in the ancestor.

In all of the species examined (Table S2.1 on CD), the *MAT* locus and *HML* are linked on the same chromosome, but homology with the ancestral *MAT* chromosome breaks down in all species on the other side of the *MAT* locus. These breakages occurred independently in each species, and lie at sites ranging from a distance of 10 ancestral genes beyond *MAT* in *C. glabrata* to 118 ancestral genes beyond *MAT* in *K. waltii*. In *S. cerevisiae* the distance is 30 ancestral genes.

We inferred the set of evolutionary rearrangements that assembled the current *S. cerevisiae* chromosome III from the ancestral pre-WGD genome (Figure 3.3). Only seven chromosomal rearrangements are required to achieve this transformation, in addition to numerous gene deletions that occurred without loss of synteny. The rearrangement route consists of one Robertsonian translocation, two reciprocal translocations, and four small inversions.



3.3.3 Erosion and Transposition (EAT) mechanism

To describe the evolution of the regions around the *MAT* locus, we will refer to the two sides of the *MAT* locus as the $\alpha 1$ side and the $\alpha 2$ side, corresponding to the locations of the *MAT* $\alpha 1$ and *MAT* $\alpha 2$ genes in a *MAT* α idiomorph (Figure 3.4). In the ancestral genome, *MAT* $\alpha 1$ is labelled Anc_1.117b and *MAT* $\alpha 2$ is labelled Anc_1.117c. In *S. cerevisiae* the gene *TAF2* is the closest neighboring gene on the $\alpha 1$ side, and *BUD5* is the closest neighbor on the $\alpha 2$ side. In most yeast species the $\alpha 1$ side is the side of the *MAT* locus that is closer to *HML*, but in *S. cerevisiae* a large inversion has turned the *MAT* locus around so that the $\alpha 2$ side is closer to *HML* (Step 4; Figure 3.3). This inversion in *S. cerevisiae* also moved *MAT* to the opposite chromosomal arm from *HML*, something that is not seen in other species.

The ancestral genes bordering the *MAT* locus are *SLA2* on the $\alpha 1$ side, and *DIC1* on the $\alpha 2$ side. Neither of these genes is still found bordering the *MAT* locus in any of the sequenced post-WGD species. Instead, the *MAT* loci of post-WGD species are bordered by genes that used to be further away on the same chromosome but have been "reeled in" towards *MAT* by a process that gradually removed the DNA between *MAT* and these genes. The process has progressed to different extents in different species.

The $\alpha 2$ side of the *MAT* locus was ancestrally flanked by *DIC1* (Anc_1.117d). In *K. polysporus* the $\alpha 2$ -side neighbor of *MAT* is now *RCY1* (Anc_1.124), a gene that was originally 8 ancestral genes away from the *MAT* locus on the $\alpha 2$ side. In *S. castellii*, *S. cerevisiae* and *C. glabrata*, the $\alpha 2$ -side neighbor of *MAT* is now *BUD5* (Anc_1.127), a gene that was previously located 11 ancestral genes further away from *MAT* on the $\alpha 2$ side.

On the $\alpha 1$ side where ancestrally *MAT* was flanked by *SLA2* (Anc_1.117), the scale of deletion has been more dramatic. Each of the four sequenced post-WGD genomes currently has a different gene flanking *MAT* on the $\alpha 1$ side: the genes are *TAF2* (Anc_1.79) in *S. cerevisiae*, *YCR043C* (Anc_1.77) in *S. castellii*, *EMG1* (Anc_1.67) in *C. glabrata*, and

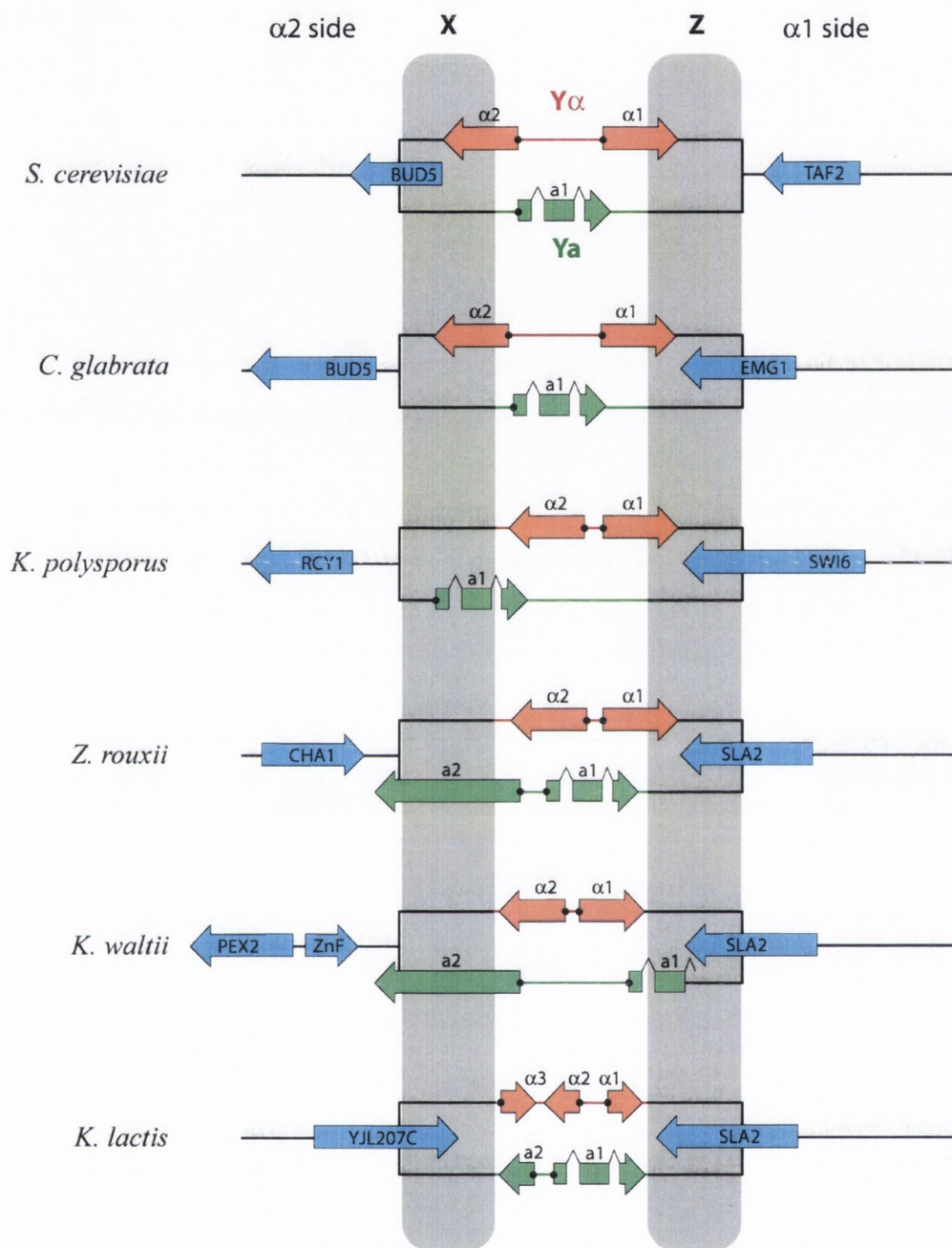


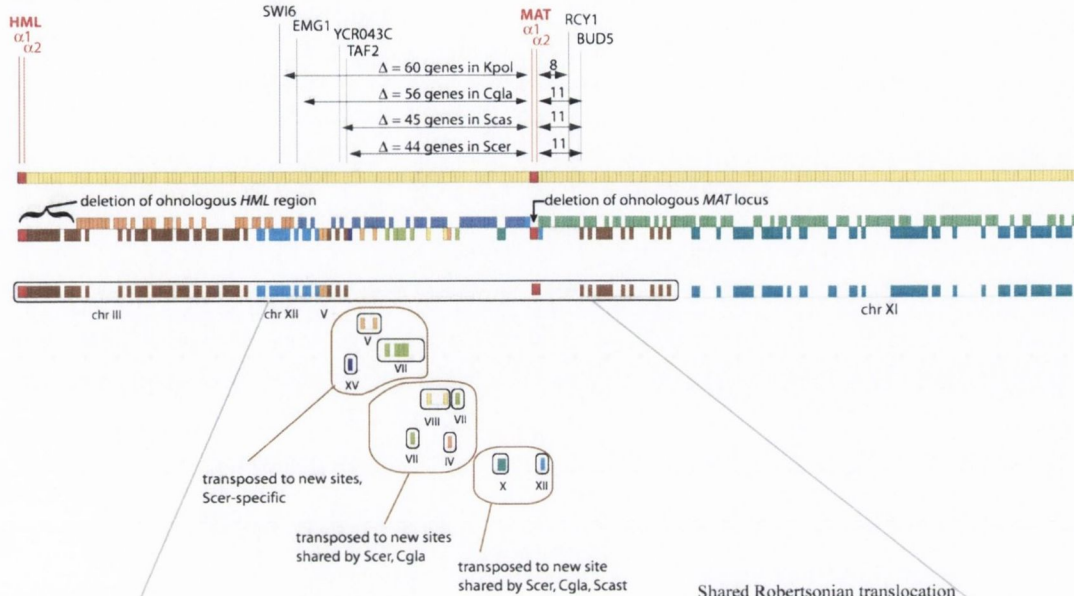
Figure 3.4 The organisation of the mating-type cassettes at the MAT locus in different hemiascomycete species. The X and Z regions (grey bars) are shared between the α - and α -cassettes, while the Y region is unique to each kind of cassette. In many of the species there has been an evident creeping of the cassette boundary at the MAT locus to overlap with neighbouring genes. The overlapping fragments of these genes are also found in the silent α - and α - HM-cassette ideomorphs located in telomeric positions in these species. Figure drawn by Ken Wolfe.

SWI6 (Anc_1.63) in *K. polysporus*. All four of these new neighbors are genes that were previously located further away from *MAT* on its $\alpha 1$ side. This means that the *S. cerevisiae* *MAT* chromosome is missing 44 ancestral genes (Anc_1.80–Anc1.117) that used to be directly adjacent to the *MAT* locus, while *C. glabrata* is missing 56 genes, *S. castellii* is missing 45 genes, and *K. polysporus* is missing 60 genes (Figure 3.5).

Many of the genes that once bordered the *MAT* locus in these post-WGD species were not deleted from their genomes, but were excised and moved. They are currently located elsewhere in the genomes of the post-WGD species, in places where they are not syntenic with the ancestral genome. These genes are often found in blocks of 2-3 adjacent genes at different locations around the genome, indicating multiple events per species rather than a single large rearrangement event per species. In each case, the relocated genes appear as insertions in their new sites, relative to the inferred ancestral genome. Their organization suggests that DNA fragments, of a few genes in size, were removed from beside the *MAT* locus and re-inserted back into the genome at random new sites. We will refer to this process of cutting out genes from beside the *MAT* locus and pasting them back into new genomic locations as Erosion and Transposition (EAT) (Figure 3.5).

All of the genes that underwent EAT were re-integrated into new sites elsewhere in the genome. In some cases, the integration sites are shared among some of the post-WGD species, which gives an indication of the relative timing of EAT events. EAT has occurred on both sides of the *MAT* locus but more prominently on the $\alpha 1$ side. On the $\alpha 2$ side, of the 8-11 genes that were removed from beside *MAT* in post-WGD species, only one (*DIC1*, Anc_1.117d) underwent EAT and became reintegrated into the genome at a new location. *DIC1* is seen as an integration relative to the ancestral arrangement, at the same relative location in *S. cerevisiae*, *C. glabrata* and *S. castellii*, while remaining beside the *MAT* locus in *K. polysporus*. This arrangement implies that the excision and reintegration occurred in the common ancestor of *S. cerevisiae*, *C. glabrata* and *S. castellii*, after the speciation of *K. polysporus*.

Fig. 3.5 A)



B)

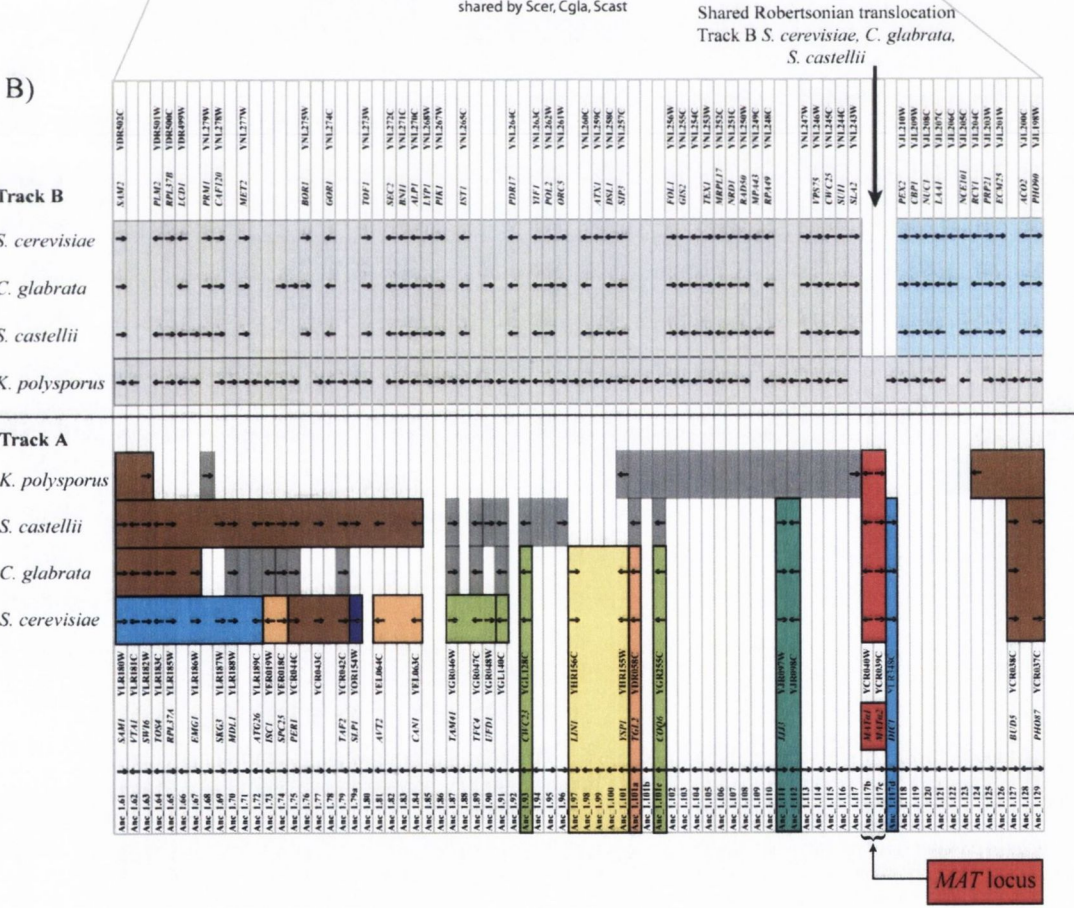


Figure 3.5 Ancestral viewpoint of *MAT* locus and paralogous regions in post-WGD species. A) The ancestral gene order (top yellow row) is used as a scaffold. The paralogous regions in *S. cerevisiae* are shown below the ancestral order. The paralogous *MAT* locus and *HML* silent cassette are missing from the paralogous chromosome. Brown blocks represent genes which are located on the extant *MAT* chromosome in *S. cerevisiae*, while other colours represent genes found on different *S. cerevisiae* chromosomes. The genes bordering the *MAT* locus in different extant post-WGD species are shown above that ancestral order. Genes in between the *MAT* locus and these loci have been deleted or relocated in the different species, some of which represent shared events between species (circled groups of genes). B) Close-up view of both tracks of the ancestral *MAT* locus region relative to all post-WGD species gene orders. The asymmetry of gene loss between Track A and Track B is apparent between the ancestral *MAT* locus location and the ancestral locations of genes that currently border the *MAT* locus. Genes in grey boxes represent unique relocations in species other than *S. cerevisiae*. Shared EAT events are in coloured boxes that cover more than one species. There has been a common Robertsonian translocation at the deleted paralogous *MAT* site in *S. cerevisiae*, *C. glabrata* and *S. castellii* on Track B, while *K. polysporus* retains the ancestral order in this region.

On the $\alpha 1$ side, of the 44-60 ancestral genes that were removed from beside the *MAT* locus in post-WGD species, the ones closest to *MAT* that successfully underwent EAT (i.e., that became re-integrated somewhere else) are Anc_1.112 and Anc_1.111 (*YJR098C* and *YJR097W*; Anc_1.111 and Anc_1.112, Figure 3.5). These two ex-neighbors of *MAT* became relocated to a new site (between Anc_7.464 and Anc_7.466). The structure of this site including the two integrated genes is conserved among *S. cerevisiae*, *C. glabrata* and *S. castellii* (but not *K. polysporus*), indicating that the movement event occurred in the common ancestor of all three species. Moving further away from the ancestral *MAT* in the $\alpha 1$ direction, we then find four blocks of genes that have undergone EAT for which the new insertion sites are shared only by *S. cerevisiae* and *C. glabrata* (Figure 3.5), which indicates that the events happened in a common ancestor of these two species but after *S. castellii* had diverged. Moving even further away from *MAT* in the $\alpha 1$ direction, we find four more EAT events that are unique to *S. cerevisiae* (Figure 3.4, from Anc_1.91 to Anc_1.79a). Likewise, we find numerous events of species-specific EAT in *C. glabrata* and *S. castellii*. All the species-specific events involved genes that were located, in the ancestral genome, further away from *MAT* than the genes that were involved in relocation events shared by multiple species. This pattern of sharing of EAT events among species indicates a time course. Genes closest to the ancestral *MAT* locus were EATen first,

becoming relocated to new sites before the *S. castellii* divergence. Each deletion event placed a new neighboring gene beside *MAT*, and this gene then became vulnerable in turn to EAT.

The EAT phenomenon is not restricted to post-WGD species. In the pre-WGD species the ancestral order and content has generally been maintained. However all of them except *S. kluyveri* are missing *DIC1* from beside their *MAT* loci, and in *K. lactis* and *A. gossypii* more genes bordering the *MAT* locus on the $\alpha 2$ side (which includes *DIC1*) have been deleted or removed. The gene currently bordering the *MAT* locus in *K. lactis* is *YJL207C*, which is 4 ancestral genes from the ancestral *MAT* locus. The rearrangement in *K. lactis* is due to two genes (*DIC1* and *CBP1*) having been deleted completely from the genome, and two linked genes having been moved from beside the *MAT* locus to a new genomic location (*PEX2* and *NUC1*, which are now inserted in a subtelomeric location in *K. lactis*). *A. gossypii* is missing 8 ancestral genes from beside its *MAT* locus, as first noted by Wendland and Walther (2005). Two of these losses (*PEX2* and *NUC1*) from beside the *A. gossypii* *MAT* locus resemble independent EAT events on the ancestral $\alpha 2$ -side. Six losses on the $\alpha 2$ side, and one loss (*SLA2*) on the $\alpha 1$ side, are genes that are completely missing from the *A. gossypii* genome.

3.3.4 Asymmetric pattern of gene deletions between *MAT* and *HML*

Large-scale gene redundancy introduced by the WGD has allowed a large number of gene deletions along both sister tracks, manifesting in the mosaic pattern of the gene content where each of the sister chromosomes contains a subset of the genes found along the same region in a pre-polyploid yeast (Dietrich et al. 2004; Kellis, Birren, and Lander 2004). In most regions of the genome, approximately equal numbers of genes have been lost from the two sister tracks in post-WGD species. In contrast, the gene loss process appears to have been extremely asymmetric in the pairs of sister regions corresponding to the interval between the *MAT* locus and *HML*, specifically in the regions where EATing of blocks of genes occurred adjacent to the ancestral *MAT* locus (Figure 3.5). The most striking difference is in *K. polysporus* where 55 genes have been retained in the sister region that

did not retain a *MAT* locus (Track B in Figure 3.5B), compared to only 3 that underwent EAT adjacent to the *MAT* locus and were derived originally from Track A. Thus Track B from *K. polysporus* in this region retains almost the full set of ancestral genes (Figure 3.5B). Similarly, *S. cerevisiae*, *C. glabrata* and *S. castellii* retain 30, 34 and 28 genes respectively on Track B, but only 14, 15 and 13 genes respectively survive from the Track A region that underwent EAT and became scattered around the genome.

3.3.5 Fate of the Duplicated *MAT* locus

Post-WGD yeast all contain only one copy of the *MAT* locus and each silent cassette. The WGD occurred in a yeast species that had one copy of the *MAT* locus, theoretically creating two copies in the duplicated ancestor. We examined the locations paralogous to the *MAT* locus and silent cassettes (we will refer to these paralogous regions as *exMAT* and *exHML*) by looking at the paralogous sister tracks of the ancestral *MAT* locus and *HML* cassette. Unfortunately, the *HMR* does not have enough shared syntenic context across species to be able to track its ancestral location (in most species other than *S. cerevisiae*, *HMR* is not on the same chromosome as *HML* and *MAT*; (Fabre et al. 2005)). In *S. cerevisiae*, *C. glabrata* and *S. castellii* the *exMAT* is located between the genes *YNL243W* (Anc_1.117) and *YNL242W* (Anc_2.1). This arrangement differs from the ancestral order, and was created by a Robertsonian translocation event whereby the *MAT* locus was broken at Anc_1.118 and rejoined to the telomere of ancestral chromosome 2 at Anc_2.1. In *K. polysporus* both the *MAT* and *exMAT* have maintained the ancestral order.

HML is located at the telomere of ancestral chromosome 1 (Anc_1.1 and Anc_1.2). The *exHML* would be located at the equivalent location on the paralogous track (Track B from Figure 3.5) if it had been retained. Interestingly in all the post-WGD species, several genes from the region beside *exHML* have been deleted. In fact, none of the genes between Anc_1.1 and Anc_1.14 have been retained in any of the post-WGD yeast species. This observation suggests that a large deletion may have removed *exHML* and several of the neighbouring genes soon after WGD as a single event shared by all post-WGD species.

3.3.6 Truncation of genes and loss of intergenic DNA in the X and Z regions

Previous studies have shown that as well as sharing sequences of the α or **a** idiomorph information, the sequence identity between *MAT* and *HM* loci within each species frequently also extends into fragments of the neighboring genes (Jacquet et al. 1991; Oliver et al. 1992; Srikantha, Lachke, and Soll 2003; Wong et al. 2003; Butler et al. 2004). These overlaps occur in the regions usually designated as X (on the α 2 side) and Z (on the α 1 side) (Figure 3.4); these regions are triplicated in the genome and are the recombination substrates used in mating-type switching. (In *S. cerevisiae*, additional regions called W and Z2 have been defined). These lie outside the X and Z regions and correspond to the sections that are duplicated – e.g. between *MAT* and *HML* – but not triplicated in the genome. Since these regions are small and there does not seem to be any consistent pattern among species regarding whether *MAT* shares a longer region of similarity with *HML* or with *HMR*, and on which side, we did not analyze W and Z2 regions further.)

We looked at the overlaps in X and Z regions systematically. We found that on the α 1 side, the *MAT* locus overlaps with *EMG1* in *C. glabrata*, *CAN1* in *S. castellii*, *SWI6* in *K. polysporus*, and *SLA2* in the pre-WGD species *K. waltii*, *K. lactis* and *Z. rouxii*. On the α 2 side, *MAT* overlaps with *BUD5* in *S. cerevisiae*, *RCY1* in *K. polysporus*, and *YJL207C* in the pre-WGD species *K. lactis*. In all cases that we examined where a neighboring gene was partly duplicated as part of the X or Z regions, the full-length copy of the neighboring gene was beside the *MAT* locus and the truncated copy was beside *HML* and/or *HMR*. *S. castellii* contains a portion of the *CAN1* gene at its *MAT* locus, but the pattern of EAT in this species indicates that the ancestral *MAT* chromosome may have been degraded further than the ancestral *CAN1* gene. As such, the *CAN1* fragment may indicate that *CAN1* once overlapped the mating cassette at the *MAT* locus and a fragment of it became part of the mating and silent cassettes which was retained by them after the EAT continued past the *CAN1* gene.

In most cases where neighboring genes extend into the X or Z regions, especially in the post-WGD species, we found that the neighboring gene was shorter than its orthologues in

other species where it does not overlap the cassette (Yellow boxes, Table 3.1). The only non-truncated gene in the X and Z regions in post-WGD species is *RCY1* in *K. polysporus*. The truncated genes have lost their ends closest to the *MAT* locus. This effect appears to be larger on genes that lie on the α 1-side of the *MAT* locus, as is the EAT effect. In some cases the truncation is predicted to have had a dramatic effect on the protein. For example, the *S. cerevisiae* Bud5 protein has lost an SH3 protein-protein interaction domain (Wong et al. 2003), and *Z. rouxii* Sla2 protein has lost a talin (actin binding) domain that is conserved between animals and most other fungi (Yoshiuchi et al. 2006).

Table 3.1 Length of genes beside the *MAT* locus in a species (yellow blocks) and the length of the orthologue in other species.

	<i>BUD5</i>	<i>EMG1</i>	<i>RCY1</i>	<i>SWI6</i>	<i>SLA2</i>	<i>YJL207C</i>
<i>S. cerevisiae</i>	642	252	840	803	968	2014
<i>C. glabrata</i>	1247	229	833	808	952	1981
<i>S. castellii</i>	1330	252	792	751	966	792
<i>K. polysporus</i>	1421	252	801	590	805	2016
<i>Z. rouxii</i>	1434	252	781	789	470	2007
<i>K. waltii</i>	1412	252	776	664	580	378
<i>K. lactis</i>	1360	252	825	741	671	1478
<i>A. gossypii</i>	1436	253	782	700	X	X

Furthermore, we found that in most cases where *MAT*-neighboring genes extend into the X and Z regions, there is essentially no intergenic DNA between the *MAT* locus genes and the flanking genes (Table 3.2). On the α 1 side (Z regions), there are frequently small regions of overlap between the 3' end of the *MAT* locus gene (*MAT* α 1 or *MATa*1) and the 3' end of the neighboring gene.

Table 3.2 Examples of intergenic distances between *MAT* locus genes and neighboring genes that overlap the X or Z regions.

Species	$\alpha 2$ side		$\alpha 1$ side	
	Neighbour	Distance from <i>MAT</i> locus gene to neighbor	Neighbour	Distance from <i>MAT</i> locus gene to neighbor
<i>S. cerevisiae</i>	BUD5	2 bp (to MAT $\alpha 2$)	TAF2	n/a
<i>C. glabrata</i>	<i>BUD5</i>	n/a	EMG1	28 bp overlap between MAT $\alpha 1$ and <i>EMG1</i>
<i>K. polysporus</i>	<i>RCY1</i>	1131 bp	SWI6	2 bp overlap between MAT $\alpha 1$ and <i>SWI6</i>
<i>K. waltii</i>	Kwal homologue of <i>KLLA0D00396g</i>	n/a	SLA2	40 bp overlap between <i>SLA2</i> and MAT $\alpha 1$
<i>Z. rouxii</i>	ZrCHA1m	n/a	SLA2	2 bp overlap between <i>SLA2</i> and MAT $\alpha 1$

n/a, neighboring gene does not overlap the X or Z region.

3.4 Discussion

3.4.1 *Z. rouxii* *MAT* chromosome and *MAT* locus

The partial sequencing of *Z. rouxii* that was performed for this study has revealed a *MAT* chromosome that is remarkably similar to the inferred ancestral *MAT* chromosome in the region between the *MAT* locus and the silent *HML* cassette. This similarity in gene content and order agrees with sequence-based phylogenies which place *Z. rouxii* just outside the WGD clade (Kurtzman and Robnett 2003). *Z. rouxii* represents the extant pre-WGD yeast lineage that is closest to the WGD ancestor.

The *Z. rouxii* *MAT* locus is bordered on its $\alpha 1$ side by *SLA2* (Anc_1.117), which is the ancestral state at that location. On the other ($\alpha 2$) side are homologues of *S. cerevisiae* *SIR1* and *CHA1* genes, which we refer to as *ZrSIR1* and *ZrCHA1m*. Neither of these have syntenic orthologues in other species (data not shown). Next to these is a syntenic block

corresponding to a region on chromosome XIV in *S. cerevisiae* beginning with *ATG2*, which is located at the telomere of ancestral chromosome 2. We infer that the arrangement in *Z. rouxii* was created during a chromosomal translocation event where the ancestral *MAT* chromosome was broken right beside the *MAT* locus on its $\alpha 2$ -side (*PEX2* was lost), and subsequently joined to the telomere of another ancestral type chromosome at *ATG2*. One possible cause of such a rearrangement might have been recombination between the *MAT* locus and a silent cassette at the end of the other chromosome, although no evidence exists for there having been a silent cassette at that particular telomere. Another possible catalyst might have been the HO endonuclease which was gained in hemiascomycetes before the *Z. rouxii* lineage speciation (Butler et al. 2004), but HO causes a DSB on the opposite side of the *MAT* locus ($\alpha 1$ side) than the rearrangement breakpoint occurred on. It has recently been shown that DSBs in DNA caused by the HO endonuclease can capture linear fragments of DNA (Haviv-Chesner et al. 2007), so it is possible that *SIR1* and the *CHAI* homolog were captured as linear fragments during this chromosomal translocation event.

The structural and gene content changes at the *MAT* locus in *Z. rouxii* pose some interesting questions. The role of *SIR1* is in transcriptional silencing, specifically at the silent cassette loci and telomeres (Rine and Herskowitz 1987). The phylogenetic distribution of *SIR1* is similar to the HO endonuclease which was also acquired prior to the divergence of *Z. rouxii* and the WGD ancestor (Butler et al. 2004). It is possible that the location of *SIR1* in *Z. rouxii* at the *MAT* locus might somehow be related to the function of *SIR1* in mating type switching, although it may also be coincidental.

In the comparison of the pair of sister ancestral *MAT* regions created by the WGD, we noticed that the *exMAT* location on the sister tracks in *S. cerevisiae*, *C. glabrata* and *S. castellii* share almost exactly the same gene order as the *Z. rouxii* *MAT* locus (except for the presence of the *CHAI* homologue, *SIR1* and the *MAT* genes themselves). Both the *MAT* and *exMAT* sister regions in *K. polysporus* however retain the ancestral gene order. Thus in the other three post-WGD species it is likely that a chromosomal translocation parallel to the one at the *Z. rouxii* *MAT* locus occurred at some point after the speciation of *K. polysporus*. This parallel translocation is interesting in that it not only has a shared

location for the event, but is in fact precisely the same genomic event occurring twice independently (a Robertsonian translocation between the telomere beside *ATG2* and the spacer between *MAT* α 2 and *PEX2*). It has been noted previously in chemostat experiments in *S. cerevisiae*, that the same locations of interchromosomal translocation breakpoints sometimes occur in parallel in separate replicate cultures (Dunham et al. 2002). Thus a parallel rearrangement in two species (especially a Robertsonian translocation that only creates one breakpoint) may not be as unlikely as it might first seem.

The location of the *CHA1* homologue is also potentially interesting. In the reconstructed ancestral genome, *CHA1* resides beside *HML*. *Z. rouxii* retains this gene, which we refer to as *ZrCHA1h* (the *h* denoting location beside *HML*), as well as having the second copy (*ZrCHA1m*) beside its *MAT* locus. At least two explanations could explain the coincidence of the *CHA1* genes in similar relative orientations beside *MAT* and *HML*. First, it is possible that *ZrCHA1m* ended up at the *MAT* locus purely by chance, perhaps by duplication and reinsertion of the ancestral *CHA1* gene in the *Z. rouxii* lineage. Alternatively, it is possible that the coincidence of the positions of the *CHA1* genes has significance and arose from the same source. It is possible that at some point in the past the boundary on the α 2 side of the *HML* silent cassette slipped to encompass the ancestral *CHA1* gene. A copy might have then been transported to *MAT* through mating type switching. Subsequently, the cassette boundary would have had to slip back to exclude *CHA1m*, as the sequences of *ZrCHA1m* and *ZrCHA1h* are quite divergent. A maximum likelihood phylogenetic tree of the *CHA1* genes is poorly resolved and does not distinguish between these two possible events (data not shown).

3.4.2 Fate of the duplicated *MAT* locus

Immediately after the WGD, all loci in the genome were duplicated, including the *MAT* locus and the silent cassettes. As there is only one copy of each region in the extant post-WGD species, and they all exist in the same orthologous locations, we must infer that the second copy was lost relatively soon after the WGD event. We suggest that shortly after the polyploidy event, the *MAT* locus and silent cassettes contained on the *exMAT*

chromosome were deleted to allow mating type switching and regain fertility (Step 2; Figure 3.3). It has been shown that for low fertility hybrid yeasts created by mating *S. cerevisiae* and other *Saccharomyces sensu stricto* species, fertility can be restored by the deletion of one of the two *MAT* loci (Greig et al. 2002a; Greig et al. 2002b). This creates a cell that is haploid at the *MAT* locus, but diploid at the rest of the genome. Such a diploid cell can divide and undergo mating type-switching. It can then autofertilize to form a fertile tetraploid.

In all species studied, *MAT* and *HML* lie on the same chromosome. This is a possible requirement for mating-type switching. Without a functioning *MAT* locus on the *exMAT* chromosome, and due to the presence of another chromosome with both a *MAT* locus and *HML*, the selective constraint for the retention of the *HML* located on the *exMAT* chromosome would have been relaxed, allowing its deletion. None of the post-WGD species contain any of the genes paralogous to *YCL052C-YCL067C*, indicating that the genes in this region, (*i.e.* genes between *YDR528W* and the telomere of the *exMAT* chromosome) were likely deleted prior to their common ancestor speciating, perhaps in a single event. We hypothesize that a similar event occurred to delete the second version of the *HMR*, but we are unable to track this locus.

3.4.3 Evolution of the *MAT* locus and bordering genes

Our results indicate progressive erosion of the chromosomal region beside the *MAT* locus and re-integration of some genes from this region into random loci elsewhere in the genome. The postulated excision process can account for several otherwise puzzling observations: the gene truncation and loss of intergenic DNA at loci flanking *MAT*; the transposition of groups of genes that ancestrally flanked the *MAT* locus, into new sites apparently chosen at random; and the asymmetry of extents of gene deletion from the A and B tracks near the *MAT* locus in post-WGD species (Figure 3.5).

To explain these observations, we hypothesize that the *MAT* locus represents a hotspot of gene deletion. This is likely due to the process of mating-type switching. The process

appears to be more pronounced in the post-WGD species, which have the HO endonuclease (Butler et al. 2004) than in the pre-WGD species which lack HO (except for *Z. rouxii*). However, we note that the EAT process has nevertheless probably occurred, albeit less dramatically, in several pre-WGD species such as *K. lactis*. However, as the inferred pre-WGD ancestral gene order (just prior to WGD) may not exactly represent the ancestral state between the individual pre-WGD species and the post-WGD lineages, inferences about the evolution of the pre-WGD genomes should be made tentatively.

In most pre-WGD hemiascomycetes, mating-type switching occurs at a low rate (Herman and Roman 1966; Zonneveld and Steensma 2003) and is probably catalyzed by homologous recombination between *MAT* and *HM* loci. We therefore suggest that the increased frequency of mating-type switching which accompanied the gain of HO also resulted in an increased frequency of EAT. A further factor that may have made the extent of gene deletion near *MAT* more dramatic in post-WGD species is the redundancy introduced by the WGD itself. We expect that a hot-spot of deletion will have a far greater effect in a duplicated lineage with a lot of gene redundancy than in a non-polyploid lineage. *Z. rouxii* is the only pre-WGD species sequenced that contains the HO endonuclease, yet it remains similar to the inferred pre-WGD ancestral order and content surrounding the *MAT* locus.

The asymmetry that we observe in the distribution of the retained genes on the two tracks (Figure 3.5) probably reflects the balanced outcome of two processes: after WGD the two regions were identical but genes on Track A were more prone to deletion because of the mutagenic effect of being beside the *MAT* locus. Only if the Track B gene was deleted first was there selection against loss of the Track A gene, necessitating that any deletions of these Track A genes were compensated by their simultaneous re-integration into some other place in the genome - a process that is obviously going to be unlikely and hence rare. Eventually, once the number of duplicates decreases to a stable level in the post-WGD genomes, gene deletion at the *MAT* locus will become relatively unfeasible. It is possible that the EAT process will continue to shorten the distance between *MAT* and *HML*. However it is likely that certain structural requirements of mating type switching, such as

the presence of the Recombination Enhancer (RE), and possible topological constraints on mating type-switching, may eventually cause gene deletions at the *MAT* locus to cease.

In *S. cerevisiae* there has been a pericentromeric inversion of the *MAT* locus relative to the other post-WGD species. This means that the EAT effect in *S. cerevisiae* is no longer occurring in the direction of *HML*, but in the direction of *HMR* (even though the effect is still in the $\alpha 1$ direction). This inversion did not change the gene content or order in the immediate vicinity of the *MAT* locus, and presumably will not stop the process of deletion and relocation. The genes most prone to EAT in *S. cerevisiae* are still its immediate neighbors on the $\alpha 1$ side, such as *TAF2*.

3.4.4 Feedback in the triplicated *MAT-HML-HMR* system

When we examined all of the *MAT* and silent cassettes available in the sequenced species, we found that in nearly all cases the homologous regions between the *MAT* and silent cassettes contained a portion of a gene bordering the *MAT* locus. In all of the post-WGD species, the gene that overlaps the cassettes is different due to differential independent EAT events in each species. Most of the genes that overlap into the cassettes are shorter than orthologues in other species where the gene does not overlap into the cassette. The genes appear to be getting shorter at the end that is overlapping the cassette. This observation combined with the EAT at the *MAT* locus allows us to infer that this process has been slowly eating away the chromosome at the *MAT* locus, absorbing and consuming parts of genes along the way. This effect is especially predominant in the post-WGD species on the $\alpha 1$ -side of the *MAT* locus where many genes have likely undergone such a fate.

Another inference we can make is that information shared between the *MAT* locus and silent cassettes is derived from the *MAT* locus. The genes that overlap the cassettes are all found in the reconstructed ancestral genome close to the *MAT* locus and not near to *HML*. Only the fragment of the gene that overlaps the mating-cassette is transferred to the silent loci. There has been regular slippage of the boundary in the pre-WGD species on the $\alpha 1$ -side of the *MAT* cassettes to include genic sequence from beside the *MAT*, fragments of

which are found in homologous regions of the silent cassettes. We can conclude from this that there must be a feedback mechanism from the *MAT* locus to the silent cassettes which maintains the homology between them that is a requisite part of mating-type switching. There appears to have been no feedback from the silent cassettes, with the possible exception of *CHAI* in *Z. rouxii* which ancestrally lies next to *HML*.

3.5 Methods

Inference of the gene order in the immediate pre-WGD ancestor is described in the previous chapter. We named the chromosomes of the ancestral genome Anc1-Anc8, with Anc1 being the ancestral chromosome containing the *MAT* locus. Genes in the ancestral genome were given names such as Anc_1.10, indicating the 10th gene along chromosome Anc1 (counting from one end). Manual checking of the ancestral genome led to a few genes being added to, or deleted from, the reconstruction after we had assigned these sequential names to loci; there are therefore some small gaps in the numbering system and some non-standard names such as Anc_1.117a. The ancestral genome can be browsed using YGOB (wolfe.gen.tcd.ie/ygob) (Byrne and Wolfe 2005).

DNA isolation, library construction and sequencing from *Z. rouxii* were done commercially by AGOWA (www.agowa.de). A plasmid library of random fragments of genomic DNA (*Sau*3AI partial digestion) from the type strain (CBS 732) of *Z. rouxii* was made. The average insert size is approximately 8 kb. Both ends of approximately 5000 plasmids were sequenced. A set of 4936 further plasmid end-reads from the same strain of *Z. rouxii* from the Genolevures-1 project (de Montigny et al. 2000) was downloaded from <http://cbi.labri.fr/Genolevures/>. All of the sequences were vector trimmed and low quality sequences were removed, giving a total of 14981 high quality reads. Contigs were then assembled using PHRAP (www.phrap.org). *S. cerevisiae* genes were identified on the contigs using BLAST against the Saccharomyces Genome Database (SGD) annotation of the *S. cerevisiae* genome (excluding genes annotated as “hypothetical”, genes with no experimental evidence, genes rejected by Brachat *et. al.* (Brachat et al. 2003), and genes that failed the reading frame conservation test (RFC)) (Kellis et al. 2003; Christie et al.

2004). Using the *MAT* chromosome in *S. cerevisiae* and the *MAT* chromosome in pre-WGD species as scaffolds, we identified plasmids whose end reads were either homologous to genes in these regions, or that overlapped with contigs that contained such genes. Tiling paths of overlapping plasmids were then chosen for complete sequencing by primer walking. Some gaps between plasmids were closed using contigs from the 14981 random reads, or by completely sequencing plasmids obtained from the Genolevures clone set. The plasmid end reads and the large sequenced contigs are on the Supplementary CD (Zr_CBS732_reads.fasta and Zr_CBS732_contigs.fasta).

The *MAT* locus of *K. waltii* was not included in the genome sequence released by Kellis et al. (2004), probably because two idiomorphs were present and could not be assembled into a single contig. We noticed that homologues of *SLA2* and *PEX2* (two genes that lie close to the *MAT* locus on either side in the ancestral genome) were located at the ends of contigs, and found primary sequence data in the NCBI Trace Archive indicating that *MAT*-related sequences lay between these genes. We amplified and sequenced the region between *K. waltii* *SLA2* and *PEX2* by PCR and found two types of sequence corresponding to *MAT* α and *MAT* α idiomorphs. The two *MAT*-related regions identified by Fabre *et al.* (2005) on *K. waltii* contig 123 correspond to *HM* loci.

Genomic Evolution of Polyploid Yeast Species

Chapter 4

The *Zygosaccharomyces rouxii* strain ATCC 42981 is an allopolyploid hybrid between two closely related *Zygosaccharomyces* species

4.1 Abstract

The *Zygosaccharomyces rouxii* strain ATCC 42981 has been reported to have two copies of the *SOD2* and *HOG2* genes, whereas the *Z. rouxii* type strain (CBS 732) only has one. To investigate the structure of its genome we analyzed paired plasmid-end sequences from strain ATCC 42981. We find that this strain contains two copies of the ribosomal rRNA array, one of which has a 26S rRNA gene identical to that of the *Z. rouxii* CBS 732, and the other of which has 26S rRNA identical to a related species, *Z. pseudorouxii*. We find that most regions from *Z. rouxii* CBS 732 map in a one-to-two fashion to regions in ATCC 42981. Complete sequencing of paired regions from ATCC 42981 and comparison to CBS 732 showed no evidence of the gene losses or pseudogene formation that might be expected if ATCC 42981 had undergone whole-genome duplication several million years ago and was in the early stages of gene loss. We conclude that ATCC 42981 is an interspecies hybrid between a *Z. rouxii* strain similar to CBS 732 and a *Z. pseudorouxii* strain, that was formed so recently that its genome has not had time to decay.

4.2 Introduction

Organisms that live in environments containing high concentrations of solutes are termed osmophilic. One such organism is *Zygosaccharomyces rouxii*, a species of hemiascomycete yeast that is used industrially in the production of soy sauce and miso paste (Pribylova, de Montigny, and Sychrova 2007). *Z. rouxii* is capable of growing in high concentrations of salt and sugar as well as at a range of pH values of between 1.8 and

8.0 (Membre, Kubaczka, and Chene 1999). It is also of industrial importance as a causative agent of spoilage in foods, where it is the first species to spoil food with high sugar concentrations (Membre, Kubaczka, and Chene 1999; Pribylova, de Montigny, and Sychrova 2007; Solieri, Cassanelli, and Giudici 2007). One mechanism that allows *Z. rouxii* to survive in high concentration solutions is the production and accumulation of glycerol as a compatible solute, which protects the cell against osmotic lysis (Iwaki, Tamai, and Watanabe 1999; Iwaki et al. 2001). Another key mechanism is the efflux of Na^+ from cells in high concentrations of salt. Several genes involved in these processes have been characterised in *Z. rouxii*. The characterised genes involved in glycerol production and accumulation include *ZrHOG1* (Iwaki, Tamai, and Watanabe 1999), *ZrFPS1* (Tang, Kayingo, and Prior 2005) and *ZrGPD1* (Iwaki et al. 2001). Those involved in Na^+ efflux from the cell are *ZrENA1* (Watanabe et al. 1999) and *ZrSOD2* (Watanabe, Miwa, and Tamai 1995).

Several early papers that published cloned genes from *Z. rouxii* used strain ATCC 49281, which was originally isolated from the miso industry. These papers reported that ATCC 42981 contained two copies of several genes that are single-copy in *S. cerevisiae* and involved in osmotolerance, including *HOG1* (Iwaki, Tamai, and Watanabe 1999), *SOD2* (Iwaki et al. 1998) and *GPD1* (Iwaki et al. 2001). The osmotolerance of *Z. rouxii* was originally attributed to the increased copy number of genes involved in osmotic homeostasis. The Genolevures-1 project subsequently included the type strain of *Z. rouxii* (CBS 732) among the species it surveyed by random plasmid end-sequencing (de Montigny et al. 2000), and rather surprisingly, their data did not confirm any of the reported gene duplications. Kincolva *et al.* (2001) sorted out the confusion by showing that *HOG1* and *SOD2* are both duplicated in ATCC 42981 but not in CBS 732.

Recent studies have found that these two strains differ in terms of some of their morphological, physiological and genomic properties (Pribylova, de Montigny, and Sychrova 2007; Pribylova et al. 2007). The ATCC 42981 strain grows better than the CBS 732 strain in high lithium and sodium salt concentrations, at high sucrose concentration, and it produces more glycerol in both the presence and absence of salt (Pribylova, de

Montigny, and Sychrova 2007). *Z. rouxii* ATCC 42981 is also capable of assimilating and using glycerol as a carbon source, which CBS 732 cannot do. The growth curves of the two strains in rich medium are similar, except that CBS 732 can reach a higher cell density (OD₆₀₀) than ATCC 42981. This implies that CBS 732 can utilise nutrients more efficiently than ATCC 42981. While similar in colony appearance, the CBS 732 cells are smaller and more elongated than ATCC 42981, and also form chains of cells, which ATCC 42981 does not do (Pribylova, de Montigny, and Sychrova 2007). Differences in the properties of the cell walls of these strains have also been discovered, which are exemplified by the requirement of different cell transformation protocols for the two strains (Pribylova and Sychrova 2003; Pribylova et al. 2007). It has been reported that the more osmotolerant ATCC 42981 has a cell wall that is more flexible and elastic than that of CBS 732.

It has previously been suggested, based on the presence of duplicated genes in ATCC 42981 and in conjunction with some karyotypic analyses showing different numbers of chromosomes, that *Z. rouxii* ATCC 42981 is a hybrid between two *Zygosaccharomyces* species (James et al. 2005; Pribylova, de Montigny, and Sychrova 2007). At least three other isolated strains of *Z. rouxii* have previously been identified as hybrids based on the presence of two divergent copies of the *ADE2*, *HIS3* and *SOD2* in each strain (James et al. 2005). We were interested in the possibility that *Z. rouxii* ATCC 42981 might be a species that had undergone WGD by allopolyploidization, and which might therefore provide a way to study the evolution of a WGD event in yeast species that is completely independent of the WGD that occurred in the common ancestor of *S. cerevisiae* and other species in Kurtzman's Clades 1-6. In this Chapter I test whether *Z. rouxii* ATCC 42981 is a hybrid by comparing sequenced regions from the CBS 732 wild-type strain (Chapter 3) to sequences from a random genomic clone library and assembled contigs from the ATCC 42981 strain.

4.3 Results

4.3.1 *Z. rouxii* ATCC 42981 is a recent hybrid between two *Zygosaccharomyces* species

We obtained more than 10,000 random sequencing reads from a plasmid library of *Z. rouxii* ATCC 42981. The contigs assembled from these reads revealed the presence of two different types of rDNA array in this strain. One rDNA array included a 26S rRNA gene identical to that already reported from CBS 732, while the other contains a 26S rRNA sequence identical to that deposited in the databases as *Zygosaccharomyces pseudorouxii*. The latter species has not been formally described and the name was introduced by James *et al.* (2005) to refer to species with this particular 26S gene. The two 26S sequences present in ATCC 42981 are 97% identical to each other, as are the sequences of *Z. rouxii* CBS 732 and *Z. pseudorouxii* to each other.

Many of the assembled contigs in *Z. rouxii* ATCC 42981 are highly similar but not identical to other contigs in the same strain. When compared to contigs assembled from CBS 732 reads, one of the homologous regions in ATCC 42981 in each case exhibits sequence similarity of 95-100%, while the second region generally has 80-90% identity to both the CBS 732 sequence and to its homologous region in ATCC 42981 (Figure 4.1). A similar pattern is seen when we map plasmid end-reads from ATCC 42981 onto the the sequences of the long sequenced regions from *Z. rouxii* CBS 732 in the previous chapter.

We identified and completely sequenced some plasmids from the ATCC 42981 library (in total, 105476 bp) that allowed us to make three-way comparisons between single regions in CBS 732 (CBS Contigs 7, 8 and 10; data from the previous chapter) and pairs of regions from ATCC 42981. Such trios of *Z. rouxii* sequences were collected from three different genomic regions, all of which are homologous to the ancestral *MAT* chromosome, and show the same pattern of sequence identity as the contigs and reads data (Figure 4.2). The Contig7 spanned a 9-gene region from Anc_6.371 (*CDC39*, *ABR112C*) to Anc_6.379 (*ESF2*, *ABR121C*). The Contig8 region spanned a 9-gene region from to Anc_6.350 (*YCR079W*, *AGL223C*) through Anc_6.354 (*ZRG17*, *AGL227C*), followed by Anc_1.142

Table 4.1. Genes in the paired chromosomal regions completely sequenced from strain ATCC 42981.

<i>A. gossypii</i> homologue	Ancestral name	Present in CBS 732?	Present in <i>Z. rouxii</i> -like clones from ATCC 42981?	Present in <i>Z. pseudorouxii</i> -like clones from ATCC 42981?
Contig7:				
ABR112C	Anc_6.371	✓	✓	✓
ADR170C	Anc_6.372	✓	✓	✓
ABR115C	Anc_6.373	✓	✓	✓
ABR116C	Anc_6.374	✓	✓	✓
ABR117C	Anc_6.375	✓	✓	✓
ABR118C	Anc_6.376	✓	✓	✓
ABR119C	Anc_6.377	✓	✓	✓
ABR120C	Anc_6.378	✓	✓	✓
ABR121C	Anc_6.379	✓	✓	✓
Contig8:				
AGL223C	Anc_6.350	✓	✓	✓
AGL224W	Anc_6.351	✓	✓	✓
AGL225C	Anc_6.352	✓	✓	✓
AGL226C	Anc_6.353	✓	✓	✓
AGL227W	Anc_6.354	✓	✓	✓
AEL149C	Anc_1.142	✓	✓	✓
AEL150W	Anc_1.141	✓	✓	✓
AEL151C	Anc_1.140	✓	✓	✓
AEL152W	Anc_1.139	✓	✓	✓
Contig10:				
AAR070C	Anc_6.308	✓	✓	✓
AAR069W	Anc_6.309	✓	✓	✓
AAR068C	Anc_6.310	✓	✓	✓
AAR067W	Anc_6.311	✓	✓	✓
AAR066W	Anc_6.312	✓	✓	✓
AAR065C	Anc_6.313	✓	✓	✓
AAR064W	Anc_6.314	✓	✓	✓
AAR063C	Anc_6.315	✓	✓	✓
AAR062C	Anc_6.316	✓	✓	✓
AAR061W	Anc_6.317	✓	✓	✓
AAR060C	Anc_6.318	✓	✓	✓
AAR059C	Anc_6.319	✓	✓	✓

(*SWE1*, *AEL149C*) through Anc_1.139 (*RPS14A/B*, *AEL152W*); this region includes a breakpoint in both strains of *Z. rouxii* relative to the ancestral *MAT* chromosome. The Contig10 region spanned a 12-gene region from Anc_6.308 (homologous to *S. cerevisiae* *MAS6* and *A. gossypii* *AAR070C*) to Anc_6.319 (homologous to *S. cerevisiae* *THR4* and *A. gossypii* *AAR059C*). None of the genes in these regions showed any evidence of gene structure change (Table 4.1): for all 31 loci strain ATCC 42981 appeared to contain two copies of the gene that were both completely intact, and one of which was almost identical to the sequence from strain CBS 732.

4.3.2 Comparison between *Z. rouxii* ATCC 42981 and other *Zygosaccharomyces* hybrids

SOD2 is the only gene for which there is sequence data for both copies in ATCC 42981 and other hybrid species. The sequence similarity between the *SOD2* genes in the strains analysed is shown in Table 4.2. The *SOD2* genes from *Z. rouxii* CBS 732 and *Z. pseudorouxii* NCYC 3042 were also included as they are likely candidates of both ancestral-type genes in *Z. rouxii* ATCC 42981. Two distinct groups of sequence similarity were observed in the data; those genes that are similar to the *Z. rouxii* CBS 732 *SOD2* gene, and those that are similar to the *Z. pseudorouxii* NCYC 3042 *SOD2* gene. There are consistently high sequence similarities (>99%) between the *Z. rouxii* CBS 732 *SOD2* gene and one gene from each of the hybrid strains, and similarly between *Z. pseudorouxii* *SOD2* and the second gene from the hybrids. The sequence identity between genes within these groups is also consistently high, while the similarity found between any of the members of one of these groups compared to any member from the other is consistently lower (81-85%).

4.4 Discussion

While previous studies that uncovered two copies of certain genes in *Z. rouxii* ATC 42981 may have alluded to the possibility that this strain represents a hybrid of two related species, the existence of duplicate genes mostly involved in osmotolerance in an osmophilic yeast is not necessarily strong evidence for this hypothesis. Likewise,

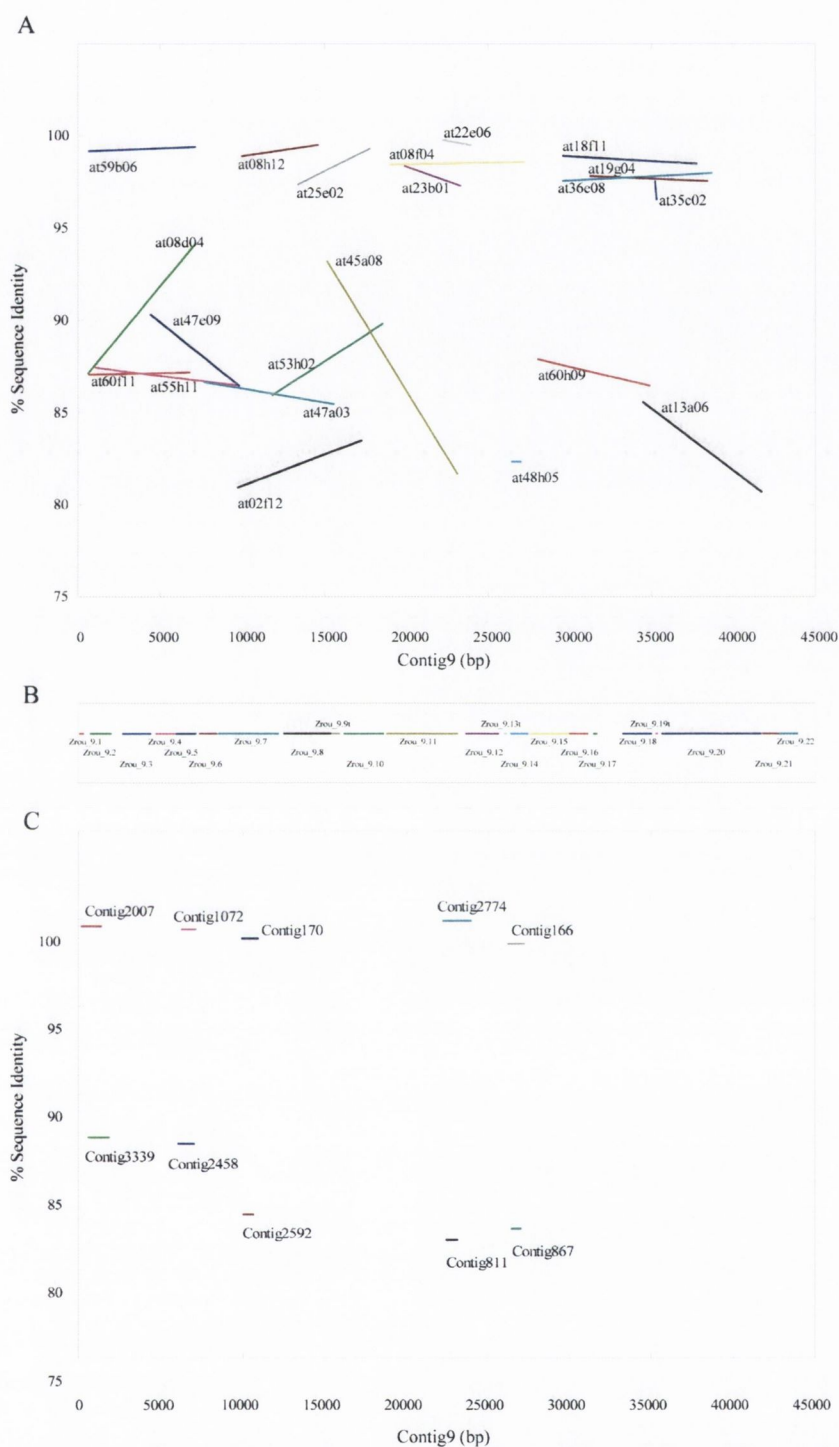


Figure 4.1 Comparison between genomic regions of *Z. rouxii* CBS 732 and *Z. rouxii* ATCC 42981. A) Plot of sequence identity (Y-axis) against position along assembled Contig9 from CBS 732 (X-axis) for random plasmid end reads in *Z. rouxii* ATCC 42981. B) Gene positions along Contig9 are represented by coloured bars (gene names above the bars are tRNA genes). C) Plot of sequence identity against position along Contig9 for *Z. rouxii* ATCC 42981 contigs assembled from the end reads sequence data.

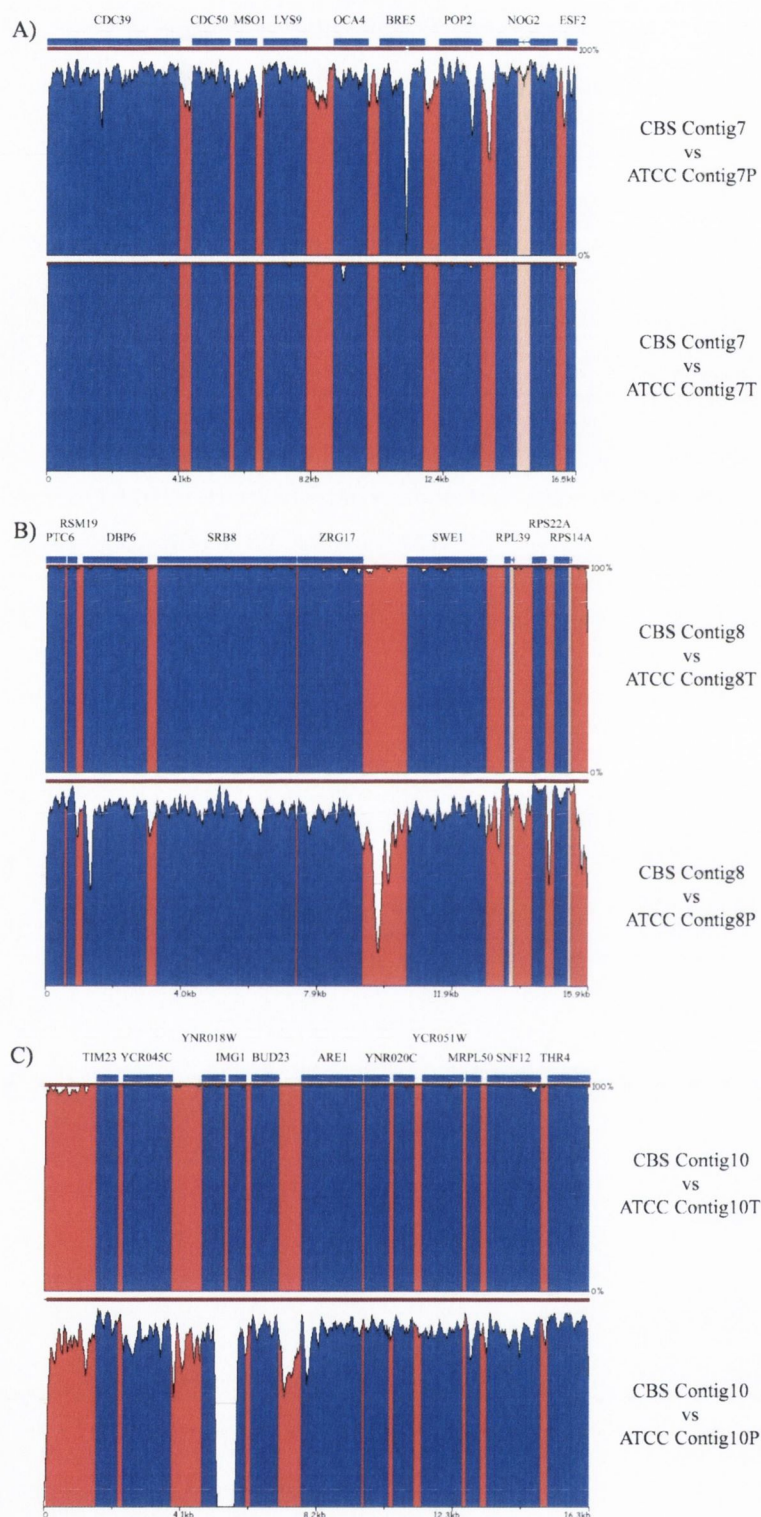


Figure 4.2 Comparison between sequenced trios of genomic regions of *Z. rouxii* CBS 732 and *Z. rouxii* ATCC 42981. A) zPicture plot of sequence identity (Y-axis) against position along assembled Contig7 from CBS 732 (X-axis) against sequenced Contig7P and Contig7T in *Z. rouxii* ATCC 42981. B) Plot of CBS 732 Contig8 against ATCC 42981 Contig8T and Contig8P. C) Plot of CBS 732 Contig10 against ATCC 42981 Contig10T and Contig10P. Gene positions along the contigs are represented by blue bars above each plot, which correspond with the blue regions of the plots. Red regions in the plots represent intergenic regions, while pink regions represent introns. In each case, the ATCC 42981 Contigs are postscripted with a "T" or "P" depending on whether they resemble the type-strain (CBS 732) or *Z. pseudorouxii* genomes.

Table 4.2 Sequence identities (%) for comparisons of the *SOD2* gene duplicates (a and b) between different hybrid *Zygosaccharomyces* species.

	CBS 732	ATCC 42981a	NCYC 1682a	ABT601a	ABT301a	ATCC 42981b	NCYC 1682b	ABT601b	ABT301b	NCYC 3042
CBS 732										
ATCC 42981a	99.87									
NCYC 1682a	99.59	100								
ABT601a	99.45	99.39	99.4							
ABT301a	99.45	99.39	99.4	100						
ATCC 42981b	85.38	85.55	82.78	82.21	82.21					
NCYC 1682b	82.21	82.78	82.78	82.21	82.21	100				
ABT601b	81.76	82.26	82.26	81.76	81.76	99.64	99.63			
ABT301b	82	82.53	82.53	82	82	99.82	99.81	99.82		
NCYC 3042	82	82.78	82.78	82	82	99.81	99.81	99.43	99.62	

chromosome number differences between strains may indeed be a by-product of hybridisation (Pribylova, de Montigny, and Sychrova 2007), but it can also occur by other mechanisms such as aneuploidy or chromosomal fission or fusion. In discovering two 26S rRNA genes that clade with different extant species in the *Zygosaccharomyces* phylogeny (Figure 4.3), as well as other duplicated regions in ATCC 42981 for which one of the regions is almost identical to the CBS 732 sequence, I verify that *Z. rouxii* ATCC 42981 indeed does represent a hybrid between ancestors that are close relatives of *Z. rouxii* and *Z. pseudorouxii*. The finding that there is 100% sequence identity between the 26S rRNAs of two extant species (*Z. rouxii* and *Z. pseudorouxii*) and the different copies in ATCC 42981 indicates that the hybridisation of these lineages must have been a recent evolutionary event, because no sequence divergence has accumulated in the 26S rRNA genes since ATCC 42981 split from its parents.

The comparisons of contigs and plasmid end reads from ATCC 42981 with contigs from CBS 732 show consistently that any region in CBS 732 corresponds to two regions in ATCC 42981, providing further evidence of hybridisation. These sister regions in ATCC 42981 can be partitioned into mutually exclusive sets by their sequence identity to the CBS 732 contigs (Figure 4.1, Figure 4.2). In each case, the CBS 732 contig has a high sequence identity (95-100%) with one ATCC 42981 sequence and a lower identity (80-90%) with the other. In these cases, we infer that the region from ATCC 42981 that is more similar to the CBS 732 sequence was contributed by the *Z. rouxii*-like ancestor, while the more divergent

sequence was contributed by the *Z. pseudorouxii*-like ancestor. It is feasible that every region in the genome of ATCC 42981 could be assigned to its progenitor lineage based solely upon the sequence similarity to either *Z. rouxii* or *Z. pseudorouxii*.

Our analysis of how the random sequence reads from ATCC 42981 map onto the 600 kb region that we sequenced from CBS 732 in the previous chapter suggests that the entirety of this region is present in two copies in ATCC 42981. There is no evidence for any variation in copy number in different regions, as might happen under aneuploidy or if large segmental deletions had occurred.

Several natural *Zygosaccharomyces* hybrids have been observed to date. James *et al.* (2005) reported three hybrids (strains NCYC 1682, NCYC 3060 and NCYC 3061) which possess two divergent copies of the *SOD2*, *HIS3* and *ADE2* genes. Recently it has been demonstrated that two strains (designated *Zygosacchchromyces pseudorouxii* ABT301 and ABT601) isolated from balsamic vinegar have two copies of the *HIS3*, *SOD2* and 5.8S rRNA (Solieri, Cassanelli, and Giudici 2007). A sequence-based phylogeny of these duplicate *SOD2* genes and those from ATCC 42981 shows that each duplicate in the *Z. pseudorouxii* ABT301 and ABT601 strains clades with one of the duplicates in *Z. rouxii* ATCC 42981 (Figure 4.4) (Solieri, Cassanelli, and Giudici 2007). This is also the case for NCYC 1682 and presumably also for the other *Z. rouxii* hybrid strains, although sequence data was not available to perform the analyses. While it is possible that these hybrids were all created by independent hybridisations of similar species, it is more parsimonious that the hybrid strains ABT301, ABT601, NCYC 1682, NCYC 3060 and NCYC 3061 were created by the same hybridisation event and now represent a new species. The NCYC 1682, NCYC 3060 and NCYC 3061 hybrids also each contain two copies of the *HIS3* gene, which clade differentially with *Z. rouxii* CBS 732 and *Z. pseudorouxii* NCYC 3042 *HIS3* genes in a phylogenetic tree (James *et al.* 2005).

Yeast species are often identified by using sequence comparisons of the 26S rRNA D1/D2 region (Kurtzman and Robnett 1998). In the case of hybrid species, it is likely that misidentification will occur when using only single gene sequences, unless two copies of

the 26S rRNA gene are found (James et al. 2005). If only one 26S rRNA gene is sequenced from a hybrid, then sequence-based comparisons will identify the hybrid species as one of its parental ancestors, rather than a separate species. This is most likely the case with the initial identifications of *Z. pseudorouxii* and *Z. rouxii* hybrid strains, including ATCC 42981.

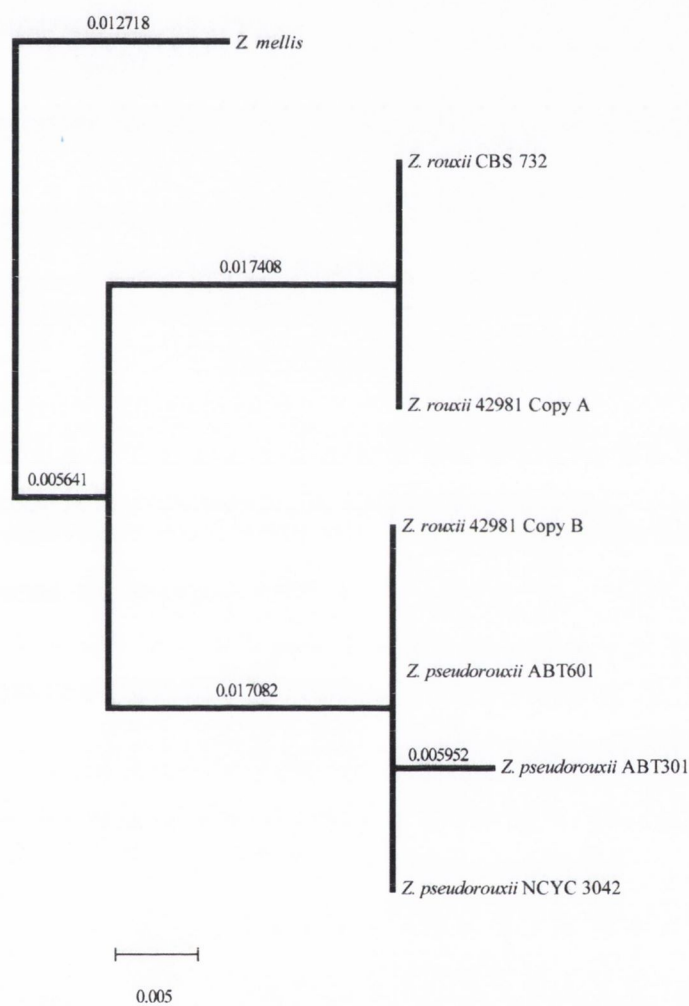


Figure 4.3 Maximum likelihood phylogenetic tree of the 26S rRNA gene sequence in different *Zygosaccharomyces rouxii* and *Zygosaccharomyces pseudorouxii* strains. The *Z. rouxii* ATCC 42981 strain has two 26S rRNAs, one is identical to the *Z. rouxii* CBS 732 sequence, and the other is identical to the *Z. pseudorouxii* NCYC 3042 sequence.

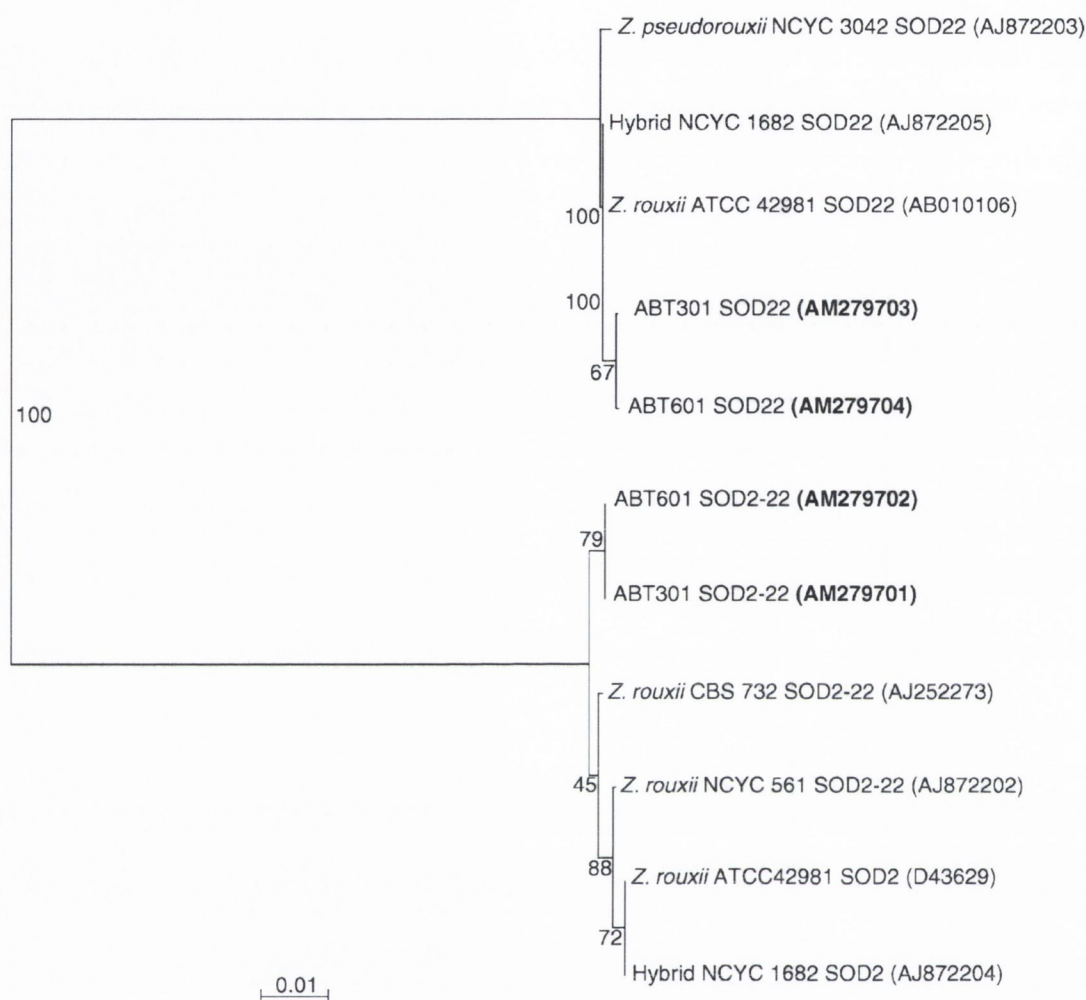


Figure 4.4 Phylogenetic neighbour-joining tree of the *SOD2* gene duplicate sequences from different strains of *Zygosaccharomyces rouxii* and *Zygosaccharomyces pseudorouxii*. In all cases the pair of genes from each strain separate into two distinct clades. Within each clade, the gene sequences are highly similar. Taken from Solieri et al. (2007).

Pribylova *et al.* (2007) documented numerous morphological and physiological differences between *Z. rouxii* CBS 732 and *Z. rouxii* ATCC 42981 which could be the results of the duplicated nature of ATCC 42981. Some of these differences may reflect the different histories of these two strains: ATCC 42981 originates from industrial miso fermentations, whereas CBS 732 was isolated from concentrated grape must (Kurtzman and Robnett

1998). The ATCC 42981 cells are larger than CBS 732 cells, which is expected for a hybrid cell with a larger genome. In terms of physiological differences between these organisms, ATCC 42981 is capable of growing in more extreme conditions than CBS 732, although it grows less efficiently in normal media. The cell wall of ATCC 42981 is more flexible and elastic than that of CBS 732, which may be linked to both the larger size of the cell and its ability to survive in harsher conditions than its parental species (Pribylova, de Montigny, and Sychrova 2007). The data from comparisons between *Z. rouxii* ATCC 42981 and CBS 732 indicate that there is possibly a cost in growth efficiency in exchange for the ability to withstand harsh environments. If this is true, then polyploidisation could be an evolutionary device to effectively deal with large selection pressures during times of environmental instability by doubling all genes, including any which may increase the survival of the organism. In many respects the relationship between *Z. rouxii* strains CBS 732 and ATCC 42981 resembles that between laboratory strains of *S. cerevisiae* and the allopolyploid *S. pastorianus* strains used in the brewing industry, although the latter show evidence of a more complex history than simple allopolyploidy (James et al. 2005).

Although the evolutionary outcomes of WGD in yeasts have been studied extensively using the *S. cerevisiae* model, the molecular origin of this event remains obscure because of the large amount of time that has passed since the *S. cerevisiae* WGD. Our laboratory has reported that the sequence divergence between the two fusing genomes must have been low (Scannell et al. 2007), but the exact nature of the event is still unknown. The attractive model of deletion of one copy of a *MAT* locus in a diploid cell, followed by mating-type switching (Greig et al. 2002a; Greig et al. 2002b; Scannell et al. 2006) remains unconfirmed. We had hoped that if a parallel WGD had occurred in *Z. rouxii*, comparison of the two events would provide insight into common underlying mechanisms. However, the ATCC 42981 strain's WGD seems to have been so recent that its post-WGD evolution lies largely in the future, not the past. Nevertheless, the recent nature of this hybridisation would make ATCC 42981 an ideal organism in which to study the early stages of evolution of a WGD, for example in a chemostat system (Dunham et al. 2002). There is enough sequence identity between the duplicated regions with the parental-type sequences to allow identification of the ancestry of each region in the hybrid. Different morphologies and

tolerance to growth in extreme conditions are apparent in the *Z. rouxii* ATCC 42981 hybrid species when compared to the *Z. rouxii* type-strain, and it may be possible to elucidate the mechanisms behind these differences with further study.

4.5 Methods

Z. rouxii strain ATCC 42981 was purchased from the American Type Culture Collection. A clone library of random fragments (Sau3AI partial digestion) of genomic DNA was constructed by AGOWA (www.agowa.de). The average insert size is approximately 8 kb. Both ends of approximately 5000 plasmids were sequenced by AGOWA. All of the sequences were vector trimmed and low quality sequences were removed, giving a total of 10752 high quality reads. Contigs were then assembled using PHRAP (www.phrap.org). The reads and contigs are included on the Supplementary CD (Zr_ATCC42981_reads.fasta and Zr_ATCC42981_contigs.fasta). Further sequencing was conducted for regions homologous to Contig7, Contig8 and Contig10 from the CBS 732 strain by fully sequencing clones that covered homologous regions in ATCC 42981.

Sequence comparisons between the ATCC 42981 and CBS 732 strains were conducted using the BLASTN program. Plasmid end reads and assembled contigs from ATCC 42981 were matched to contigs from CBS 732 that were assembled in the work for Chapter 4 of this thesis. In the case of the reads, they were chosen only if both ends from the same plasmid had homologous sequence in the CBS 732 contig. The sequence coordinates of each ATCC 42981 plasmid end read and contig that maps to a CBS 732 contig region (X-axis) were plotted against the percentage sequence identity (Y-axis) between the ATCC 42981 sequence and the homologous region of the CBS contig using Gnuplot (<http://www.gnuplot.info/>). Sequence identity comparisons between the larger sequenced homologous regions of CBS 732 and ATCC 42981 were performed with zPicture (<http://zpicture.dcode.org/>).

Two 26S rRNA genes were found in the ATCC 42981 strain by using BLAST against the CBS 732 26S rRNA gene sequence (AY046112.1), and the sequences were obtained by

manually merging overlapping plasmid end reads. A maximum likelihood phylogenetic tree was constructed for these 26S rRNA sequences as well as those from several other *Zygosaccharomyces* species using the PAUP* package (<http://paup.csit.fsu.edu/>). The strains/species and accession numbers for the 26S rRNA sequences used are as follows: *Zygosaccharomyces rouxii* CBS 732 (AY046112.1), *Zygosaccharomyces pseudorouxii* NCYC 3042 (AJ555406.1), *Zygosaccharomyces mellis* CBS 5499 (AJ557254.1), *Zygosaccharomyces pseudorouxii* ABT301 (AJ966342) and *Zygosaccharomyces pseudorouxii* ABT601 (AJ966517).

SOD2 gene sequences from the ABT301 and ABT 601 strains of *Z. pseudorouxii* were obtained from Genbank. The accession numbers for the sequences are as follows: ABT301 (AM279701) and ABT601 (AM279703). Sequence comparisons were performed using BLAST.

References

- Abramova, N., O. Sertil, S. Mehta, and C. V. Lowry. 2001a. Reciprocal regulation of anaerobic and aerobic cell wall mannoprotein gene expression in *Saccharomyces cerevisiae*. *J Bacteriol* **183**:2881-2887.
- Abramova, N. E., B. D. Cohen, O. Sertil, R. Kapoor, K. J. Davies, and C. V. Lowry. 2001b. Regulatory mechanisms controlling expression of the DAN/TIR mannoprotein genes during anaerobic remodeling of the cell wall in *Saccharomyces cerevisiae*. *Genetics* **157**:1169-1177.
- Admire, A., L. Shanks, N. Danzl, M. Wang, U. Weier, W. Stevens, E. Hunt, and T. Weinert. 2006. Cycles of chromosome instability are associated with a fragile site and are increased by defects in DNA replication and checkpoint controls in yeast. *Genes Dev* **20**:159-173.
- Andersson, J. O. 2005. Lateral gene transfer in eukaryotes. *Cell Mol Life Sci* **62**:1182-1197.
- Bachman, N., Y. Eby, and J. D. Boeke. 2004. Local definition of Ty1 target preference by long terminal repeats and clustered tRNA genes. *Genome Res* **14**:1232-1247.
- Batada, N. N., and L. D. Hurst. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet* **39**:945-949.
- Batada, N. N., A. O. Urrutia, and L. D. Hurst. 2007. Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet*.
- Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B* **57**:289-300.
- Blanchard, J. L., and M. Lynch. 2000. Organellar genes: why do they end up in the nucleus? *Trends Genet* **16**:315-320.
- Blanchette, M., T. Kunisawa, and D. Sankoff. 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J Mol Evol* **49**:193-203.
- Blattner, F. R., G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**:1453-1474.
- Bork, P., and A. Bairoch. 1996. Go hunting in sequence databases but watch out for the traps. *Trends Genet* **12**:425-427.
- Bourque, G., G. Tesler, and P. A. Pevzner. 2006. The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction. *Genome Res* **16**:311-313.
- Bowers, J. E., B. A. Chapman, J. Rong, and A. H. Paterson. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**:433-438.
- Brachat, S., F. S. Dietrich, S. Voegeli, Z. Zhang, L. Stuart, A. Lerch, K. Gates, T. Gaffney, and P. Philippsen. 2003. Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biol* **4**:R45.

- Bressan, D. A., J. Vazquez, and J. E. Haber. 2004. Mating type-dependent constraints on the mobility of the left arm of yeast chromosome III. *J Cell Biol* **164**:361-371.
- Burt, D. W., C. Bruley, I. C. Dunn, C. T. Jones, A. Ramage, A. S. Law, D. R. Morrice, I. R. Paton, J. Smith, D. Windsor, A. Sazanov, R. Fries, and D. Waddington. 1999. The dynamics of chromosome evolution in birds and mammals. *Nature* **402**:411-413.
- Butler, G., C. Kenny, A. Fagan, C. Kurischko, C. Gaillardin, and K. H. Wolfe. 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., and K. H. Wolfe. 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* **175**:1341-1350.
- Byrne, K. P., and K. H. Wolfe. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15**:1456-1461.
- Byrnes, J. K., G. P. Morris, and W. H. Li. 2006. Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. *Mol Biol Evol* **23**:1136-1143.
- Canizares, J. V., C. Pallotti, I. Sainz-Pardo, M. Iranzo, and S. Mormeneo. 2002. The SRD2 gene is involved in *Saccharomyces cerevisiae* morphogenesis. *Arch Microbiol* **177**:352-357.
- Casaregola, S., H. V. Nguyen, A. Lepingle, P. Brignon, F. Gendre, and C. Gaillardin. 1998. A family of laboratory strains of *Saccharomyces cerevisiae* carry rearrangements involving chromosomes I and III. *Yeast* **14**:551-564.
- Chen, L., A. L. DeVries, and C. H. Cheng. 1997. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc Natl Acad Sci U S A* **94**:3811-3816.
- Cheng, C. H., and L. Chen. 1999. Evolution of an antifreeze glycoprotein. *Nature* **401**:443-444.
- Chowdhary, B. P., T. Raudsepp, L. Fronicke, and H. Scherthan. 1998. Emerging patterns of comparative genome organization in some mammalian species as revealed by Zoo-FISH. *Genome Res* **8**:577-589.
- Christie, K. R., S. Weng, R. Balakrishnan, M. C. Costanzo, K. Dolinski, S. S. Dwight, S. R. Engel, B. Feierbach, D. G. Fisk, J. E. Hirschman, E. L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C. L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein, and J. M. Cherry. 2004. *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res* **32**:D311-314.
- Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**:71-76.
- Cliften, P. F., L. W. Hillier, L. Fulton, T. Graves, T. Miner, W. R. Gish, R. H. Waterston, and M. Johnston. 2001. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res* **11**:1175-1186.

- Cohen, B. A., R. D. Mitra, J. D. Hughes, and G. M. Church. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* **26**:183-186.
- Colson, I., D. Delneri, and S. G. Oliver. 2004. Effects of reciprocal chromosomal translocations on the fitness of *Saccharomyces cerevisiae*. *EMBO Rep* **5**:392-398.
- Conant, G. C., and A. Wagner. 2005. The rarity of gene shuffling in conserved genes. *Genome Biol* **6**:R50.
- Conant, G. C., and A. Wagner. 2003. Asymmetric sequence divergence of duplicate genes. *Genome Res* **13**:2052-2058.
- Conant, G. C., and K. H. Wolfe. 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol Syst Biol* **3**:129.
- Cooper, T. G. 2002. Transmitting the signal of excess nitrogen in *Saccharomyces cerevisiae* from the Tor proteins to the GATA factors: connecting the dots. *FEMS Microbiol Rev* **26**:223-238.
- Copeland, N. G., N. A. Jenkins, D. J. Gilbert, J. T. Eppig, L. J. Maltais, J. C. Miller, W. F. Dietrich, A. Weaver, S. E. Lincoln, R. G. Steen, and et al. 1993. A genetic linkage map of the mouse: current applications and future prospects. *Science* **262**:57-66.
- Cosma, M. P. 2004. Daughter-specific repression of *Saccharomyces cerevisiae* HO: Ash1 is the commander. *EMBO Rep* **5**:953-957.
- Cusack, B. P., and K. H. Wolfe. 2007. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol Biol Evol* **24**:679-686.
- David, L., W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**:5320-5325.
- de Montigny, J., M. Straub, S. Potier, F. Tekaia, B. Dujon, P. Wincker, F. Artiguenave, and J. Souciet. 2000. Genomic exploration of the hemiascomycetous yeasts: 8. *Zygosaccharomyces rouxii*. *FEBS Lett* **487**:52-55.
- de Rosa, R., and B. Labedan. 1998. The evolutionary relationships between the two bacteria *Escherichia coli* and *Haemophilus influenzae* and their putative last common ancestor. *Mol Biol Evol* **15**:17-27.
- Deak, T., and L. R. Beauchat. 1996. *Handbook of Food Spoilage Yeasts*. CRC Press, Inc., Boca Raton, FL.
- DeBry, R. W., and M. F. Seldin. 1996. Human/mouse homology relationships. *Genomics* **33**:337-351.
- Dehal, P., and J. L. Boore. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**:e314.
- Derr, L. K., and J. N. Strathern. 1993. A role for reverse transcripts in gene conversion. *Nature* **361**:170-173.
- Derr, L. K., J. N. Strathern, and D. J. Garfinkel. 1991. RNA-mediated recombination in *S. cerevisiae*. *Cell* **67**:355-364.
- Despons, L., B. Wirth, V. L. Louis, S. Potier, and J. L. Souciet. 2006. An evolutionary scenario for one of the largest yeast gene families. *Trends Genet* **22**:10-15.
- Devine, S. E., and J. D. Boeke. 1996. Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III. *Genes Dev* **10**:620-633.

- Dietrich, F. S., S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pohlmann, P. Luedi, S. Choi, R. A. Wing, A. Flavier, T. D. Gaffney, and P. Philippsen. 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**:304-307.
- Doolittle, W. F. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* **14**:307-311.
- Dranginis, A. M. 1989. Regulation of STA1 gene expression by MAT during the life cycle of *Saccharomyces cerevisiae*. *Mol Cell Biol* **9**:3992-3998.
- Drgon, T., L. Sabova, N. Nelson, and J. Kolarov. 1991. ADP/ATP translocator is essential only for anaerobic growth of yeast *Saccharomyces cerevisiae*. *FEBS Lett* **289**:159-162.
- Duarte, J. M., L. Cui, P. K. Wall, Q. Zhang, X. Zhang, J. Leebens-Mack, H. Ma, N. Altman, and C. W. dePamphilis. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Mol Biol Evol* **23**:469-478.
- Dujon, B. 2005. Hemiascomycetous yeasts at the forefront of comparative genomics. *Curr Opin Genet Dev* **15**:614-620.
- Dujon, B. 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet* **22**:375-387.
- Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neugeglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J. M. Beckerich, E. Beyne, C. Bleykasten, A. Boisrame, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J. M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G. F. Richard, M. L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wesolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, and J. L. Souciet. 2004. Genome evolution in yeasts. *Nature* **430**:35-44.
- Dunham, M. J., H. Badrane, T. Ferea, J. Adams, P. O. Brown, F. Rosenzweig, and D. Botstein. 2002. Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* **99**:16144-16149.
- Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**:14863-14868.
- Fabre, E., H. Muller, P. Therizols, I. Lafontaine, B. Dujon, and C. Fairhead. 2005. Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. *Mol Biol Evol* **22**:856-873.
- Fischer, G., S. A. James, I. N. Roberts, S. G. Oliver, and E. J. Louis. 2000. Chromosomal evolution in *Saccharomyces*. *Nature* **405**:451-454.
- Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**:496-512.

- Fogel, S., and J. W. Welch. 1982. Tandem gene amplification mediates copper resistance in yeast. *Proc Natl Acad Sci U S A* **79**:5342-5346.
- Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**:1531-1545.
- Froenicke, L. 2005. Origins of primate chromosomes - as delineated by Zoo-FISH and alignments of human and mouse draft genome sequences. *Cytogenet Genome Res* **108**:122-138.
- Froenicke, L., M. G. Caldes, A. Graphodatsky, S. Muller, L. A. Lyons, T. J. Robinson, M. Volleth, F. Yang, and J. Wienberg. 2006. Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes? *Genome Res* **16**:306-310.
- Futcher, A. B. 1988. The 2 micron circle plasmid of *Saccharomyces cerevisiae*. *Yeast* **4**:27-40.
- Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. 1996. Life with 6000 genes. *Science* **274**:546, 563-547.
- Gojkovic, Z., W. Knecht, E. Zameitat, J. Warneboldt, J. B. Coutelis, Y. Pynyaha, C. Neuveglise, K. Moller, M. Loffler, and J. Piskur. 2004. Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts. *Mol Genet Genomics* **271**:387-393.
- Goutte, C., and A. D. Johnson. 1994. Recognition of a DNA operator by a dimer composed of two different homeodomain proteins. *Embo J* **13**:1434-1442.
- Goutte, C., and A. D. Johnson. 1993. Yeast $\alpha 1$ and $\alpha 2$ homeodomain proteins form a DNA-binding activity with properties distinct from those of either protein. *J Mol Biol* **233**:359-371.
- Greig, D., R. H. Borts, E. J. Louis, and M. Travisano. 2002a. Epistasis and hybrid sterility in *Saccharomyces*. *Proc Biol Sci* **269**:1167-1171.
- Greig, D., E. J. Louis, R. H. Borts, and M. Travisano. 2002b. Hybrid speciation in experimental populations of yeast. *Science* **298**:1773-1775.
- Haber, J. E. 1998. Mating-type gene switching in *Saccharomyces cerevisiae*. *Annu Rev Genet* **32**:561-599.
- Haber, J. E. 2002. Switching of *Saccharomyces cerevisiae* mating-type genes. Pp. 927-952 in N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. *Mobile DNA II*. ASM Press, Washington, D.C.
- Haber, J. E., G. Ira, A. Malkova, and N. Sugawara. 2004. Repairing a double-strand chromosome break by homologous recombination: revisiting Robin Holliday's model. *Philos Trans R Soc Lond B Biol Sci* **359**:79-86.
- Haber, J. E., Wolfe K. H. 2005. *Function and Evolution of HO and VDE Endonucleases in Fungi*. Springer-Verlag Berlin Heidelberg.
- Hagen, D. C., L. Bruhn, C. A. Westby, and G. F. Sprague, Jr. 1993. Transcription of α -specific genes in *Saccharomyces cerevisiae*: DNA sequence requirements for activity of the coregulator $\alpha 1$. *Mol Cell Biol* **13**:6866-6875.
- Haldane, J. 1933. The part played by recurrent mutation in evolution. *The American Naturalist* **67**:5-19.

- Hall, C., S. Brachat, and F. S. Dietrich. 2005. Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot Cell* **4**:1102-1115.
- Haviv-Chesner, A., Y. Kobayashi, A. Gabriel, and M. Kupiec. 2007. Capture of linear fragments at a double-strand break in yeast. *Nucleic Acids Res.*
- He, X., and J. Zhang. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**:1157-1164.
- Herman, A., and H. Roman. 1966. Allele specific determinants of homothallism in *Saccharomyces lactis*. *Genetics* **53**:727-740.
- Herskowitz, I. 1988. Life cycle of the budding yeast *Saccharomyces cerevisiae*. *Microbiol Rev* **52**:536-553.
- Hess, S. M., D. R. Stanford, and A. K. Hopper. 1994. SRD1, a *S. cerevisiae* gene affecting pre-rRNA processing contains a C2/C2 zinc finger motif. *Nucleic Acids Res* **22**:1265-1271.
- Hicks, J., B., Strathern, J., N., Herskowitz, I. 1977. The Cassette Model of Mating-type Interconversion. Cold Spring Harbor Laboratory, NY.
- Hicks, J., J. N. Strathern, and A. J. Klar. 1979. Transposable mating type genes in *Saccharomyces cerevisiae*. *Nature* **282**:478-473.
- Hicks, J. B., J. N. Strathern, and I. Herskowitz. 1977. Interconversion of Yeast Mating Types III. Action of the Homothallism (HO) Gene in Cells Homozygous for the Mating Type Locus. *Genetics* **85**:395-405.
- Hittinger, C. T., A. Rokas, and S. B. Carroll. 2004. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci U S A* **101**:14144-14149.
- Hohmann, S., and P. A. Meacock. 1998. Thiamin metabolism and thiamin diphosphate-dependent enzymes in the yeast *Saccharomyces cerevisiae*: genetic regulation. *Biochim Biophys Acta* **1385**:201-219.
- Hotopp, J. C., M. E. Clark, D. C. Oliveira, J. M. Foster, P. Fischer, M. C. Torres, J. D. Giebel, N. Kumar, N. Ishmael, S. Wang, J. Ingram, R. V. Nene, J. Shepard, J. Tomkins, S. Richards, D. J. Spiro, E. Ghedin, B. E. Slatko, H. Tettelin, and J. H. Werren. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**:1753-1756.
- Hughes, A. L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**:119-124.
- Hughes, T. R., C. J. Roberts, H. Dai, A. R. Jones, M. R. Meyer, D. Slade, J. Burchard, S. Dow, T. R. Ward, M. J. Kidd, S. H. Friend, and M. J. Marton. 2000. Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat Genet* **25**:333-337.
- Hurst, L. D., E. J. Williams, and C. Pal. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet* **18**:604-606.
- Iwaki, T., Y. Higashida, H. Tsuji, Y. Tamai, and Y. Watanabe. 1998. Characterization of a second gene (ZSOD22) of Na⁺/H⁺ antiporter from salt-tolerant yeast *Zygosaccharomyces rouxii* and functional expression of ZSOD2 and ZSOD22 in *Saccharomyces cerevisiae*. *Yeast* **14**:1167-1174.
- Iwaki, T., S. Kurono, Y. Yokose, K. Kubota, Y. Tamai, and Y. Watanabe. 2001. Cloning of glycerol-3-phosphate dehydrogenase genes (ZrGPD1 and ZrGPD2) and glycerol

- dehydrogenase genes (ZrGCY1 and ZrGCY2) from the salt-tolerant yeast *Zygosaccharomyces rouxii*. *Yeast* **18**:737-744.
- Iwaki, T., Y. Tamai, and Y. Watanabe. 1999. Two putative MAP kinase genes, ZrHOG1 and ZrHOG2, cloned from the salt-tolerant yeast *Zygosaccharomyces rouxii* are functionally homologous to the *Saccharomyces cerevisiae* HOG1 gene. *Microbiology* **145** (Pt 1):241-248.
- Jacquet, M., J. M. Buhler, F. Iborra, M. C. Francingues-Gaillard, and C. Soustelle. 1991. The MAT locus revisited within a 9.8 kb fragment of chromosome III containing BUD5 and two new open reading frames. *Yeast* **7**:881-888.
- Jaillon, O., J. M. Aury, F. Brunet, J. L. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot, S. Nicaud, D. Jaffe, S. Fisher, G. Lutfalla, C. Dossat, B. Segurens, C. Dasilva, M. Salanoubat, M. Levy, N. Boudet, S. Castellano, V. Anthouard, C. Jubin, V. Castelli, M. Katinka, B. Vacherie, C. Biemont, Z. Skalli, L. Cattolico, J. Poulain, V. De Berardinis, C. Cruaud, S. Duprat, P. Brottier, J. P. Coutanceau, J. Gouzy, G. Parra, G. Lardier, C. Chapple, K. J. McKernan, P. McEwan, S. Bosak, M. Kellis, J. N. Volff, R. Guigo, M. C. Zody, J. Mesirov, K. Lindblad-Toh, B. Birren, C. Nusbaum, D. Kahn, M. Robinson-Rechavi, V. Laudet, V. Schachter, F. Quetier, W. Saurin, C. Scarpelli, P. Wincker, E. S. Lander, J. Weissenbach, and H. Roest Crolius. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**:946-957.
- James, S. A., C. J. Bond, M. Stratford, and I. N. Roberts. 2005. Molecular evidence for the existence of natural hybrids in the genus *Zygosaccharomyces*. *FEMS Yeast Res* **5**:747-755.
- Jardine, O., J. Gough, C. Chothia, and S. A. Teichmann. 2002. Comparison of the small molecule metabolic enzymes of *Escherichia coli* and *Saccharomyces cerevisiae*. *Genome Res* **12**:916-929.
- Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet* **22**:225-231.
- Jones, C. D., and D. J. Begun. 2005. Parallel evolution of chimeric fusion genes. *Proc Natl Acad Sci U S A* **102**:11373-11378.
- Jungwirth, H., and K. Kuchler. 2006. Yeast ABC transporters-- a tale of sex, stress, drugs and aging. *FEBS Lett* **580**:1131-1138.
- Karin, M., R. Najarian, A. Haslinger, P. Valenzuela, J. Welch, and S. Fogel. 1984. Primary structure and transcription of an amplified genetic locus: the CUP1 locus of yeast. *Proc Natl Acad Sci U S A* **81**:337-341.
- Kasahara, M. 2007. The 2R hypothesis: an update. *Curr Opin Immunol*.
- Katinka, M. D., S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretailade, P. Brottier, P. Wincker, F. Delbac, H. El Alaoui, P. Peyret, W. Saurin, M. Gouy, J. Weissenbach, and C. P. Vivares. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**:450-453.
- Keeling, P. J. 2004. Reduction and compaction in the genome of the apicomplexan parasite *Cryptosporidium parvum*. *Dev Cell* **6**:614-616.

- Kellis, M., B. W. Birren, and E. S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**:617-624.
- Kellis, M., N. Patterson, B. Birren, B. Berger, and E. S. Lander. 2004. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J Comput Biol* **11**:319-355.
- Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**:241-254.
- Keogh, R. S., C. Seoighe, and K. H. Wolfe. 1998. Evolution of gene order and chromosome number in *Saccharomyces*, *Kluyveromyces* and related fungi. *Yeast* **14**:443-457.
- Kim, J. M., S. Vanguri, J. D. Boeke, A. Gabriel, and D. F. Voytas. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res* **8**:464-478.
- Kinclova, O., S. Potier, and H. Sychrova. 2001. The *Zygosaccharomyces rouxii* strain CBS732 contains only one copy of the HOG1 and the SOD2 genes. *J Biotechnol* **88**:151-158.
- Kohn, M., J. Hogel, W. Vogel, P. Minich, H. Kehrer-Sawatzki, J. A. Graves, and H. Hameister. 2006. Reconstruction of a 450-My-old ancestral vertebrate protokaryotype. *Trends Genet* **22**:203-210.
- Kondrashov, F. A., and A. S. Kondrashov. 2006. Role of selection in fixation of gene duplications. *J Theor Biol* **239**:141-151.
- Koonin, E. V., R. L. Tatusov, and K. E. Rudd. 1995. Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. *Proc Natl Acad Sci U S A* **92**:11921-11925.
- Koszul, R., S. Caburet, B. Dujon, and G. Fischer. 2004. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *Embo J* **23**:234-243.
- Krylov, D. M., Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* **13**:2229-2235.
- Kunz, B. A., G. R. Taylor, and R. H. Haynes. 1985. Mating-type switching in yeast is induced by thymine nucleotide depletion. *Mol Gen Genet* **199**:540-542.
- Kurtzman, C. P., and C. J. Robnett. 2003. Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Res* **3**:417-432.
- Kurtzman, C. P., and C. J. Robnett. 1998. Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. *Antonie Van Leeuwenhoek* **73**:331-371.
- Kwast, K. E., L. C. Lai, N. Menda, D. T. James, 3rd, S. Aref, and P. V. Burke. 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**:250-265.

- Labib, K., and B. Hodgson. 2007. Replication fork barriers: pausing for a break or stalling for time? *EMBO Rep* **8**:346-353.
- Lai, L. C., A. L. Kosorukoff, P. V. Burke, and K. E. Kwast. 2005. Dynamical remodeling of the transcriptome during short-term anaerobiosis in *Saccharomyces cerevisiae*: differential response and role of Msn2 and/or Msn4 and other factors in galactose and glucose media. *Mol Cell Biol* **25**:4075-4091.
- Larkin, D. M., A. Everts-van der Wind, M. Rebeiz, P. A. Schweitzer, S. Bachman, C. Green, C. L. Wright, E. J. Campos, L. D. Benson, J. Edwards, L. Liu, K. Osoegawa, J. E. Womack, P. J. de Jong, and H. A. Lewin. 2003. A cattle-human comparative map built with cattle BAC-ends and human genome sequence. *Genome Res* **13**:1966-1972.
- Lee, H. L., R. K. Jansen, T. W. Chumley, and K. J. Kim. 2007. Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Mol Biol Evol* **24**:1161-1180.
- Lee, J. M., and E. L. Sonnhammer. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* **13**:875-882.
- Leh-Louis, V., B. Wirth, S. Potier, J. L. Souciet, and L. Despons. 2004. Expansion and contraction of the DUP240 multigene family in *Saccharomyces cerevisiae* populations. *Genetics* **167**:1611-1619.
- Lemoine, F. J., N. P. Degtyareva, K. Lobachev, and T. D. Petes. 2005. Chromosomal translocations in yeast induced by low levels of DNA polymerase α model for chromosome fragile sites. *Cell* **120**:587-598.
- Leppert, G., R. McDevitt, S. C. Falco, T. K. Van Dyk, M. B. Ficke, and J. Golin. 1990. Cloning by gene amplification of two loci conferring multiple drug resistance in *Saccharomyces*. *Genetics* **125**:13-20.
- Lercher, M. J., T. Blumenthal, and L. D. Hurst. 2003. Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. *Genome Res* **13**:238-243.
- Levine, M. T., C. D. Jones, A. D. Kern, H. A. Lindfors, and D. J. Begun. 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**:9935-9939.
- Li, W. H., J. Yang, and X. Gu. 2005. Expression divergence between duplicate genes. *Trends Genet* **21**:602-607.
- Llorrente, B., P. Durrens, A. Malpertuy, M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara, E. Bon, P. Brottier, S. Casaregola, B. Dujon, J. de Montigny, A. Lepingle, C. Neugeglise, O. Ozier-Kalogeropoulos, S. Potier, W. Saurin, F. Tekaia, C. Toffano-Nioche, M. Wesolowski-Louvel, P. Wincker, J. Weissenbach, J. Souciet, and C. Gaillardin. 2000. Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*. *FEBS Lett* **487**:122-133.
- Llorrente, B., C. Fairhead, and B. Dujon. 1999. Genetic redundancy and gene fusion in the genome of the Baker's yeast *Saccharomyces cerevisiae*: functional characterization of a three-member gene family involved in the thiamine biosynthetic pathway. *Mol Microbiol* **32**:1140-1152.

- Long, M., E. Betran, K. Thornton, and W. Wang. 2003a. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**:865-875.
- Long, M., M. Deutsch, W. Wang, E. Betran, F. G. Brunet, and J. Zhang. 2003b. Origin of new genes: evidence from experimental and computational analyses. *Genetica* **118**:171-182.
- Lopez, P. J., and B. Seraphin. 2000. YIDB: the Yeast Intron DataBase. *Nucleic Acids Res* **28**:85-86.
- Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**:1151-1155.
- Lynch, M., and A. Force. 2000a. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**:459-473.
- Lynch, M., and A. G. Force. 2000b. The origin of intraspecific genomic incompatibility via gene duplication. *American Naturalist* **156**:590-605.
- Lynch, M., and V. Katju. 2004. The altered evolutionary trajectories of gene duplicates. *Trends Genet* **20**:544-549.
- Ma, J., L. Zhang, B. B. Suh, B. J. Raney, R. C. Burhans, W. J. Kent, M. Blanchette, D. Haussler, and W. Miller. 2006. Reconstructing contiguous regions of an ancestral genome. *Genome Res* **16**:1557-1565.
- Membre, J. M., M. Kubaczka, and C. Chene. 1999. Combined effects of pH and sugar on growth rate of *Zygosaccharomyces rouxii*, a bakery product spoilage yeast. *Appl Environ Microbiol* **65**:4921-4925.
- Merico, A., P. Sulo, J. Piskur, and C. Compagno. 2007. Fermentative lifestyle in yeasts belonging to the *Saccharomyces* complex. *Febs J* **274**:976-989.
- Metzenberg, R. L., and N. L. Glass. 1990. Mating type and mating strategies in *Neurospora*. *Bioessays* **12**:53-59.
- Mojzita, D., and S. Hohmann. 2006. Pdc2 coordinates expression of the THI regulon in the yeast *Saccharomyces cerevisiae*. *Mol Genet Genomics* **276**:147-161.
- Morse, R. H., S. Y. Roth, and R. T. Simpson. 1992. A transcriptionally active tRNA gene interferes with nucleosome positioning in vivo. *Mol Cell Biol* **12**:4015-4025.
- Muller, S., R. Stanyon, P. C. O'Brien, M. A. Ferguson-Smith, R. Plesker, and J. Wienberg. 1999. Defining the ancestral karyotype of all primates by multidirectional chromosome painting between tree shrews, lemurs and humans. *Chromosoma* **108**:393-400.
- Murata, Y., T. Homma, E. Kitagawa, Y. Momose, M. S. Sato, M. Odani, H. Shimizu, M. Hasegawa-Mizusawa, R. Matsumoto, S. Mizukami, K. Fujita, M. Parveen, Y. Komatsu, and H. Iwahashi. 2006. Genome-wide expression analysis of yeast response during exposure to 4 degrees C. *Extremophiles* **10**:117-128.
- Murphy, W. J., D. M. Larkin, A. Everts-van der Wind, G. Bourque, G. Tesler, L. Auvil, J. E. Beever, B. P. Chowdhary, F. Galibert, L. Gatzke, C. Hitte, S. N. Meyers, D. Milan, E. A. Ostrander, G. Pape, H. G. Parker, T. Raudsepp, M. B. Rogatcheva, L. B. Schook, L. C. Skow, M. Welge, J. E. Womack, J. O'Brien S, P. A. Pevzner, and H. A. Lewin. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**:613-617.
- Murphy, W. J., R. Stanyon, and S. J. O'Brien. 2001. Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol* **2**:REVIEWS0005.

- Mushegian, A. R., and E. V. Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* **93**:10268-10273.
- Nadeau, J. H., and B. A. Taylor. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci U S A* **81**:814-818.
- Nakatani, Y., H. Takeda, Y. Kohara, and S. Morishita. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**:1254-1265.
- Nash, R., S. Weng, B. Hitz, R. Balakrishnan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, M. S. Livstone, R. Oughtred, J. Park, M. Skrzypek, C. L. Theesfeld, G. Binkley, Q. Dong, C. Lane, S. Miyasato, A. Sethuraman, M. Schroeder, K. Dolinski, D. Botstein, and J. M. Cherry. 2007. Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res* **35**:D468-471.
- Nelissen, B., R. De Wachter, and A. Goffeau. 1997. Classification of all putative permeases and other membrane plurispanners of the major facilitator superfamily encoded by the complete genome of *Saccharomyces cerevisiae*. *FEMS Microbiol Rev* **21**:113-134.
- Nickoloff, J. A., E. Y. Chen, and F. Heffron. 1986. A 24-base-pair DNA sequence from the MAT locus stimulates intergenic recombination in yeast. *Proc Natl Acad Sci U S A* **83**:7831-7835.
- Nieduszynski, C. A., S. Hiraga, P. Ak, C. J. Benham, and A. D. Donaldson. 2007. OriDB: a DNA replication origin database. *Nucleic Acids Res* **35**:D40-46.
- Nosaka, K., M. Onozuka, H. Konno, Y. Kawasaki, H. Nishimura, M. Sano, and K. Akaji. 2005. Genetic regulation mediated by thiamin pyrophosphate-binding motif in *Saccharomyces cerevisiae*. *Mol Microbiol* **58**:467-479.
- Nourani, A., M. Wesolowski-Louvel, T. Delaveau, C. Jacq, and A. Delahodde. 1997. Multiple-drug-resistance phenomenon in the yeast *Saccharomyces cerevisiae*: involvement of two hexose transporters. *Mol Cell Biol* **17**:5453-5460.
- Ohno, S. 1973. Ancient linkage groups and frozen accidents. *Nature* **244**:259-262.
- Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, Berlin.
- Oliver, S. G., Q. J. van der Aart, M. L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J. P. Ballesta, P. Benit, and et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357**:38-46.
- Onishi, H. 1963. Osmophilic Yeasts. *Adv Food Res* **12**:53-94.
- Pal, C., and L. D. Hurst. 2003. Evidence for co-evolution of gene order and recombination rate. *Nat Genet* **33**:392-395.
- Paluszynski, J. P., R. Klassen, M. Rohe, and F. Meinhardt. 2006. Various cytosine/adenine permease homologues are involved in the toxicity of 5-fluorocytosine in *Saccharomyces cerevisiae*. *Yeast* **23**:707-715.
- Papp, B., C. Pal, and L. D. Hurst. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**:194-197.
- Paques, F., and J. E. Haber. 1999. Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol Mol Biol Rev* **63**:349-404.
- Parry, J. M., D. Sharp, and E. M. Parry. 1979. Detection of mitotic and meiotic aneuploidy in the yeast *Saccharomyces cerevisiae*. *Environ Health Perspect* **31**:97-111.

- Pevzner, P., and G. Tesler. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci U S A* **100**:7672-7677.
- Piskur, J., E. Rozpedowska, S. Polakova, A. Merico, and C. Compagno. 2006. How did *Saccharomyces* evolve to become a good brewer? *Trends Genet* **22**:183-186.
- Pratt-Hyatt, M. J., K. M. Kapadia, T. E. Wilson, and D. R. Engelke. 2006. Increased recombination between active tRNA genes. *DNA Cell Biol* **25**:359-364.
- Pribylova, L., J. de Montigny, and H. Sychrova. 2007. Osmoresistant yeast *Zygosaccharomyces rouxii*: the two most studied wild-type strains (ATCC 2623 and ATCC 42981) differ in osmotolerance and glycerol metabolism. *Yeast* **24**:171-180.
- Pribylova, L., V. Farkas, I. Slaninova, J. de Montigny, and H. Sychrova. 2007. Differences in osmotolerant and cell-wall properties of two *Zygosaccharomyces rouxii* strains. *Folia Microbiol (Praha)* **52**:241-245.
- Pribylova, L., and H. Sychrova. 2003. Efficient transformation of the osmotolerant yeast *Zygosaccharomyces rouxii* by electroporation. *J Microbiol Methods* **55**:481-484.
- Rachidi, N., M. J. Martinez, P. Barre, and B. Blondin. 2000. *Saccharomyces cerevisiae* PAU genes are induced by anaerobiosis. *Mol Microbiol* **35**:1421-1430.
- Raychaudhuri, S., and W. A. Prinz. 2006. Uptake and trafficking of exogenous sterols in *Saccharomyces cerevisiae*. *Biochem Soc Trans* **34**:359-362.
- Reiner, S., D. Micolod, G. Zellnig, and R. Schneiter. 2006. A genomewide screen reveals a role of mitochondria in anaerobic uptake of sterols in yeast. *Mol Biol Cell* **17**:90-103.
- Ricchetti, M., C. Fairhead, and B. Dujon. 1999. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* **402**:96-100.
- Richard, F., M. Lombard, and B. Dutrillaux. 2003. Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res* **11**:605-618.
- Rine, J., and I. Herskowitz. 1987. Four genes responsible for a position effect on expression from HML and HMR in *Saccharomyces cerevisiae*. *Genetics* **116**:9-22.
- Robinson, T. J., A. Ruiz-Herrera, and L. Froenicke. 2006. Dissecting the mammalian genome--new insights into chromosomal evolution. *Trends Genet* **22**:297-301.
- Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**:798-804.
- Rudin, N., and J. E. Haber. 1988. Efficient repair of HO-induced chromosomal breaks in *Saccharomyces cerevisiae* by recombination between flanking homologous sequences. *Mol Cell Biol* **8**:3918-3928.
- Ryu, S. L., Y. Murooka, and Y. Kaneko. 1998. Reciprocal translocation at duplicated RPL2 loci might cause speciation of *Saccharomyces bayanus* and *Saccharomyces cerevisiae*. *Curr Genet* **33**:345-351.
- Sankoff, D., C. Zheng, and Q. Zhu. 2007. Polyploids, genome halving and phylogeny. *Bioinformatics* **23**:i433-439.
- Scannell, D. R., K. P. Byrne, J. L. Gordon, S. Wong, and K. H. Wolfe. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**:341-345.
- Scannell, D. R., A. C. Frank, G. C. Conant, K. P. Byrne, M. Woolfit, and K. H. Wolfe. 2007. 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* **104**:8397-8402.
- Scazzocchio, C. 2000. The fungal GATA factors. *Curr Opin Microbiol* **3**:126-131.
- Schacherer, J., J. de Montigny, A. Welcker, J. L. Souciet, and S. Potier. 2005. Duplication processes in *Saccharomyces cerevisiae* haploid strains. *Nucleic Acids Res* **33**:6319-6326.
- Schacherer, J., Y. Tourrette, J. L. Souciet, S. Potier, and J. De Montigny. 2004. Recovery of a function involving gene duplication by retroposition in *Saccharomyces cerevisiae*. *Genome Res* **14**:1291-1297.
- Schmitt, M. J., and F. Breinig. 2002. The viral killer system in yeast: from molecular biology to application. *FEMS Microbiol Rev* **26**:257-276.
- Sherman, J. M., and L. Pillus. 1997. An uncertain silence. *Trends Genet* **13**:308-313.
- Smith, M. M. 1987. Molecular evolution of the *Saccharomyces cerevisiae* histone gene loci. *J Mol Evol* **24**:252-259.
- Solieri, L., S. Cassanelli, and P. Giudici. 2007. A new putative *Zygosaccharomyces* yeast species isolated from traditional balsamic vinegar. *Yeast* **24**:403-417.
- Srikantha, T., S. A. Lachke, and D. R. Soll. 2003. Three mating type-like loci in *Candida glabrata*. *Eukaryot Cell* **2**:328-340.
- Strathern, J. N., and I. Herskowitz. 1979. Asymmetry and directionality in production of new cell types during clonal growth: the switching pattern of homothallic yeast. *Cell* **17**:371-381.
- Sychrova, H., V. Braun, S. Potier, and J. L. Souciet. 2000. Organization of specific genomic regions of *Zygosaccharomyces rouxii* and *Pichia sorbitophila*: comparison with *Saccharomyces cerevisiae*. *Yeast* **16**:1377-1385.
- Tang, X. M., G. Kayingo, and B. A. Prior. 2005. Functional analysis of the *Zygosaccharomyces rouxii* Fps1p homologue. *Yeast* **22**:571-581.
- Tatusov, R. L., A. R. Mushegian, P. Bork, N. P. Brown, W. S. Hayes, M. Borodovsky, K. E. Rudd, and E. V. Koonin. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr Biol* **6**:279-291.
- Teichmann, S. A., and R. A. Veitia. 2004. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics* **167**:2121-2125.
- Thomas, J. W., and J. W. Touchman. 2002. Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends Genet* **18**:104-108.
- Thomson, J. M., E. A. Gaucher, M. F. Burgan, D. W. De Kee, T. Li, J. P. Aris, and S. A. Benner. 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat Genet* **37**:630-635.
- Torres, E. M., T. Sokolsky, C. M. Tucker, L. Y. Chan, M. Boselli, M. J. Dunham, and A. Amon. 2007. Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science* **317**:916-924.
- Tsai, I. J., D. Bensasson, A. Burt, and V. Koufopanou. 2008. Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proc Natl Acad Sci U S A* **105**:4957-4962.
- Tsong, A. E., B. B. Tuch, H. Li, and A. D. Johnson. 2006. Evolution of alternative transcriptional circuits with identical logic. *Nature* **443**:415-420.

- Valadi, H., A. Valadi, R. Ansell, L. Gustafsson, L. Adler, J. Norbeck, and A. Blomberg. 2004. NADH-reductive stress in *Saccharomyces cerevisiae* induces the expression of the minor isoform of glyceraldehyde-3-phosphate dehydrogenase (TDH1). *Curr Genet* **45**:90-95.
- van Hoof, A. 2005. Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics* **171**:1455-1461.
- Wagner, A. 2003. How the global structure of protein interaction networks evolves. *Proc Biol Sci* **270**:457-466.
- Wagner, A. 2002. Asymmetric functional divergence of duplicate genes in yeast. *Mol Biol Evol* **19**:1760-1768.
- Wagner, A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* **18**:1283-1292.
- Wang, W., H. Zheng, C. Fan, J. Li, J. Shi, Z. Cai, G. Zhang, D. Liu, J. Zhang, S. Vang, Z. Lu, G. K. Wong, M. Long, and J. Wang. 2006. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **18**:1791-1802.
- Watanabe, Y., T. Iwaki, Y. Shimono, A. Ichimiya, Y. Nagaoka, and Y. Tamai. 1999. Characterization of the Na⁺-ATPase gene (ZENA1) from the salt-tolerant yeast *Zygosaccharomyces rouxii*. *J Biosci Bioeng* **88**:136-142.
- Watanabe, Y., S. Miwa, and Y. Tamai. 1995. Characterization of Na⁺/H⁺-antiporter gene closely related to the salt-tolerance of yeast *Zygosaccharomyces rouxii*. *Yeast* **11**:829-838.
- Wei, W., J. H. McCusker, R. W. Hyman, T. Jones, Y. Ning, Z. Cao, Z. Gu, D. Bruno, M. Miranda, M. Nguyen, J. Wilhelmy, C. Komp, R. Tamse, X. Wang, P. Jia, P. Luedi, P. J. Oefner, L. David, F. S. Dietrich, Y. Li, R. W. Davis, and L. M. Steinmetz. 2007. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc Natl Acad Sci U S A* **104**:12825-12830.
- Wendland, J., and A. Walther. 2005. *Ashbya gossypii*: a model for fungal developmental biology. *Nat Rev Microbiol* **3**:421-429.
- Wickner, R. B. 1996. Double-stranded RNA viruses of *Saccharomyces cerevisiae*. *Microbiol Rev* **60**:250-265.
- Wilhelm, F. X., M. Wilhelm, and A. Gabriel. 2005. Reverse transcriptase and integrase of the *Saccharomyces cerevisiae* Ty1 element. *Cytogenet Genome Res* **110**:269-287.
- Winzler, E. A., D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Veronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston, and R. W. Davis. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**:901-906.
- Wolfe, K. H. 2006. Comparative genomics and genome evolution in yeasts. *Philos Trans R Soc Lond B Biol Sci* **361**:403-412.

- Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**:708-713.
- Wong, S., G. Butler, and K. H. Wolfe. 2002. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc Natl Acad Sci U S A* **99**:9272-9277.
- Wong, S., M. A. Fares, W. Zimmermann, G. Butler, and K. H. Wolfe. 2003. Evidence from comparative genomics for a complete sexual cycle in the 'asexual' pathogenic yeast *Candida glabrata*. *Genome Biol* **4**:R10.
- Wong, S., and K. H. Wolfe. 2005. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat Genet* **37**:777-782.
- Woolfit, M., E. Rozpedowska, J. Piskur, and K. H. Wolfe. 2007. Genome survey sequencing of the wine spoilage yeast *Dekkera (Brettanomyces) bruxellensis*. *Eukaryot Cell* **6**:721-733.
- Wyrick, J. J., J. G. Aparicio, T. Chen, J. D. Barnett, E. G. Jennings, R. A. Young, S. P. Bell, and O. M. Aparicio. 2001. Genome-wide distribution of ORC and MCM proteins in *S. cerevisiae*: high-resolution mapping of replication origins. *Science* **294**:2357-2360.
- Yamada-Inagawa, T., A. J. Klar, and J. Z. Dalgaard. 2007. *Schizosaccharomyces pombe* Switches Mating Type by the Synthesis-Dependent Strand-Annealing Mechanism. *Genetics* **177**:255-265.
- Yamashita, I., T. Maemura, T. Hatano, and S. Fukui. 1985. Polymorphic extracellular glucoamylase genes and their evolutionary origin in the yeast *Saccharomyces diastaticus*. *J Bacteriol* **161**:574-582.
- Yamashita, I., M. Nakamura, and S. Fukui. 1987. Gene fusion is a possible mechanism underlying the evolution of STA1. *J Bacteriol* **169**:2142-2149.
- Yang, F., E. Z. Alkalaeva, P. L. Perelman, A. T. Pardini, W. R. Harrison, P. C. O'Brien, B. Fu, A. S. Graphodatsky, M. A. Ferguson-Smith, and T. J. Robinson. 2003. Reciprocal chromosome painting among human, aardvark, and elephant (superorder Afrotheria) reveals the likely eutherian ancestral karyotype. *Proc Natl Acad Sci U S A* **100**:1062-1066.
- Yoshiuchi, S., T. Yamamoto, H. Sakane, J. Kadota, J. Mochida, M. Asaka, and K. Tanaka. 2006. Identification of novel mutations in ACT1 and SLA2 that suppress the actin-cable-overproducing phenotype caused by overexpression of a dominant active form of Bni1p in *Saccharomyces cerevisiae*. *Genetics* **173**:527-539.
- Yu, J., P. K. Chang, K. C. Ehrlich, J. W. Cary, D. Bhatnagar, T. E. Cleveland, G. A. Payne, J. E. Linz, C. P. Woloshuk, and J. W. Bennett. 2004. Clustered pathway genes in aflatoxin biosynthesis. *Appl Environ Microbiol* **70**:1253-1262.
- Zhang, J. 2003. Evolution by gene duplication: an update. *Trends in Ecology and Evolution* **18**:292-298.
- Zhao, S., J. Shetty, L. Hou, A. Delcher, B. Zhu, K. Osoegawa, P. de Jong, W. C. Nierman, R. L. Strausberg, and C. M. Fraser. 2004. Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome Res* **14**:1851-1860.
- Zonneveld, B. J. M., and H. Y. Steensma. 2003. Mating, sporulation, and tetrad analysis in *Kluyveromyces lactis* in K. Wolf, K. Breunig, and G. Barth, eds. *Non-conventional yeasts in genetics, biochemistry and biotechnology*. Springer-Verlag, Berlin.