LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

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.

Automated Annotation of Yeast Genomes

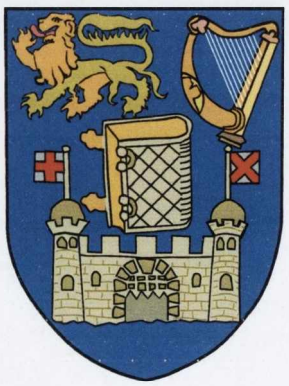
by Estelle Proux

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

Supervised by Professor Kenneth H. Wolfe

Department of Genetics
University of Dublin
Trinity College

October 2011



TRINITY COLLEGE
22 OCT 2012
LIBRARY DUBLIN

Thesis 9716

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgment.

Signature of Author.....

roux

2011

Summary

I have built an automated annotation pipeline (named YGAP – Yeast Genome Annotation Pipeline), designed specifically for new yeast genome sequences lacking transcriptome data. YGAP uses homology and synteny information from other yeast species stored in the Yeast Gene Order Browser (YGOB) database to annotate yeast genomes using the outputs from high-throughput sequencing methods. The pipeline is able to detect probable frameshift sequencing errors and can propose corrections for them. YGAP searches intelligently for introns, and detects tRNA genes and Ty-like elements. Using YGAP, I have automatically reannotated the genomes of *Saccharomyces cerevisiae* and *Naumovozyma castellii*, and compared the annotation results to those obtained with another popular annotation pipeline, AUGUSTUS. For *S. cerevisiae* and *N. castellii*, 91-93% of YGAP's predicted gene structures were identical to those in previous manually curated gene sets, and it outperforms AUGUSTUS.

I then used YGAP to automatically annotate six other new yeast genomes from the Saccharomycetaceae family, and participated in the manual correction of the genome sequence annotations. I examined the genes automatically annotated as singletons by YGAP in these six genome sequences and in the *N. castellii* genome, and classified them depending of their inferred origins. Singletons in these species represent 7% of the genes on average. Most of them seem to have arisen from duplication events, even if the original gene is not always identifiable. Some of them have become amplified, such as a 14-member family of singletons in *Tetrapisispora blattae.* Genes gained by horizontal gene transfer are extremely rare in our datasets, except in *Torulaspora delbrueckii*. Orphan genes (those lacking homologs in other lineages) represent up to 46% of the singletons. Some orphans belong to orphan gene families that can have up to 20 members (in *N. castellii*), but they are usually part of smaller families (two to five members).

The addition of the six newly sequenced genomes into the YGOB database helped improve our knowledge of the Saccharomycetaceae family. We added one or two genome sequences to clades that were previously empty. I used data from these new genomes to study how the location of the ribosomal DNA (rDNA) array has changed

during yeast evolution. I identify an ancestral rDNA location at an internal chromosomal site in Saccharomycetaceae species, and show that the rDNA array has been relocated from this site to telomeric or different internal locations in some species. Interestingly, rDNA arrays located in internal sites (ancestral or not) are flanked by genes coding for proteins that maintain the specialized chromatin structure of rDNA. Based on these observations, I suggest that the location of the rDNA array is determined by natural selection. The mechanism responsible for the relocation of the array is still unknown.

Acknowledgments

Many thanks to everyone in the Wolfe lab, past and present: David, Séan, Kevin, Jon, Aine, Nadia and Gael. Thank you for the serious and less serious talks and for your precious advices on how to survive in Ireland. It was a real pleasure to work with you and you made my days less rainy.

Many thanks also to my friends (bioinformaticians or not) all around the world: Aude, Pierre-Yves, François, Jennifer and Sophie. Thank you for your support and for sharing advices and feelings. You made me realize many times how lucky I was here in my lab.

Thanks to Ken Wolfe for being a great supervisor. Thank you for your patience, your advices and your encouragements. I have learnt a lot during those past three years.

Finally, many thanks to my family for their support, and especially to my mother who has always encouraged me to do what I wanted to do. A big thank you to Julien for his love and support 24/7.

Table of Contents

CHAPTER 1 GENERAL INTRODUCTION **9**

1.1 AN OVERVIEW OF THE SACCHAROMYCOTINA YEASTS 9

1.1.1 Scientific value of Saccharomycotina species *9*

1.1.2 Phylogenetics *11*

1.1.2.1 Saccharomycotina among the fungi kingdom 11

1.1.2.2 Phylogenetics relationship among the Saccharomycotina 12

1.1.3 Genomics of the Saccharomycotina *15*

1.1.4 YGOB: the Yeast Gene Order Browser *16*

1.2 YEAST GENOME EVOLUTION 18

1.2.1 Gene content evolution *18*

1.2.1.1 Gene gain 18

1.2.1.2 Gene loss 24

1.2.2 Structural evolution *26*

1.3 COMPARATIVE GENOMICS 28

1.3.1 Genome annotation improvement *29*

1.3.2 Nonprotein-coding elements *29*

1.3.3 Evolution on a genome-wide scale *30*

1.3.4 Changes in metabolic pathway *31*

1.3.5 Evolution of yeast lifestyle *32*

1.4 RIBOSOMAL DNA 33

1.4.1 Ribosomal RNA genes *33*

1.4.2 Evolution of rDNA array in eukaryotes *35*

1.4.3 Evolution of solo 5S genes (outside of the rDNA array) in eukaryotes *37*

1.4.4 rDNA as a lifespan determinant *39*

1.5 NEW SEQUENCING TECHNOLOGIES AND GENOME ANNOTATION 41

1.5.1 New sequencing technologies and their error rates *41*

1.5.2 Genome annotation *42*

CHAPTER 2 YGAP: THE YEAST GENOME ANNOTATION PIPELINE **44**

2.1 ABSTRACT 44

2.2 INTRODUCTION 45

2.3 METHODS 47
2.3.1 Input data 47
2.3.2 Checking the integrity of scaffolds 48
2.3.3 Locating genes 48
2.3.4 Gene models 49
2.3.5 Intron annotation 52
2.3.6 Frameshift correction 53
2.3.7 SearchDOGS and large ORF steps 54
2.3.8 Retrotransposons 54
2.3.9 Mini-YGOB browser 55
2.4 RESULTS 56
2.4.1 Tests with S. cerevisiae 56
2.4.2 Automatic annotations of Naumovozyma castellii and Tetrapisispora blattae 59
2.5 DISCUSSION 62

CHAPTER 3 STUDY OF THE SINGLETON GENES IN NEWLY SEQUENCED SACCHAROMYCETACEAE SPECIES 65

3.1 ABSTRACT 65
3.2 INTRODUCTION 65
3.3 METHODS 67
3.4 RESULTS 68
3.4.1 Horizontal gene transfers into T. delbrueckii 69
3.4.1.1 Multiple transposase-related genes 69
3.4.1.2 3-methyl-adenine DNA glycosylase gene in *T. delbrueckii* 74
3.4.1.3 A *T. delbrueckii* gene related to lignin degradation 77
3.4.2 Multiple duplications in Tetrapisispora blattae 78
3.4.2.1 Duplicated genes of the *HSP30* family 79
3.4.2.2 Duplicated genes of the PIR family in *T. blattae* 83
3.4.3 Orphan genes in the new species 85

CHAPTER 4 EVOLUTION OF THE LOCATION OF THE RIBOSOMAL DNA ARRAY IN SACCHAROMYCOTINA 89

4.1 ABSTRACT 89

4.2 INTRODUCTION 89
4.3 METHODS 90
4.4 RESULTS AND DISCUSSION 92
4.4.1 An ancestral rDNA location in Saccharomycetaceae 92
4.4.2 Species with subtelomeric rDNA arrays 99
4.4.3 Evolutionary movement of rDNA in L. thermotolerans 100
4.4.4 rDNA locations in the CTG clade 103
CHAPTER 5 CONCLUSIONS 108
REFERENCES 109
APPENDIX I EVOLUTIONARY EROSION OF YEAST SEX CHROMOSOMES BY MATING-TYPE SWITCHING ACCIDENTS 123

Abbreviations

CGOB	Candida Gene Order Browser
HGT	Horizontal Gene Transfer
Mb	Megabase
NGS	Next-Generation Sequencing
ORF	Open Reading Frame
WGD	Whole Genome Duplication
YGAP	Yeast Genome Annotation Pipeline
YGOB	Yeast Gene Order Browser

Chapter 1
General introduction

In this chapter, I give an overview on the Saccharomycotina, their phylogenetic relationships, their genomes and how they evolve. In addition, I summarize the reasons Saccharomycotina species are useful for comparative genomics studies. I also review the literature on eukaryotic ribosomal RNA genes and their evolution, with a focus on yeast ribosomal RNA genes. Finally, I describe information about new sequencing technologies and genome annotation.

1.1 An overview of the Saccharomycotina yeasts

The word "yeast" is a generic term used to designate fungal species that live exclusively or almost exclusively in a single-celled mode (Dujon 2006). Today, more than a thousand yeast species have been identified. Yeast species exist in both of the major fungal phyla, the Basidiomycota (e.g. *Cryptococcus neoformans*) and the Ascomycota (e.g. *Saccharomyces cerevisiae*). But most of the yeasts sequenced to date belong to the subphylum Saccharomycotina (Kurtzman 2011), a group in which *Saccharomyces cerevisiae*, *Candida albicans* and *Yarrowia lipolytica* are members.

1.1.1 Scientific value of Saccharomycotina species

Saccharomycotina yeasts are ideal organisms for comparative genomics studies, due to their genomic properties and the broad range of phylogenetic distances covered - the evolutionary distances within the Saccharomycotina yeasts are similar to the entire phylum of chordates (Dujon et al. 2004). However, the hemiascomycetous species that have been sequenced and most studied are also of biomedical or industrial interest (Ahearn 1998; Demain et al. 1998).

The most widespread hemiascomycetous yeast pathogenic for human is *C. albicans* (Barnett 2008). This species is endogenous in the oral, gastrointestinal or urinogenital tracts of humans and other warm blooded animals. It may be present in up to 62% of healthy people (Barnett 2008). *C. albicans* causes millions of superficial

infections per year. However, its pathogenic effects can affect several organs and are often experienced after treatment with antibiotics or among people who have a suppressed immune system (iatrogenically or from infection). *C. tropicalis* is probably the third most important *Candida* pathogen of human, causing diseases in neutropenic patients in tropical countries (Ann Chai et al. 2010). It is taxonomically close to *C. albicans* and they share many pathogenic traits. Other *Candida* species are involved in human infections, such as *C. dubliniensis* (Gutierrez et al. 2002), *C. parapsilosis* (Nosek et al. 2009) and *C. glabrata* – the second most common *Candida* pathogen after *C. albicans* (Li et al. 2007). Even *S. cerevisiae* can produce fatal complications in immune deficient patients (Aucott et al. 1990).

Yeasts have been used for many centuries in many industrial processes, especially the production of ethanol. Different strains of *S. cerevisiae* were chosen and adapted for different fermentations: baker's yeast, wine yeasts, sake yeast, brewing strains and distiller's strains (Demain et al. 1998; Liti et al. 2009). *S. cerevisiae* uses the monosaccharide glucose to produce ethanol, but other yeasts can degrade different substrates: *Clavispora lusitania* can convert cellobiose into ethanol (Freer and Detroy 1983), *Kluyveromyces marxianus* and *Torulaspora delbrueckii* can ferment the plant polysaccharide inulin (e.g. from Jerusalem artichokes) to ethanol (Demain et al. 1998). Several species of the genus *Kluyveromyces* are lactose fermenters producing ethanol from by-products of the milk industry. Moreover, some yeast enzymes have been useful in industry, such as invertase for candy and jam manufacture, β-galactosidase for production of syrups from milk or whey, and α-galactosidase for crystallization of beet sugar (Demain et al. 1998). Furthermore, yeasts can produce or even overproduce molecules of industrial interest such as riboflavin (vitamin B2), citric acid, extracellular polysaccharide, and several flavor compounds (Demain et al. 1998). In the biomedical industry, recombinant DNA technology in yeast has allowed the expression of heterologous gene products such as hormones and vaccines (Kjeldsen et al. 1997; Souciet et al. 2000).

1.1.2 Phylogenetics

1.1.2.1 Saccharomycotina among the fungi kingdom

Eukaryotes can be divided into four kingdoms: animals, plants, fungi and protists. The opisthokonta hypothesis states that fungi and animals (and some protists) are more closely related to each other than they are to plants (Cavalier-Smith and Chao 1995; Steenkamp et al. 2006) and is well supported by molecular data (Baldauf et al. 2000; James et al. 2006). This hypothesis agrees with the taxonomic data: fungi (and animals) are heterotrophic, whereas plants are autotrophic; and fungi absorb (rather than ingest, like animals) food. Fungal cell walls are structurally unique and are composed of varying proportions of chitin (from 1% of the cell wall in certain yeasts to 20% in some filamentous fungi) and glucan (mainly β-1,3-glucan) (Bowman and Free 2006).

The known fungal species are typically divided into five phyla: Ascomycota, Basidiomycota, Glomeromycota, Zygomycota and Chytridomycota (**Figure 1.1**). However, about three quarters of the described fungal species are members of the phylum Ascomycota, and the majority of the remainder are member of the phylum Basidiomycota. A combined gene phylogeny of the fungi supported monophyly of the Ascomycota, Basidiomycota and Glomeromycota (James et al. 2006). The Ascomycota and Basidiomycota form the dikarya clade, meaning that these fungi have a portion of their life cycle with paired nuclei (James et al. 2006).

Ascomycota is the largest phylum among fungi. The ascomycetous yeasts (or sac fungi) are characterized by the ascus, a structure in which meiosis takes place, and that surrounds the formed ascospores (James et al. 2006). The phylum Ascomycota is divided into three monophyletic subphyla: Taphrinomycotina (also known as Archiascomycotina, which includes *Schizosaccharomyces pombe*), Pezizomycotina (which includes filamentous fungi such as *Neurospora crassa*) and Saccharomycotina (or Hemiascomycetes, which includes *S. cerevisiae*).

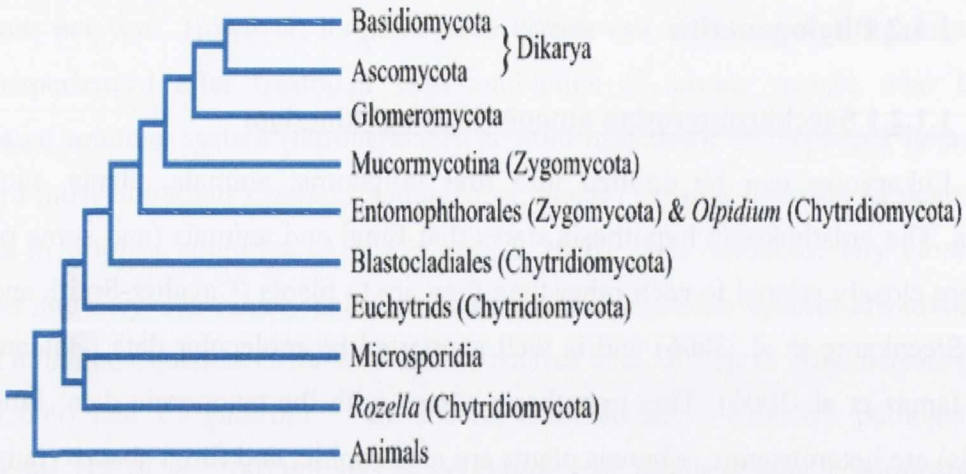


Figure 1.1 The main branches of the kingdom Fungi, derived from combined data for six genes of 199 fungi (adapted from Bruns 2006; James et al. 2006). The kingdom Animals here is used as an outgroup.

1.1.2.2 Phylogenetics relationship among the Saccharomycotina

The Saccharomycotina can be divided into 3 major clusters. First, there is one cluster whose only known member is *Yarrowia lipolytica* (Barth and Gaillardin 1997). Second, there is a cluster usually referred to as the "CTG clade" which consists of species that translate the CTG codon as serine rather than leucine (Miranda et al. 2006). The CTG clade includes some species, such as those in the genera *Debaryomyces*, *Lodderomyces*, *Meterozyma* and *Scheffersomyces*, that have well-described sexual cycles and are classified as the family Debaryomycetaceae (Kurtzman 2011). The CTG clade also includes many pathogens in the genus *Candida*, such as *C. albicans*, that are either asexual or that only mate under highly restrictive and rare circumstances. For historical reasons, such asexual species are not classified into any family; the genus *Candida* is polyphyletic and is defined only by the failure of its members to mate under standard microbiological conditions. However, molecular phylogenetic analyses show very consistently that the CTG clade is a monophyletic group containing the Debaryomycetaceae and many *Candida* species related to *C. albicans* (Butler et al. 2009). The third clade is the family Saccharomycetaceae, which includes *S. cerevisiae* and is the main focus of this thesis. In addition to these three major clades, the subphylum Saccharomycotina also includes many other, smaller, families (Kurtzman 2011) such as the Phaffomycetaceae which includes the biotechnologically important species *Komagataella pastoris* (formerly called *Pichia pastoris*, used as a host for expression of recombinant proteins) (Küberl et al. 2011).

The major division within the Saccharomycetaceae is between the species that share a common ancestor that underwent a whole-genome duplication (WGD) (Wolfe and Shields 1997), and the species that diverged prior to this event. We call these species post-WGD species and non-WGD species respectively. But the Saccharomycetaceae can be further divided in twelve genera (Kurtzman and Robnett 2003): six post-WGD genera (*Saccharomyces*, *Kazachstania*, *Naumovozyma*, *Nakaseomyces*, *Tetrapisispora* and *Vanderwaltozyma*) and six non-WGD genera (*Zygosaccharomyces*, *Zygotorulaspora*, *Torulaspora*, *Lachancea*, *Kluyveromyces* and *Eremothecium*). In general, each of these genera is a monophyletic clade. This revised genus nomenclature vastly simplified the taxonomic treatment of the family, because under the previous scheme (Kurtzman and Fell 1998) the large genera *Kluyveromyces* and *Saccharomyces* were polyphyletic. Under the revised scheme (Kurtzman and Robnett 2003) terms such as "*Saccharomyces sensu stricto*" and "*Saccharomyces sensu lato*" became obsolete because the "*sensu lato*" species were reassigned to new monoplyletic genera, leaving the "*sensu stricto*" species in a monophyletic genus, *Saccharomyces*. A phylogenetic tree of the family Saccharomycetaceae is presented in **Figure 1.2**.

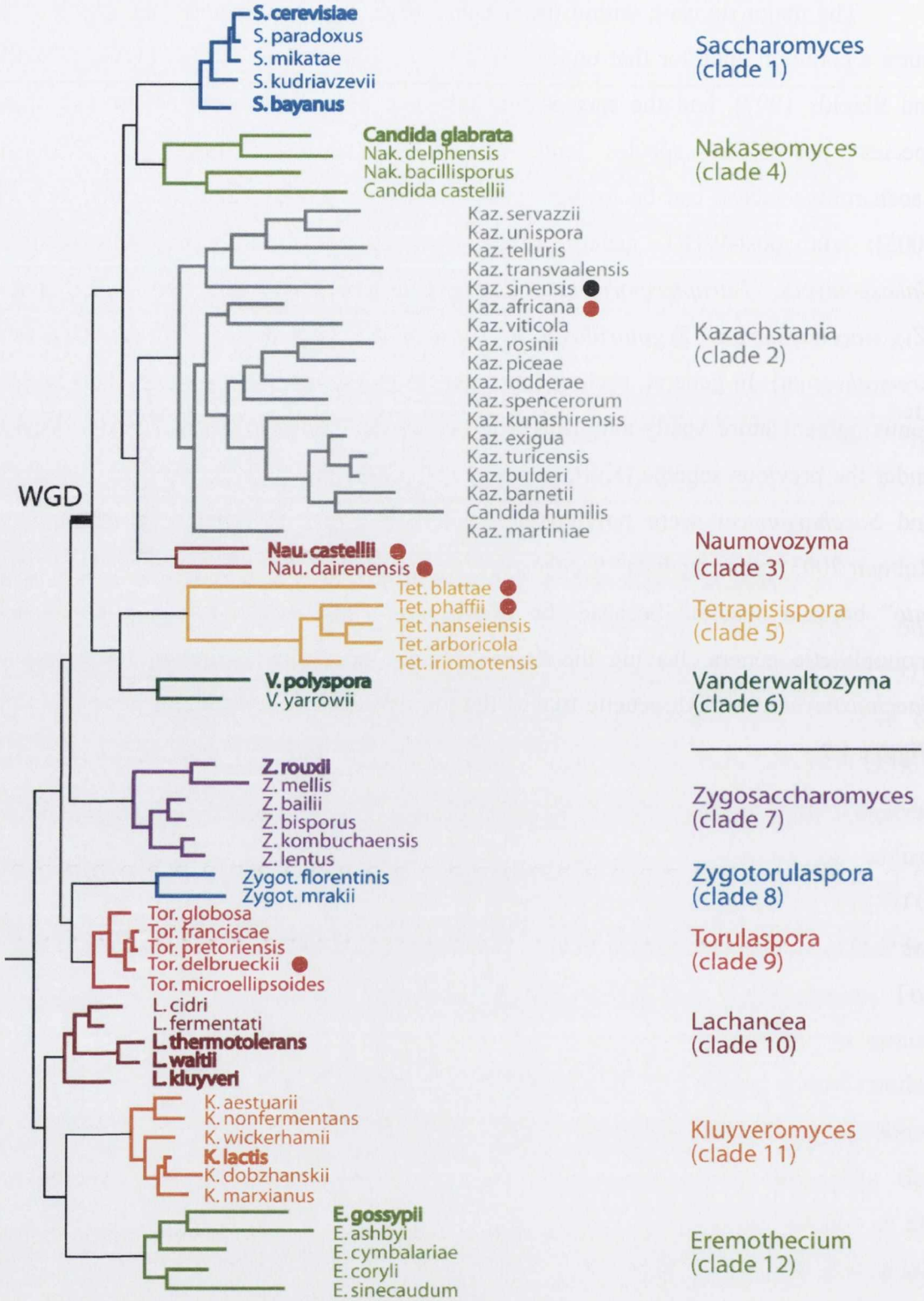


Figure 1.2 Phylogenetic relationships among the family Saccharomycetaceae (modified from Kurtzman 2011). Species in clades 1-6 are post-WGD, and species in clades 7-12 are non-WGD. The species in bold have their genomes fully sequenced, annotated and present in YGOB. In **Chapters 2 and 3** we are going to focus on seven yeast species: six of them are present in this tree (red dots); the last one, *Kazachstania naganishii*, is not present in this tree. Its closest related species is marked by a black dot.

1.1.3 Genomics of the Saccharomycotina

S. cerevisiae was the first eukaryote to have its genome fully sequenced (Goffeau et al. 1996), largely because it is smaller and simpler than other eukaryotic genomes. Its size is about 12 Mb, is composed of 16 chromosomes, and contains about 5600 protein-encoding genes. The genome is very compact: more than 72% of the genome is occupied by ORFs (excluding rDNA repeats), leaving relatively little noncoding DNA and other structural and functional elements (Dujon 1996). With the development of sequencing technologies, many other eukaryotic genomes are now sequenced, and show that hemiascomycetes share some genomic characteristics that are not observed in other eukaryotes, even other ascomycetes. For example, in ascomycete yeasts, fewer than 5% of the genes contain introns, usually a single one frequently located at the 5' end (Bon et al. 2003). However, in *Schizosaccharomyces pombe* (Taphrinomycotina), more than 40% of the genes contain introns (Wood et al. 2002). Similarly, other fungal species do not have a genome as compact as the genome of *S. cerevisiae*: in *N. crassa*, genes cover about 45% of the genome sequence, and its genome is larger (40 Mb) and contains more genes (about 10,000) (Galagan et al. 2003).

The genome of *S. cerevisiae* also contains 40 genes for small nuclear RNA molecules, 275 transfer RNA (tRNA) genes and about 140 copies of rRNA genes in a tandem array on chromosome XII that cover about 7% of the genome (Goffeau et al. 1996; Johnston et al. 1997). The tRNA genes are found scattered throughout the genome, usually as single transcriptional units, and 57 (21%) of them contain introns in the tRNA precursors (Hani and Feldmann 1998).

Finally, the genome of *S. cerevisiae* also contains 51 complete retrotransposons (or Ty elements), belonging to five families: Ty1 (32), Ty2 (13), Ty3 (2), Ty4 (3) and Ty5 (1). Each Ty element is about 6 kb in length, including long terminal direct repeats (LTRs) about 330 bp located at each end of the element (Mieczkowski et al. 2006). Two genes are commonly found in LTR retrotransposons, homologs of the retroviral *gag* and *pol* genes. The *gag* gene of retroviruses encodes structural proteins of the viral particle, and the retroviral *pol* locus encodes a polyprotein with protease, integrase, reverse transcriptase, and RNAseH catalytic domains (Neuveglise et al. 2002). Only three of these five Ty families are known to be transpositionally active in *S. cerevisiae*: Ty1,

Ty2, and Ty3. The families are defined by the sequences of their LTRs. Ty5 elements from *S. cerevisiae* are either solo LTRs or degenerate elements that have accumulated several deleterious mutations (Neuveglise et al. 2002). However, intact and active copies of Ty5 have been found in *S. paradoxus,* a closely related species to *S. cerevisiae* (Zou et al. 1996). The genome of *S. cerevisiae* also contains about 280 solo LTRs or LTR fragments (Kim et al. 1998). Solo LTRs are the remnants of sites from which a Ty element has been lost during evolution, due to excision by recombination between the two LTRs at each end of the element. Overall, retrotransposon sequences (complete or not) constitute about 3% of the genome. They are found scattered throughout the genome, either in the form of one single element, or up to ten elements accumulated in "hotspots of transposition" (Hani and Feldmann 1998). At least in *Saccharomyces*, the location of Ty elements varies extensively from strain to strain within a species.

1.1.4 YGOB: the Yeast Gene Order Browser

In 2005 our laboratory released YGOB, an online tool designed to facilitate the visualization of synteny within and between the genomes of several yeast species belonging to the "*Saccharomyces* complex", including both post-WGD and non-WGD species (Byrne and Wolfe 2005).

YGOB consists of (1) a database of homologous genes that have been intensively manually curated based on both sequence similarity and synteny, and (2) software that calculates and displays the synteny relationships between any gene in one species and its orthologs in other species. The latest version of the database (version 4 (Gordon et al. 2009)) contains the genomes of eleven species: five post-WGD species (*S. cerevisiae*, *S. bayanus*, *C. glabrata*, *N. castelli* and *V. polyspora*) and six non-WGD species (*Z. rouxii*, *L. thermotolerans*, *L. waltii*, *L. kluyveri*, *K. lactis* and *E. gossypii*). It also contains a reconstructed ancestral genome, inferred to be the genome organization of an ancestor that existed about 100 million years ago, at the point immediately before the WGD occurred (Gordon et al. 2009).

YGOB is based on two structural features: tracks and pillars (**Figure 1.3**). Tracks (horizontal elements) represent portions of chromosome and allow the synteny relations among genomes to be shown. Each non-WGD species has one track, and each post-WGD species has two tracks (A and B). Pillars (vertical elements) show sets of

orthologous genes among the species or WGD-derived paralogs (ohnologs) within a genome.

In this thesis (**Chapter 2**), I have automatically annotated genome sequences of six new Saccharomycotina yeast species (five post-WGD species and one non-WGD species) and reannotated the genome of *N. castelli*. Those species have been integrated into the laboratory's private version of YGOB, and will be added to the public YGOB server after manual curation is complete.

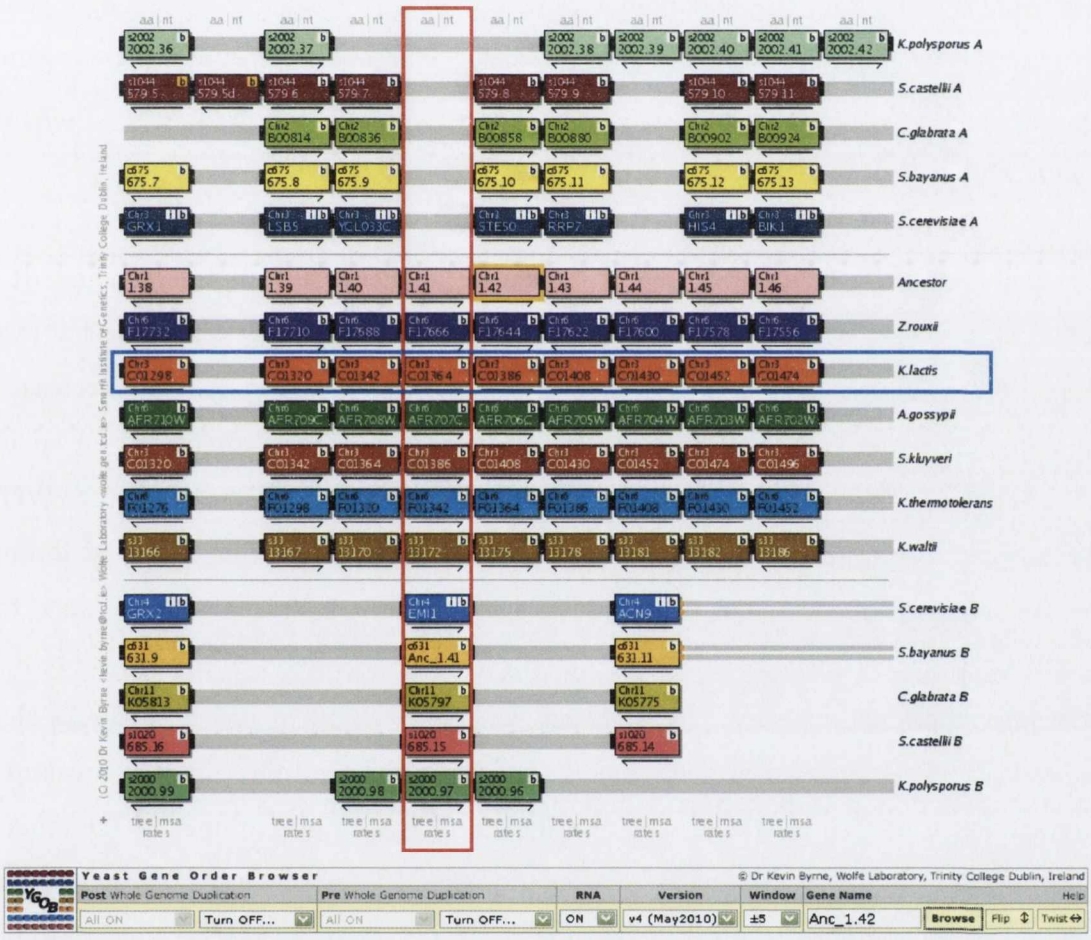


Figure 1.3 Screenshot from YGOB. Each small rectangle represents a gene. Tracks (for example the blue outline box) show a portion of one chromosome from one species (here, *K. lactis*) and the gene order on this fragment. Pillars (for example, the red outline box) show sets of orthologous genes among species, or paralog pairs (ohnologs) formed by the WGD.

1.2 Yeast genome evolution

1.2.1 Gene content evolution

1.2.1.1 Gene gain

The key mechanism of gene gain in all species is gene duplication (Lynch and Conery 2000). Several kinds of duplications can be found in yeast genomes (Liti and Louis 2005): (i) large duplicated blocks where the gene orientation is conserved, resulting from WGD and subsequent rearrangements; (ii) spontaneously formed segmental duplications; (iii) subtelomeric duplications; (iv) duplication of single genes, dispersed in the genome; (v) tandem gene repeats. Moreover, genes can also be gained by horizontal gene transfer (HGT) or (very rarely) can be formed *de novo* from non-coding sequence or by the fusion of two existing genes (Dujon 2010).

In the early 1990s, by analyzing partial sequences from the genome of *S. cerevisiae*, several large duplicated blocks with conserved synteny were detected (Lalo et al. 1993; Melnick and Sherman 1993; Wolfe and Lohan 1994). After the release of the complete genome, it was suggested that the baker's yeast evolved via an ancient whole-genome duplication (WGD) (Wolfe and Shields 1997). After the WGD, chromosomes were dramatically rearranged, duplicated blocks were scattered through the genome, and the original block order was lost. 55 duplicated blocks were first identified using a sequence similarity method (Wolfe and Shields 1997), and more were found by additional analyses (Seoighe and Wolfe 1999). 50 of these 55 blocks share orientation of the two copies with respect to the centromere, indicating the breakup of original large blocks by reciprocal translocation. The origin of the WGD remains uncertain. *De novo* WGD has been experimentally tested in *S. cerevisiae*. An autotetraploidization event (duplication of the entire chromosome set) seems not to be a good explanation, because autopolyploid derivatives of *S. cerevisiae* have an elevated rate of chromosome instability and die in stationary phase, failing to arrest their mitotic growth (Andalis et al. 2004). Alternatively, an allotetraploidization event (hybridization between two distinct but related species) has also been tested. Allotetraploids between *S. cerevisiae* and *S. paradoxus* have been produced, but in most cases they have limited meiotic viability (Greig et al. 2002). However, none of these experiments is conclusive,

because they involved the descendants from the ancestral duplication, and not their nonduplicated progenitors (Dujon 2006).

But the 55 duplicated blocks represent only one third of *S. cerevisiae* paralogs (Llorente et al. 2000a). Segmental duplications constitute a major signature of various eukaryotic genomes, unevenly dispersed along the different chromosomes, with clustering near subtelomeric and pericentromeric regions (Marques-Bonet et al. 2009; Dujon 2010). However, in yeast genomes, if one excludes subtelomeric regions, segmental duplications are usually rare (Dujon 2010) – but they can be frequent in some species, such as *C. glabrata* (Poláková et al. 2009). Spontaneous duplications of large segments containing many genes can be generated at high frequency in experimental cultures of *S. cerevisiae* (Koszul et al. 2004). Different types of duplications were observed (**Figure 1.4**): tandem intrachromosomal duplications are frequently found and are unstable at meiosis; interchromosomal duplications forming a new chromosomal end are also frequently observed; interchromosomal duplications with reciprocal translocations of distal chromosome segments are stable; finally, duplications can also lead to the formation of supernumerary chromosomes or circular episomes (Koszul et al. 2006; Dujon 2010). It has been shown that in *S. cerevisiae* segmental duplications arise from untimely DNA synthesis events requiring the major replicative polymerase δ (Pol δ), more precisely the Pol32 subunit of Pol δ (Payen et al. 2008). Two different mechanisms have been described: a classical RAD52-dependent break-induced replication (BIR) mechanism when the duplications are anchored by dispersed repeated elements, such as remnants of Ty elements; and a RAD52-independent microhomology/microsatellite-induced replication (MMIR) mechanism in the combined absence of homologous recombination, microhomology-mediated, and nonhomologous end joining machineries (Payen et al. 2008).

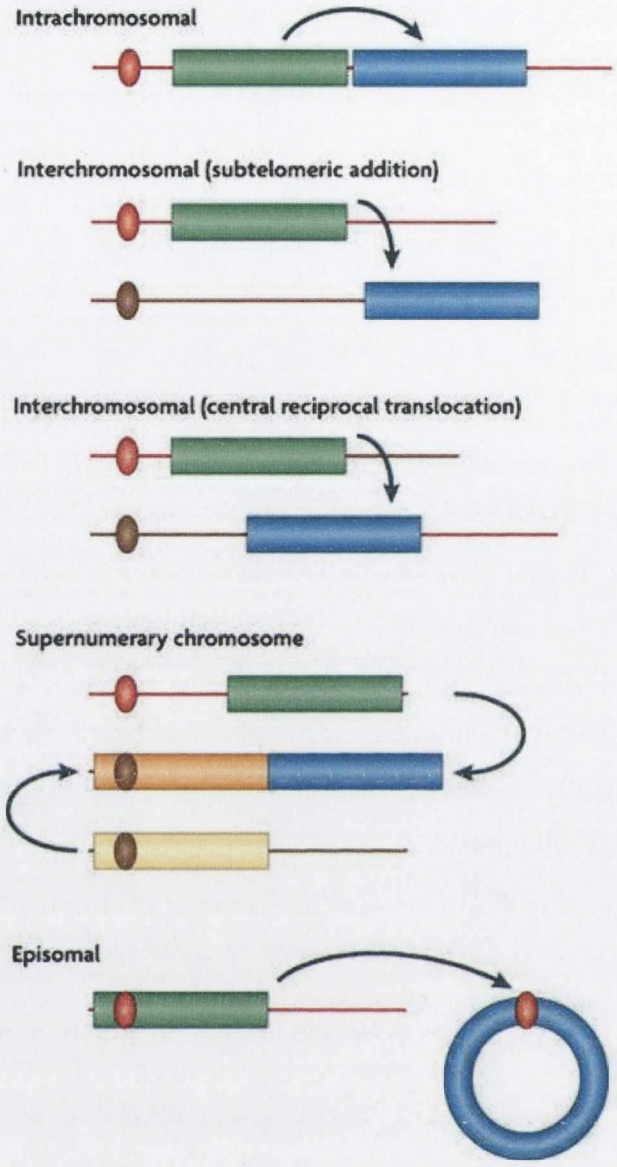


Figure 1.4 Types of segmental duplications in *S. cerevisiae* as observed in experimental evolution assays (Dujon 2010). Intrachromosomal duplications and interchromosomal duplications with subtelomeric addition are the most frequent structures observed.

In *S. cerevisiae*, a third of gene duplications occur in the subtelomeres. These regions are often lacking from so-called "whole genome" sequences because their high repeat content and extensive sequence similarity (between chromosomes) make it difficult to assemble these regions and to distinguish orthologs and paralogs, and as a result they remain relatively understudied. They are known to be highly recombinogenic (Pryde et al. 1997). A study of three subtelomeric gene families that were historically linked with maltose metabolism in yeasts from the *Saccharomyces* genus showed that,

in some yeasts, the *MAL* genes have completely disappeared, while in others, they show multiple recent duplication events. Moreover, the evolutionary rate at which these changes have taken place is exceptional, with wide differences in copy number within closely related species of the group (Brown et al. 2010). Many of the duplicated genes in subtelomeric regions are involved in secondary metabolism. They are not essential, but they play an important role in the adaptation to new environmental conditions (Liti and Louis 2005). The origin of duplicated genes within the subtelomeric regions remains unknown. One explanation could be the high frequency of ectopic recombination (Liti and Louis 2005). A high frequency of recombinational exchange has been measured between the Y′ elements of subtelomeric regions (Louis and Haber 1990). The Y′ repeated sequence family is composed of long and short elements that reside at 17 of the 32 chromosome ends. This recombination should result in the homogenization of the two forms of Y′ (long and short) and the disappearance of one of them. However both of them are maintained in many yeast strains, implying that Y′ elements tend to recombine preferentially with elements from the same size class (Louis et al. 1994). This mechanism allows for amplification and diversification in the subtelomeres. Moreover, subtelomeric specific epigenetic effects, like chromatin-dependent silencing, may also participate to the evolutionary potential of these regions, for example, by allowing swift divergence of the transcriptional regulation of the duplicated copies (Brown et al. 2010).

Another contribution of the redundancy of yeast genomes is made by dispersed copies of paralogous genes. These copies could have originated from ancient segmental duplications followed by extensive gene loss and chromosomal rearrangements (Dujon 2006). No direct mechanisms operating at the chromosomal level have been identified that would create dispersed copies of single genes, except by the activity of transposons as an intermediate. This phenomenon has been described in *S. cerevisiae*, in which Ty retrotransposons can duplicate part of a gene and integrate it into a new location in the genome (Schacherer et al. 2004). In this experiment, dispersed intra- or inter-chromosomal duplications of the terminal domain of the studied gene (a particular mutated allele of the *URA2* gene) were found. The duplicated sequence is generally flanked by a poly(A) tract, which suggests that these duplication events are RNA-mediated. Moreover, the duplicated copy is always located in Ty1 sequence, and the artificial activation of a Ty1 transcription process increased the frequency of the

duplication events. These results suggest that these duplications involve template switching of the polymerase during the reverse transcription phase and the subsequent integration of the cDNA into a Ty1 element. However, due to the scarcity of active retrotransposons in yeast genomes (Bleykasten-Grosshans and Neuveglise 2011), this mechanism should play only a limited part compared to the other mechanisms of duplication.

Clusters of identical or similar genes exist in many eukaryotes, and undergo continuous reorganization by duplication and deletion of gene copies. In yeasts, shorter tandem arrays (two to three copies) are more numerous than longer ones. Tandem gene repeats are significantly more numerous in some lineages, such as *D. hansenii* (Dujon 2006). In most cases the sequences of the gene copies are diverged, which is more consistent with functional diversification than with the simple need for copy-number increases (Dujon 2010). Several examples are consistent with a role of tandem duplication in adaptive evolution (Dujon et al. 2004). For example, in *C. glabrata*, the expansion of six additional YPS (yapsin, or yeast aspartyl protease) genes, which encode proteases required for virulence, can be observed, whereas there are only two of them in *S. cerevisiae* (Kaur et al. 2007). The expansion of this cluster is consistent with the role of these enzymes in processing a cell wall adhesion necessary for the adherence of the yeast to mammalian cells. In other cases, such as the *CUP1* and *ENA1* tandem gene arrays of *S. cerevisiae*, the sequences of the gene copies are virtually identical and the cluster is able to expand or contract to match gene expression levels to environmental conditions (Fogel and Welch 1982; Wieland et al. 1995; Warringer et al. 2011).

There are several possible outcomes in the evolution of duplicate genes (Lynch and Conery 2000): one copy can accumulate degenerative mutations and be turned into a pseudogene (nonfunctionalization); one copy can acquire a useful new function while the other performs the ancestral function (neofunctionalization); or, both copies can partition the ancestral function between them and their total capacity is reduced to the level of the single-copy ancestral gene (subfunctionalization). The latter two mechanisms have been documented in yeast. For example, neofunctionalization was reported in a study that resurrects the ancestor of the homologous genes *ADH1* and *ADH2*, which differ by 24 amino acids (Thomson et al. 2005). Adh1 (alcohol

dehydrogenase 1) reduces acetaldehyde to ethanol, while Adh2 consumes the accumulated ethanol. The kinetic behavior of the ancestor suggests that it was optimized to produce but not consume ethanol. These results suggest that *ADH2* was preserved because it had a novel function; however it is also possible that *ADH2* was originally kept for another reason, and gained its novel function later. Another study identified in *S. cerevisiae* four pairs of duplicated genes created by the WGD (*ORC1/SIR3*, *SNF12/RSC6*, *RNR2/RNR4* and *HBS1/SKI7*) that have been preserved by subfunctionalization (van Hoof 2005); in each case, both of the *S. cerevisiae* gene knockouts could be complemented by the single ortholog from the non-WGD species *Lachancea kluyveri*.

Duplication may be the major mechanism of gene gain, but other processes for the acquisition of new genes exist. Horizontal gene transfer (HGT) is important in prokaryotes but rarer in eukaryotes. A study on the acquisition of prokaryotic genes in 60 fungal genomes detected 713 transferred genes, with particularly high HGT rates in Pezizomycotina (Marcet-Houben and Gabaldon 2010). The most notable example is probably the acquisition of the bacterial *URA1* gene, coding for the enzyme dihydroorotate dehydrogenase, which functions in the biosynthesis of uracil. This gene was gained by HGT in the Saccharomycotina lineage from a *Lactococcus* bacterium, and then was vertically transmitted to many extant members of this family (Gojkovic et al. 2004; Hall et al. 2005). This gene coexisted with the ancestral eukaryotic gene *URA9*, and eventually replaced it in some lineages, including *S. cerevisiae* (Scannell et al. 2007a). Hall et al. (2005) searched for the presence of genes transferred from bacteria to the genomes of *S. cerevisiae* and *Eremothecium gossypii*, and only found eleven genes (including *URA1*). Interestingly, eight of these eleven genes are located in subtelomeric regions. This result suggests that subtelomeric regions may be preferred sites for the acquisition and insertion of novel DNA.

Finally, genes can also be gained by *de novo* formation. These genes are generally species-specific, of unknown function and unclear origins. The molecular assembly of novel genes has been shown to be facilitated by intron shuffling or exonization of mobile sequences in complex eukaryotes (Long et al. 2003). However, due to the scarcity of spliceosomal introns in Saccharomycotina and the absence of non LTR-retrotransposons – and the loss of all active copies of LTR-retrotransposons in

many lineages – these two mechanisms should have a limited role in yeast evolution. They could be partly involved in the fusions and fissions of protein domains that have been observed in yeast genomes (Durrens et al. 2008). The *de novo* formation of genes is also possible by mutational sequence changes. Several studies have illustrated these mechanisms. In *S. cerevisiae*, a short protein-coding gene, *BSC4*, has no homologs in all the other sequenced fungal genomes. This gene may be involved in the DNA repair pathway during the stationary phase and contribute to the robustness of *S. cerevisiae*, when shifted to a nutrient-poor environment. It was proposed that *BSC4* may have evolved from a previously expressed noncoding sequence (Cai et al. 2008). Another example involved the *MDF1* gene (Mating Depressing Factor 1) and its anti-sense partner *ADF1* (Antisense of Depressing Factor 1). It was demonstrated that *MDF1* most likely originated de novo from a previously non-coding sequence and can significantly suppress the mating efficiency of baker's yeast in rich medium by binding *MAT α2* and thus promote vegetative growth (Li et al. 2010).

1.2.1.2 Gene loss

Gene loss, as well as gene gain, modifies the content of a genome. Two circumstances could allow a gene to be lost (Scannell et al. 2007a): (i) if the selection pressure that maintains the gene disappears; and (ii) if a new selection pressure appears and causes the gene to be maladaptive. Several examples of gene loss have been documented in yeasts, illustrating these situations.

A study on the GAL pathway in yeast genomes has found that at least three independent lineages have inactivated or lost most or all the genes of this pathway, while leaving interacting genes intact (Hittinger et al. 2004). The GAL pathway converts galactose into glucose-6-phosphate, which enters glycolysis (Bhat and Murthy 2001). In *S. cerevisiae*, the GAL pathway is encoded by seven regulatory and structural genes. Interestingly, S*accharomyces kudriavzevii* retains remnants of all the seven *GAL* genes as syntenic pseudogenes, containing multiple stop codons, frameshifts and deletions. The degree of sequence degeneration of the *GAL* pseudogenes suggests that their inactivation occurred soon after *S. kudriavzevii* diverged from *S. cerevisiae*, and that a change in niche may have relieved the purifying selection acting on the *GAL* genes, making them unnecessary. The *GAL* genes were also lost independently in other lineages (Hittinger et al. 2004). Compared to most *Saccharomyces* species which have

been isolated from sugar-rich substrates, *S. kudriavzevii* was isolated from soil and rotten leaves. The degeneration of the GAL pathway in this species may represent only one set of the physiological changes that happened within this lineage (Hittinger et al. 2004). Interestingly, Hittinger et al. (2010) demonstrated that the GAL pathway in *S. kudriavzevii* was example of a balanced unlinked gene network polymorphism: in Portuguese strains, the GAL gene network is functional, and in Japanese strains, the network is non-functional and composed of allelic pseudogenes. The average genome-wide divergence of synonymous sites (*dS*) between the Japanese and Portuguese strains across all annotated genes is 0.021, while divergence among all sites is 0.011, which in both cases is low. Peaks of extreme sequence divergence between populations are found centered on the *GAL* genes, which can be explained by strong balancing selection on the *GAL* genes. This suggests that non-functional alleles are more fit in some genetic backgrounds or environmental conditions (Hittinger et al. 2010).

A second example of gene loss concerns the *AQY1* and *AQY2* genes, two water transporters (or aquaporins) in some strains of *S. cerevisiae* (Will et al. 2010). The paralogous *AQY1* and *AQY2* genes are implicated in freeze-thaw stress: rapid export of water through aquaporins is thought to increase freeze-thaw survival by preventing intracellular shearing due to water crystallization (Tanghe et al. 2004). However, whereas freeze-thaw–tolerant *S. cerevisiae* strains (isolated from oak soil in the North-eastern United States) harbor functional aquaporin genes, sensitive strains (isolated from warm environments) lost aquaporin function (Will et al. 2010). Will et al. (2010) showed that *AQY1* and *AQY2* genes have been lost at least 6 independent times, through 2 partial selective sweeps at *AQY2* and possibly others at *AQY1*. The high variation between strain groups, and the non-random retention or loss of both paralogs in diverse strains, is consistent with the establishment of balanced polymorphism. Moreover, the loss of both aquaporin genes provides a major fitness advantage on high-sugar substrates common to many strains' natural niche. The antagonistic effect of aquaporin function contributes to the maintenance of both functional and non-functional alleles in *S. cerevisiae* (Will et al. 2010).

Another example of gene loss is the loss of all the *BNA* genes in *C. glabrata* (Domergue et al. 2005). *S. cerevisiae* can synthesize the nicotinic acid (NA) moiety of NAD+ from tryptophan using the kynurenine pathway, or *BNA* pathway (Panozzo et al.

2002), but *C. glabrata* relies solely on exogenous sources of NA for growth. Considering the importance of this pathway, the loss of the *BNA* genes in *C. glabrata* is likely to have occurred under strong positive selection (Scannell et al. 2007a). A last well-known example involves the gene *a2* from the ancestral mating-type (*MAT*) locus in yeast. In *C. albicans*, this gene is required to activate **a**-specific genes in **a** cells. However, this *a2* gene is missing in the genome of *S. cerevisiae* and other post-WGD yeasts (Tsong et al. 2003). In the baker's yeast, the **a**-specific genes are expressed by default in **a** cells, and are repressed by the *α2* gene in α cells. It has been shown that during the evolution from one state to another, an intermediate stage is likely to have involved redundant control by both systems (Tsong et al. 2006). In this scenario, the loss of the *a2* gene was complemented by the activity of another one. However, it is also difficult to exclude the possibility that the change was favored by selection for unknown reason (Scannell et al. 2007a).

1.2.2 Structural evolution

Chromosomal rearrangements are involved in one of the primary causes of reproductive isolation, called the chromosomal speciation model (Fischer et al. 2000). They induce the formation of multivalents (an association of more than two homologous chromosomes) during meiosis, which are prone to mis-segregation and can result in the production of aneuploid gametes (containing a number of chromosomes different from the normal haploid complement) and therefore the decrease of fertility. It has been demonstrated that most rearrangement events that have occurred in the *S. cerevisiae* lineage since the WGD could be classified as either reciprocal translocations or inversions (Gordon et al. 2009).

A study on the species belonging to the *Saccharomyces* genus (Fischer et al. 2000) detected reciprocal translocations by separating chromosomes by electrophoresis and hybridization to single-gene probes. *S. paradoxus* and *S. kudriavzevii* showed no detectable translocations compared to *S. cerevisiae*. However, four reciprocal translocations were identified in *S. cariocanus*, and one non-reciprocal and three reciprocal translocations in *S. bayanus*. Moreover, two different translocations were found in *S. mikatae* in two different isolates. Interestingly, in *S. cariocanus*, one of the reciprocal translocation is probably due to an ectopic recombination event between two

highly similar duplicated genes, *TEF1* and *TEF2*; similarly, one reciprocal translocation in *S. bayanus* results from an ectopic recombination between duplicated genes. In the other cases, Ty elements have been found within 9 of the 15 remaining breakpoint intervals, and tRNA sequences are present within three intervals where no repeated elements have been characterized. tRNA genes have previously been linked to sites of genomic rearrangements between *E. gossypii* and *S. cerevisiae* (Dietrich et al. 2004). There are several possible mechanisms by which tRNA sequences could induce genomic rearrangements (Gordon et al. 2009): for example, illegitimate recombination between these multiple identical sequences could result in reciprocal translocations (Pratt-Hyatt et al. 2006); Ty elements tend to integrate besides tRNA sequences, and Ty-mediated reciprocal translocations are found in industrial strains (Rachidi et al. 1999); replication forks have a tendency to stall near highly-expressed sequences, like tRNA genes, and defects in chromosome replication can lead to translocations that are thought to result from recombination events at stalled DNA replication forks (Labib and Hodgson 2007). Moreover, crossovers between Ty elements or solo LTR elements located on non-homologous chromosomes can also generate translocations (Mieczkowski et al. 2006).

In contrast to prokaryotes, where the inversions are predominantly large, most of the inversions in yeast appear to be small, often involving only a single gene (Seoighe et al. 2000; Huynen et al. 2001). The proportion of local gene inversions with conservation of synteny was calculated for 13 yeast species (Llorente et al. 2000b). This study demonstrated that inversions are rare even over relatively long evolutionary distances (from *S. cerevisiae* to most *Kluyveromyces* species), but accumulate and become more numerous over even longer evolutionary distances (between *S. cerevisiae* and *Pichia*, *Candida*, *Debaryomyces* and *Yarrowia* species). Another study on gene order evolution between *S. cerevisiae* and *C. albicans* predicted about 1100 inversions, most of which involved single genes, and showed that small size inversions were a major cause of gene order change between distantly related species (Seoighe et al. 2000). The results from a comparison between *S. cerevisiae* and *S. bayanus* suggest a possible mechanism of gene inversion through an initial inverted duplication event flowed by differential divergence or loss of the two copies of the duplicated gene in two different species (Fischer et al. 2001). However, the comparison of *S. paradoxus*, *S. mikatae* and *S. bayanus* revealed a total of 20 inversions; all of them are flanked by tRNA genes in opposite transcriptional

orientation and usually of the same isoacceptor type (Kellis et al. 2003). This suggests a mechanism that generates inversions mediated by ectopic recombination.

1.3 Comparative genomics

With the development of next-generation sequencing (NGS) technologies, the number of genome sequencing projects has increased drastically during the past few years. As a result, it is almost impossible to estimate the number of genome projects currently in progress. In 2009, 5543 genome sequencing projects were registered on the Genomes OnLine Database (Liolios et al. 2010), and almost 900 of them were complete. With the availability of all these complete sequences, comparative genomics – the study of the genetic relationship between and within organisms by comparing features of their genomes (Liti and Louis 2005) – has become an invaluable tool to understand genomes. Complete genomes allow for global views, and multiple genomes increase predictive power of comparative analysis (Kellis et al. 2004b). The assumption behind comparative genomics is that the studied genomes have a common ancestor, and that every base pair in each modern sequence can be explained by the action of evolution (selection pressures and random genetic drift) during descent from this original ancestral genome (Ureta-Vidal et al. 2003).

The first sequence-based comparative genomics studies involved bacterial genomes (Koonin and Mushegian 1996; Koonin et al. 1997; Cole 1998; Brosch et al. 2000), as they were among the first organisms to be fully sequenced (Fleischmann et al. 1995; Fraser et al. 1995). At this time bacterial genomes were ideal for comparative genomics studies for practical reasons (compact genomes, lack of introns) as well as medical and industrial purposes. The first eukaryote to be completely sequenced was the baker's yeast, *Saccharomyces cerevisiae* (Goffeau et al. 1996). With the improvement of sequencing technologies, many larger genomes have been sequenced. Now comparative genomics methods have been applied to many taxa (e.g. mammals, plants). For example, the 1000 Genomes Project is the first project to sequence the genomes of a large number of people, in order to provide a comprehensive resource on human genetic variation. Recently, the participants to this project have released a comprehensive map of mobile element insertion polymorphisms in human (Stewart et al. 2011) and a study

on the diversity of human copy number variation and multicopy genes (Sudmant et al. 2010). In plants, comparative genomics analyses have allowed the identification of paleo-duplications in the rice, wheat, sorghum, and maize genomes and to propose a common ancestor for the grasses that contains five proto-chromosomes (Salse et al. 2008). Below, I summarize some of the insights that have come from comparative genomics in yeasts.

1.3.1 Genome annotation improvement

Comparative approaches have proved to be an efficient method of gene discovery and annotation improvement. A comparative analysis of *S. cerevisiae* based on three relative *Saccharomyces* species (*S. paradoxus, S. mikatae* and *S. bayanus*) led to a revision of the original annotation of the genome (Kellis et al. 2003): for example, in the original genomic annotation, genes were annotated as the longest possible open reading frame (ORF). This study suggested that in 120 cases, the methionine that produces the longest frame is not the true translational start codon. It also proposed the elimination of more than 500 annotated spurious ORFs, and the addition of new small ORFs, some of them containing introns. A similar approach based on the comparison of *S. cerevisiae* with *N. castellii* and *L. kluyveri* also suggested the suppression of more than 500 ORFs and the addition of small ORFs (fewer than 100 codons in length) (Cliften et al. 2003). More recently, using a YGOB-based approach, our laboratory was able to detect several hundred previously unannotated genes in yeast genome sequences, based on their conservation in other species (OhEigeartaigh et al. 2011).

1.3.2 Nonprotein-coding elements

Comparative genomics is also useful in identifying conserved DNA sequences that are either conserved regulatory elements, or that are functional genes for conserved noncoding RNA molecules. By surveying random DNA sequences of several yeast belonging to the *Saccharomyces sensu stricto* clade, comparative genomics aided the identification of promoter regulatory elements, nonprotein-coding RNA genes and small protein-coding genes. The sequences of some of these promoters and RNA genes was used to determine the degree of conservation of known functional elements within the genome (Cliften et al. 2001).

Another example is the secondary structure of the yeast telomerase RNA (TLC1). This sequence is highly divergent, even among closely related species: an alignment of four TLC1 sequences (from *S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus*) showed only 43% identity (Zappulla and Cech 2004). The structure of TLC1 consists of a central core and three arms that interact with other telomerase components and telomeric proteins. A comparison between six *Saccharomyces* yeasts and six *Kluyveromyces* yeasts has revealed the presence of conserved domains, CS3 and CS4, forming a central core in the RNA template (Tzfati et al. 2003; Lin et al. 2004). This would suggest that the RNA structure, more than conserved sequence, seems to play a major role in telomerase function (Chappell and Lundblad 2004).

A more recent example is the evolution of splicing patterns in yeast noncoding genes. Mitrovich *et al.* (2010) showed that in *C. albicans*, the exons of some spliced noncoding RNAs have no apparent function, but that their introns contain C/D box small nucleolar RNAs (snoRNAs) that are noncoding RNAs that target modifications to rRNA. However, in *S. cerevisiae*, nearly all snoRNAs arise from unspliced primary transcripts, and therefore require a splicing-independent processing pathway (Pieknа-Przybylska et al. 2007). This observation suggests that the transition of snoRNAs from intron sequences to unspliced and dedicated transcripts happened within the Saccharomycetaceae (Mitrovich et al. 2010). A search for orthologous snoRNAs in other species from the Saccharomycotina showed that Saccharomycetaceae species have fewer intronic snoRNAs (three to five) than non-Saccharomycetaceae species (23 to 33). These data can be explained by a massive loss of snoRNA-associated introns, most of which occurred in the common ancestor of the Saccharomycetaceae (Mitrovich et al. 2010). The introns appear to have been lost through degeneration of their splicing signals, a mechanism distinct from that which is proposed for protein-coding genes – a reverse transcription of spliced RNA, followed by homologous recombination that replaces the intron-containing genomic sequence with the intronless copy (Fink 1987).

1.3.3 Evolution on a genome-wide scale

Comparative genomics is also a powerful tool for elucidating a global genome-wide view of evolution. The analysis of the *S. cerevisiae* genome revealed the presence of an ancient WGD in an ancestor of the baker's yeast (Wolfe and Shields 1997). This WGD event has been verified by comparative genomics through the sequences of *L.*

waltii and *E. gossypii*. These two yeast species diverged prior to the WGD event and showed the predicted 1:2 mapping pattern between their genomes and pairs of duplicate blocks in *S. cerevisiae* (Dietrich et al. 2004; Kellis et al. 2004a). The comparison with *L. waltii* (Kellis et al. 2004a) showed that *S. cerevisiae* arose from complete duplication of eight ancestral chromosomes, and subsequently returned to functionally normal ploidy by massive loss of nearly 90% of duplicated genes in small deletions. These were balanced and complementary in paired regions, preserving at least one copy of virtually each gene in the ancestral gene set. In this study, 145 paired regions in *S. cerevisiae* were identified, tiling 88% of the genome and containing 457 duplicated gene pairs. More recently, through the use of YGOB, paired regions covering essentially all the genome have been identified and the number of ohnolog gene pairs has been increased to 551 (Byrne and Wolfe 2005). Another study showed that in *C. glabrata*, 18 of the 20 duplicated blocks are shared with S. cerevisiae (Dujon et al. 2004). This is consistent with the fact that *C. glabrata* diverged after the WGD.

1.3.4 Changes in metabolic pathway

The study of the ohnologs (gene pairs formed by genome duplication) present in *S. cerevisiae* (and only in a single copy in non-WGD species) showed that the genome duplication played a direct role in the adaptation of the *S. cerevisiae* lineage towards fermentation (Piskur 2001; Wolfe 2004; Piskur et al. 2006; Conant and Wolfe 2007). *S. cerevisiae* developed the ability to grow under anaerobic conditions, provided that glucose is available. Transcription of many ohnolog pairs is differentially regulated by oxygen, with one member of the pair being induced in hypoxic conditions whereas the other is induced in aerobic conditions (Kwast et al. 2002). Moreover, during its evolution, *S. cerevisiae* has gained sophisticated regulatory mechanisms to sense fluctuating levels of glucose, and as a result has many different glucose-sensing and signalling pathways (Gelade et al. 2003). The WGD event greatly increased the ability of *S. cerevisiae* to respond to glucose, with the presence of many ohnologs in those different pathways (Wolfe 2004). Moreover, all the WGD lineages (and some non-WGD species such as *Lachancea kluyveri*, *Dekkera bruxellensis* and *S. pombe*) developed the "Crabtree effect" which enabled it to use the fermentation pathway when oxygen is available; this change allows *S. cerevisiae* to monopolize extant sugar resources and to kill competing microorganisms by poisoning the environment with ethanol (Thomson et al. 2005; Piskur et al. 2006; Conant and Wolfe 2007).

Furthermore, a comparative study of four *Schizosaccharomyces* species (*S. pombe*, *S. octosporus*, *S. cryophilus* and *S. japonicus*) showed that, compared to budding yeast, fission yeast has lost several genes involved in the glyoxylate cycle, in glycogen biosynthesis, and paralogs of glycolytic genes (Rhind et al. 2011). These adaptations are consistent with the incapacity of fission yeast to consume ethanol as a sole carbon source. Unlike *S. cerevisiae*, in the *Schizosaccharomyces* species, the expression of the genes encoding the pyruvate dehydrogenase complex and *ADH1* is reduced as glucose is depleted, which prevents the effective use of pyruvate for respiration. Thus, the lack of efficient ethanol catabolism in fission yeast demonstrates that aerobic fermentation (Crabtree effect) did not evolve to create a consumable by-product, and that in fission yeast ethanol is a waste product.

1.3.5 Evolution of yeast lifestyle

Comparative genomics has provided insights into the sexual cycle of *Candida* species, and the evolution of mating systems. In *S. cerevisiae*, mating type is determined by two different alleles (*MATa* and *MATα*) of the *MAT* locus. The baker's yeast has gained the ability to switch some haploid cells in a colony from one mating type to another (Haber 1998). This switch has been described by the cassette model, which states that this change happens by a gene-conversion process that involves replacing the genetic information at *MAT* with information copied from one of two silent cassette loci, *HMLα* or *HMR*a (Hicks and Strathern 1977). This system evolved after the gain of the silent cassettes then the acquisition of the HO endonuclease from a mobile genetic element (Butler et al. 2004). In the *Candida* lineage, the transcription factor gene *IME1*, the major regulator of meiosis in *S. cerevisiae*, is absent (Tzung et al. 2001; Bennett 2009). However, some species, such as *C. lusitaniae*, possess a meiotic cycle (Reedy et al. 2009), whereas in *C. albicans*, the completion of the sexual cycle occurs by a parasexual mechanism of random chromosome loss rather than conventional meiosis (Forche et al. 2008). *C. albicans* either has a cryptic meiotic program that has yet to be identified, or conserved "meiosis-specific" genes have been reprogrammed to function in the parasexual cycle (Bennett 2009). In the sexual *Candida* species, drastic changes have been noted in the pathways involved in meiotic recombination, and in the mechanisms of chromosome pairing and crossover formation (Butler et al. 2009). These

observations indicate that meiosis in sexual *Candida* species is fundamentally distinct from that in *S. cerevisiae*, in terms of both its regulation and its molecular apparatus.

1.4 Ribosomal DNA

Ribosomes are macromolecular complexes that direct protein synthesis from messenger RNAs. They are formed by a combination of ribosomal RNA molecules (rRNAs) and ribosomal proteins. These rRNAs are often characterized in sedimentation velocity units (S, for Svedburg). The ribosome is composed of two major subunits – one large and one small. In prokaryotes, the large subunit is composed of one 5S RNA, one 23S RNA, and about 30 proteins. The small subunit is composed of one 16S RNA and 21 proteins. In eukaryotes, the large subunit is composed of one 5S RNA, one 25-28S RNA, one 5.8S RNA and about 40 proteins. The small subunit is composed of one 18S RNA and about 30 proteins.

Due to their presence in every living organism and the presence of highly conserved regions in their sequences, the rRNA genes have been very useful in phylogenetic studies (Hillis and Dixon 1991; Soltis and Soltis 2000; Bayha et al. 2010; Caccio et al. 2010). The organization on the rRNA transcription unit (rDNA array, see below) and of the 5S rDNA is also a marker of phylogeny, some particular rDNA organizational characteristics being conserved among related species (Torres-Machorro et al. 2010).

1.4.1 Ribosomal RNA genes

In prokaryotes, the ribosomal RNA genes are present in one or more copies. The genes are often organized in a single operon (in that case, genes are often separated by tRNA genes), or can be dispersed through the genome. In eukaryotes, the 28S, 18S and 5.8S genes are mainly organized in one or more large tandem arrays that can contain hundreds of individual rDNA units, occupying millions of base pairs (**Figure 1.4** (Nei and Rooney 2005)). Each repeat unit is transcribed (by RNA polymerase I) into a large ("45S") precursor RNA that is subsequently cleaved to yield the mature 18S, 25S and 5.8S rRNA molecules. The transcribed intergenic regions between the 18S gene and the 5.8S gene, and between the 5.8S gene and the 25S gene, are called Internal Transcribed

Spacers (ITS1 and ITS2 respectively). Regions at the 5' and 3' extremities of the 45S transcript, which are discarded during rRNA maturation, are called External Transcribed Spacers (ETS1 and ETS2). In yeast species, but not in most other eukaryotes, the 5S rRNA gene usually also forms part of the repeating rDNA unit. The 5S gene is usually in opposite orientation to the other rRNA genes, is transcribed by RNA polymerase III, and does not form part of the 45S primary transcript. The regions flanking the 5S gene are called Intergenic Spacers (IGS1 and IGS 2). In other eukaryotes, the 5S gene can be present inside the array or located elsewhere in the genome or one can find a combination of both, as in humans. This rDNA array is repeated in tandem (**Figure 1.5**).

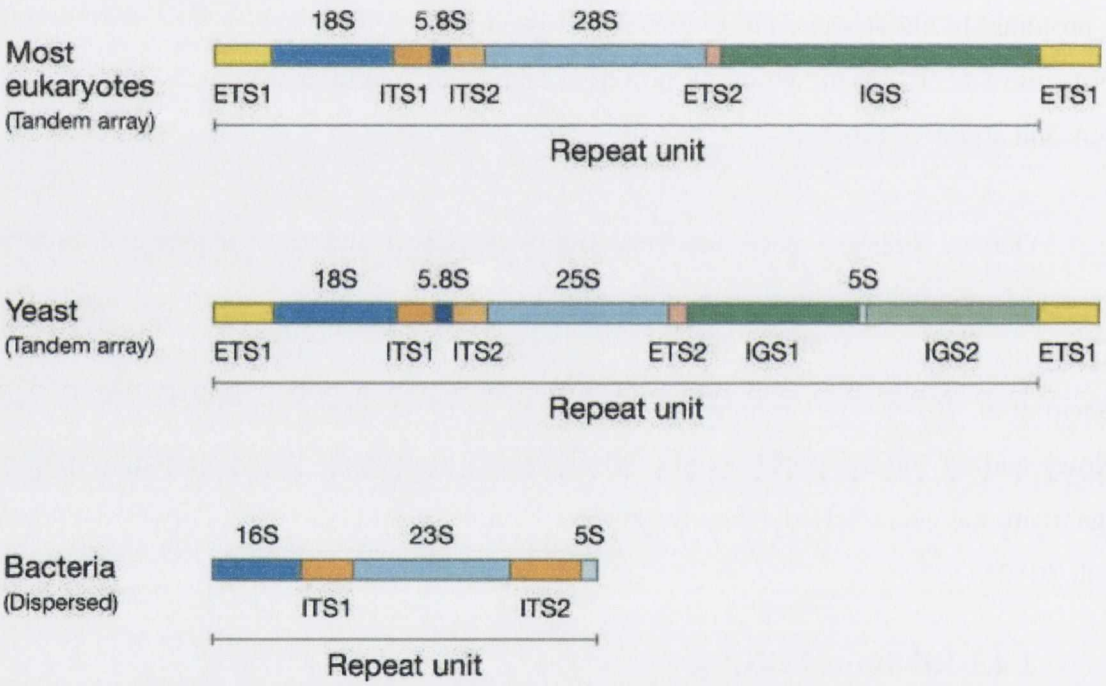


Figure 1.5 Different structures of rDNA genes (Nei and Rooney 2005). In bacteria the 3 rDNA genes (16S, 23S and 5S) are clustered together and repeated in tandem. In eukaryotes, the 5S gene can sometimes be part of the repeated rDNA array (for example in *S. cerevisiae*) or located somewhere else in the genome. ETS = External Transcribed Spacers, ITS = Internal Transcribed Spacers, IGS = Intergenic Spacers.

In *S. cerevisiae*, the rDNA array is located on chromosome XII. The rDNA cluster is composed of approximately 150 copies and occupies approximately 60% (1.5 Mb) of the chromosome and 10% of the total genome (Johnston et al. 1997). In yeasts from the Saccharomycotina, the length of the rDNA array unit is between 7.7 and

12.5 kb, and the copy number between 50 and 200 copies (Maleszka and Clark-Walker 1993; Wendland et al. 1999; Torres-Machorro et al. 2010). In species where all the 5S genes are linked to the rDNA array, the rDNA/5S rDNA copy ratio is equal to one. In *S. pombe*, in which the 5S gene is not linked to the rDNA array, this ratio is thought to be between 3.3 and 4 (the number of rDNA copies is comprise between 100 and 120) (Torres-Machorro et al. 2010). A study showed that in eukaryotes, a positive correlation exists between the genome size and rDNA copy number (Prokopowich et al. 2003).

A study of the location of the 5S gene in the Ascomycota phylum showed that 5S genes are within the rDNA arrays in all the studied species in the Taphrinomycotina subphylum, almost all the species in the Pezizomycotina subphylum, and in the basal Saccharomycotina species *Yarrowia lipolytica*, whereas in the other Saccharomycotina species the 5S genes are present in the rDNA units (Bergeron and Drouin 2008). This meant that having the 5S genes outside the rDNA array is likely to be the ancestral condition of Ascomycetes; the insertion of 5S gene into the rDNA array occurred after the divergence of other Saccharomycotina species from *Y. lipolytica*.

An interesting hypothesis to explain the pattern of 5S gene organization and dispersal in these fungal species is that 5S genes may be able to multiply and integrate around the genome through a process similar to retroposition (Rogers 1985; Brosius 1991). The 5S genes and the short interspersed elements (SINEs) have the same type of internal RNA polymerase III promoter (Paule and White 2000; Weiner 2002). Moreover, a novel class of SINE has been described in the zebrafish genome, that may have arisen from the fusion of a 5S gene or pseudogene and a LINE (Kapitonov and Jurka 2003), showing that 5S genes and retroelements can interact with one another. Furthermore, some plant retrotransposons named *Cassandra* carry conserved 5S RNA sequences and associated RNA polymerase III promoters and terminators in their long terminal repeats (LTRs) (Kalendar et al. 2008). Transposition of this element would be able to amplify 5S sequences in plant genomes.

1.4.2 Evolution of rDNA array in eukaryotes

A study published in 2007 on the evolution of rDNA arrays in five fungal species showed that the number of nucleotide sequence polymorphisms within the rDNA arrays of those five species was very low (Ganley and Kobayashi 2007). The 5S

gene was present in the rDNA arrays of four of the five species (*S. cerevisiae*, *S. paradoxus*, *E. gossypii* and *Cryptococcus neoformans*). No bias of polymorphisms toward the IGS regions (the least selective constrained regions of the rDNA) was detected. This suggests the rDNA array evolves via concerted evolution (Brown et al. 1972; Zimmer et al. 1980), a mechanism that makes repeated genes within a species more similar to each other, in structure and number, than they are to the orthologous repeated genes of other species. The molecular process responsible for concerted evolution is named homogenization (Dover 1982), a process that implies unequal crossover (Szostak and Wu 1980) or gene conversion (a form of nonreciprocal recombination in which a DNA segment of a recipient gene is copied from a donor gene) (Jeffreys 1979). Holliday junctions (Holliday 1964) were detected on a two-dimensional gel analyses of rDNA fragments from *S. cerevisiae* genome, during mitotic growth of yeast cells (Zou and Rothstein 1997). So homogenization could occur by homologous recombination between tandem repeats. However, a recent study using a *S. cerevisiae* strain lacking histone chaperone Asf1 (involved in multiple replication and repair processes) showed that these yeasts undergo reproducible rDNA repeat expansions that do not require the replication fork blocking protein Fob1, and are therefore independent of known rDNA expansion mechanisms (Houseley and Tollervey 2011). The mechanism of repeat gain is surprisingly independent of the homologous recombination proteins Rad52, Rad51 and Rad59.

To conclude, rDNA arrays evolve through concerted evolution (**Figure 1.6**). The identity of the sequences is maintained by homogenization, which could occur by homologous recombination between tandem repeats. Homologous recombination is very frequent within rDNA, and is tightly linked to DNA replication of the tandem arrays. However, there are exceptions to this rule. The first well-characterized case is in the protozoan *Plasmodium*. Studies of the genomes of *P. berghei, P. falciparum* and *P. vivax* showed that the rDNA genes, which are dispersed in different chromosomes, exist as two classes with very different nucleotide sequences, that are expressed at different stages of the parasite's life cycle (Gunderson et al. 1987; Waters et al. 1989; Qari et al. 1994). The gene sequences are similar within each class but are not similar between different classes. It was concluded that the rDNA genes in *Plasmodium* evolved by the birth-and-death model (Rooney 2004). Birth-and-death is an evolutionary model in which new genes are created by repeated gene duplication and some duplicate genes are

maintained in the genome for a long time but others are deleted or become non-functional by deleterious mutations (Nei et al. 1997) (**Figure 1.6**). The existence of several different classes of rDNA genes had also been documented in some oak species (Muir et al. 2001) and worms (Carranza et al. 1996), as well as in the hybrid yeast *Saccharomyces pastorianus* (Nakao et al. 2009).

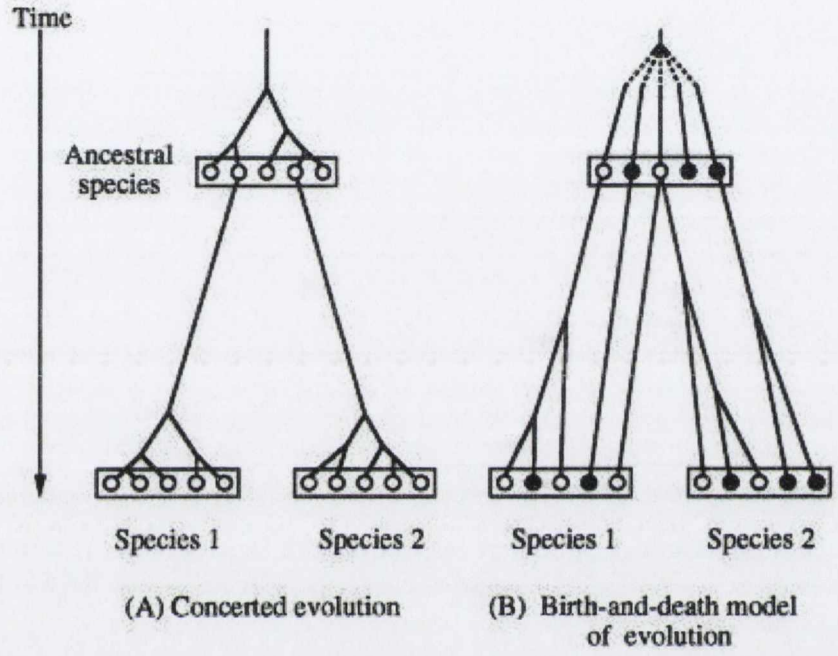


Figure 1.6 Concerted evolution versus Birth-and-death model of evolution (Nei et al. 1997). The white dots represent functional genes, the black dots pseudogenes.

1.4.3 Evolution of solo 5S genes (outside of the rDNA array) in eukaryotes

5S genes can be found dispersed in the genome, sometimes amplified in tandem repeats (separate from the main rDNA array) located at one or several places in the genome. For example, there are three clusters of 5S genes in *Drosophila virilis*, whereas *D. melanogaster* has only one cluster of about 160 tandemly repeated 5S genes (Kress et al. 2001). The 5S gene family can include up to 24,000 members in *Xenopus* (Brown and Sugimoto 1974). It has been found in several species that 5S genes, when not present in the rDNA array, is linked to other tandemly repeat gene families such as those coding for histones (in some arthropods), trans-splicing leaders (in some nematodes and protozoan species), ubiquitins (in the ciliate protozoa *Tetrahymena*

pyriformis) and small nuclear RNA genes (for example in the fish *Solea senegalensis* and in the isopod crustacean *Asellus aquaticus*) (Guerreiro et al. 1993; Drouin and de Sa 1995; Pelliccia et al. 2001; Manchado et al. 2006) (**Figure 1.7**).

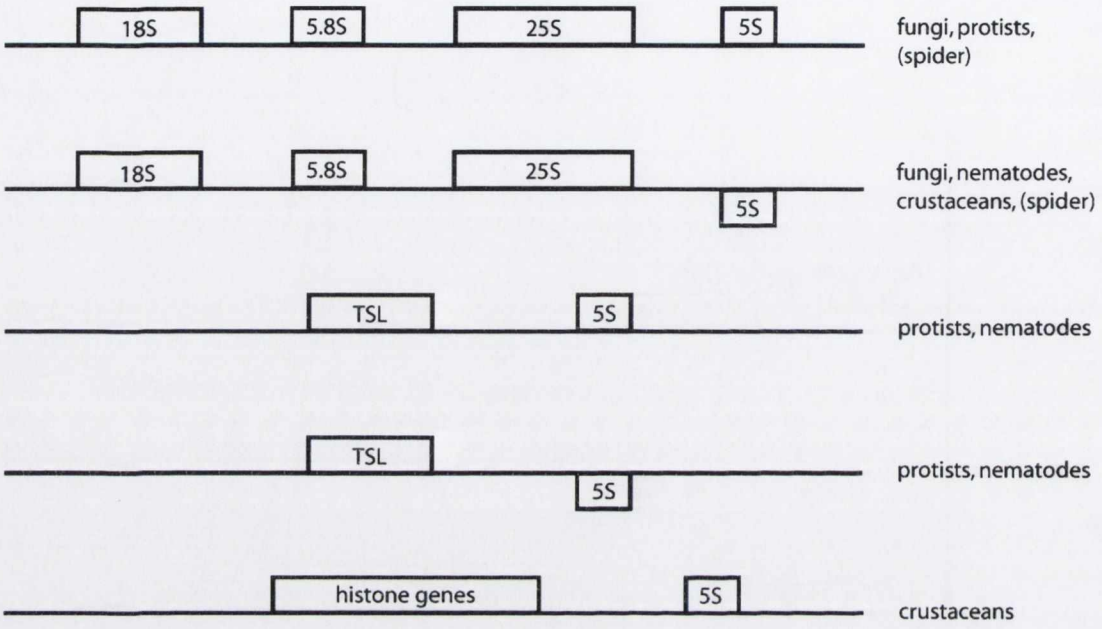


Figure 1.7 The arrangement of 5S genes within the tandem repeat units of various multigene families in different taxonomic groups (Drouin and de Sa 1995). TSL, tans-splicing leader.

However, here again there are exceptions to the rule of concerted evolution. A study of the 5S genes in some *Oryza* species showed characteristics inconsistent with concerted evolution: multiple distinct 5S rDNA types were detected within a species, and were also shared among species, and intraspecific nucleotide diversity was detected within a 5S rDNA class (Zhu et al. 2008). Furthermore, a study on four filamentous fungi within the ascomycete subphylum Pezizomycotina showed that 5S genes located at different loci were more similar to other 5S genes in other species than 5S genes in the same species (Rooney and Ward 2005). The number of copies of 5S genes in these species varies from 50 to 100, and they are dispersed in the genome rather than organized as a tandem array. Moreover, in the four studied genomes, a great portion of the total gene number was comprised of truncated 5S pseudogenes. These results suggest that instead of evolving via concerted evolution, like the large rDNA arrays, the 5S families in these species evolve via birth-and-death evolution under strong purifying selection.

1.4.4 rDNA as a lifespan determinant

In *S. cerevisiae*, it was shown that extrachromosomal rDNA circles (ECR) derived from the genomic rDNA by recombination accumulated in the mother cell, and it was proposed that this accumulation induced senescence (Johnson et al. 1999). ERCs are able to replicate via an origin of replication (ARS) contained within the rDNA repeat, and are preferentially segregated to mother cells during division. However, Ganley et al. (2009) established a yeast strain in which ERCs did not duplicate and did not accumulate in the mother cell. But even in the absence of ERCs, the life span of this strain was shortened (Ganley et al. 2009). The cause of this shortening was in fact rDNA instability, triggered by a high recombination rate, which was only detectable in the mother cell. rDNA instability is an upstream event of ERC accumulation and has a more direct impact on lifespan shortening (Kobayashi 2011). A model regarding the role of the rDNA in aging was proposed by Kobayashi in 2008. In this model, stability and copy number of the rDNA regulate cellular functions such as senescence and damage resistance (Kobayashi 2008).

Two genes are required for rDNA stability-dependent life span: *SIR2*, which encodes the NAD^+-dependent histone deacetylase, and *HDAC*, a repressor of the noncoding transcription that activates rDNA recombination as part of the amplification process (Kobayashi 2011). In contrast, the *FOB1* gene encodes a protein that inhibits the replication fork and causes double-strand breaks to trigger recombination in the rDNA (Kobayashi et al. 2004) (**Figure 1.8**). rDNA instability is known to increase the numbers of ERCs (Kaeberlein et al. 1999): yeast strains mutated in *SIR2* gene accumulates more ERCs because of the high recombination rate and have a very shortened life span.

In *S. cerevisiae*, about half the rDNA copies are not transcribed (Dammann et al. 1993). Yeast strains without these extra copies show genomic instability and became sensitive to DNA damage (Ide et al. 2010). This sensitivity is dependent on rDNA transcriptional activity: the ratio of actively transcribed rDNA genes increases when the copy number reduces. The rDNA transcriptional activity interferes with cohesion between rDNA loci of sister chromatids. The extra rDNA copies facilitate condensin

association and sister-chromatid cohesion, thereby facilitating equal recombinational repair.

In human, the genes responsible for some premature aging diseases, such as the *WRN* gene in Werner syndrome, have been identified: they are RecQ homologs that repair damaged DNA (Yu et al. 1996; Shen and Loeb 2001). The yeast homolog of the *WRN* gene is *SGS1*. The *SGS1* helicase suppresses recombination in genomic DNA, including the rDNA (Watt et al. 1996). Yeast cells mutated in *SGS1* exhibit an increase rate of recombination at the rDNA locus, contain a higher proportion of extrachromosomal rDNA circles and age prematurely compared to wild-type (Watt et al. 1995; Sinclair and Guarente 1997; Sinclair et al. 1997).

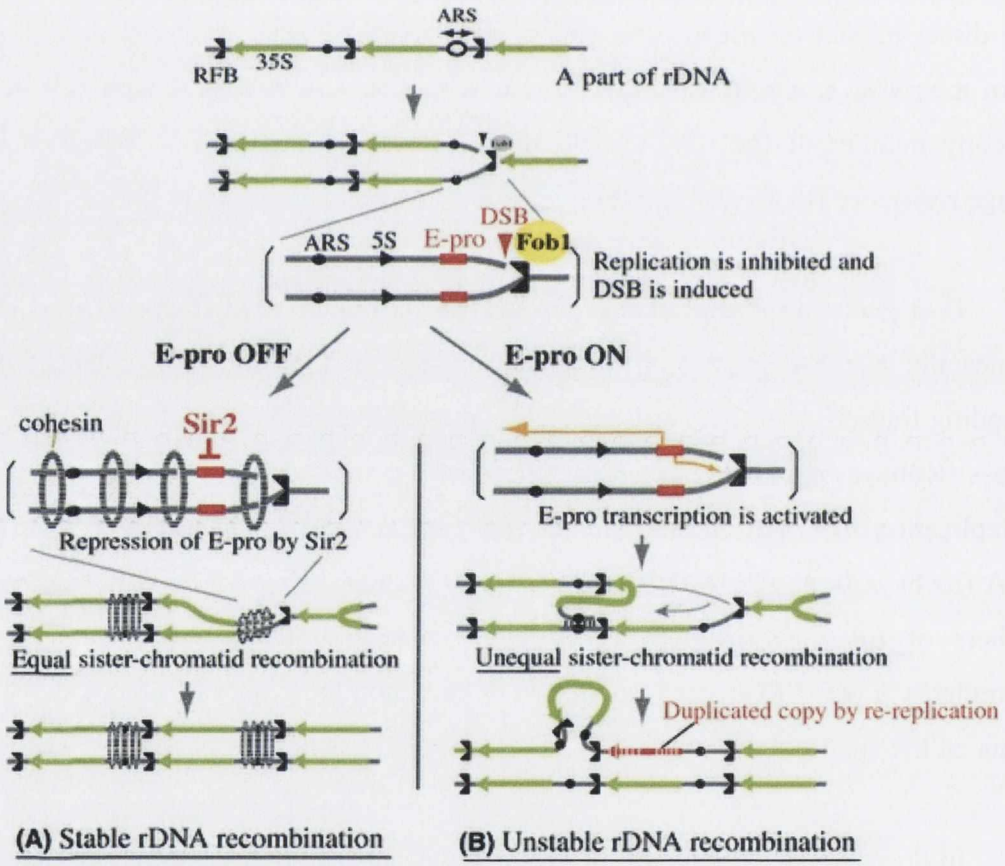


Figure 1.8 rDNA recombination for copy number change (Kobayashi 2011). In the S-phase of the cell cycle, by the function of Fob1, the DNA replication forks in the rDNA are inhibited at the RFB and DNA double-strand breaks (DSB) are induced. The broken ends are repaired by homologous recombination via sister chromatids. E-pro is a bidirectional noncoding promoter that functions in the regulation of rDNA stability (copy number). (A) In situations where copy number is normal, the silencing protein, Sir2, represses E-pro transcription, allowing the cohesin protein complex (dotted ellipse) to associate with the IGS. DSBs are repaired by equal sister-chromatid recombination, with no change in rDNA copy number. (B) In situations where copy number is reduced, Sir2 is not functioning and E-pro is activated. This E-pro

transcription displaces cohesin from the IGS. The lack of cohesion means that unequal sister chromatids can be used as templates for the repair of DSBs, resulting in changes in rDNA copy number. The gray lines represent single chromatids (double-strand DNA).

1.5 New sequencing technologies and genome annotation

During the past few years, the number of genome sequencing projects has increased drastically, due to the development of next-generation sequencing (NGS) technologies. These methods can produce large amounts of sequence data quickly and inexpensively compared to conventional Sanger methods. As a result, it is almost impossible to estimate the number of genome projects currently in progress. However, the error profiles and the limitations of NGS technologies differ significantly from those of previous sequencing technologies (Kircher and Kelso 2010), and in general they are more error-prone than Sanger sequencing.

1.5.1 New sequencing technologies and their error rates

The two major NGS technologies in use today are Illumina (formerly called Solexa) and Roche (formerly called 454). The Illumina platform is based on massively parallel sequencing of millions of fragments using a reversible terminator technology that employs a sequencing-by-synthesis concept similar to that used in Sanger sequencing. It generates highly accurate short (50 bp on average) reads that are used for resequencing applications (*i.e.*, sequencing when an assembly can be done using a highly similar reference genome). Paired sequence reads are possible using the Illumina technology, but the maximum distance between pairs is limited to approximately 3 kb. The Roche platform is based on pyrosequencing technology (Ronaghi et al. 1996; Margulies et al. 2005). It generates longer reads (400 bp on average) that are less accurate. Its longer read length and its ability to obtain data from read-pairs that are separated by relatively large distances in the genome(8-20 kb) make the Roche platform the system of choice for *de novo* genome sequencing (assembly without a reference sequence). The other new sequencing platforms in used today are SOLiD from Applied Biosystems (Pandey et al. 2008), HeliScope from Helicos (Hart et al. 2010), and the Pacific Bioscience platform (Korlach et al. 2010).

In this thesis, I use *de novo* assemblies of new yeast genomes from the Roche platform. The average substitution (excluding insertion/deletion, indel) error rate for this technology has been reported to be in the range of 10^{-3} to 10^{-4} (Kircher and Kelso 2010). This is higher than the rates observed for Sanger sequencing (10^{-5} to 10^{-6}), but lower than the substitution error rate for the Illumina platform (10^{-2} to 10^{-3}) and for other NGS technologies (Kircher and Kelso 2010). Most of the errors observed with the Roche platform are small indels (Quinlan et al. 2008), particularly in homopolymer tracts.

1.5.2 Genome annotation

With the development of next-generation sequencing technologies, the need for automated annotation has become urgent. Several automatic genome annotation tools have been developed, based on the identification of orthologous genes among species (Bryson et al. 2006). The detection of homologs is usually based on sequence similarity, calculated with one of the programs of the BLAST suite (Altschul et al. 1990), using single-directional best hit or bi-directional best hit annotation methods. In addition, some programs consider other parameters: synteny (Vallenet et al. 2006), or textual annotation similarity (once matches to the target gene are selected by sequence similarity, the next step is to compare textual annotation information of these matches by using cosine similarity: in this case the annotations are transformed in vectors of words. Each word in texts defines a dimension in Euclidian space and the frequency of each word corresponds to the value in the dimension. The cosine of 2 vectors is defined as $\cos(\vec{t_1}, \vec{t_2}) = \frac{\vec{t_1}.\vec{t_2}}{\|\vec{t_1}\| \|\vec{t_2}\|}$ and is used as a similarity of 2 annotations) (Yang et al. 2010). Some annotation pipelines refine candidate regions originally found by BLAST using other methods (for instance CEGMA (Parra et al. 2007) refines gene structures using a combination of GeneWise (Birney et al. 2004), HMMER (Durbin et al. 1998) and GeneID (Parra et al. 2000)). These tools can be either quite specific (organellar genomes, microbial genomes) or are able to work both on eukaryotic and prokaryotic genomes. Moreover, until recently, among all the automatic annotation tools, only one uses synteny data: MaGe (Vallenet et al. 2006), which uses an exact graph-theoretical approach and using all available microbial proteomes (Boyer et al. 2005).

Schneider et al. (2010) recently released RAPYD – Rapid Annotation Platform for Yeast Data, an annotation tool that has been developed specifically for yeast species. RAPYD used three software modules for yeast annotation, metabolic pathway reconstruction, and comparative genomics. In the annotation module, the gene structures are mainly predicted by AUGUSTUS (Stanke and Waack 2003). Gene function predictions are based on (i) BLAST comparisons on several databases: nr and nt (NCBI), Swiss-Prot (Boutet et al. 2007), KEGG (Ogata et al. 1999), KOG (Tatusov et al. 2003), CDD (Marchler-Bauer et al. 2009) and others with a conservative E-value threshold of 1e-30; and (ii) Hidden Markov Model (HMM) based sequence analyses using the PFAM (Finn et al. 2010) and TIGRFAMS (Haft et al. 2003) databases. Recently, new features were added to Augustus, such as information on gene and transcript annotations from closely related species syntenically mapped to the target genome, if available (Stanke et al. 2008). In **Chapter 2** I develop an automated annotation pipeline for yeast genome sequences and compare its results to those obtained using AUGUSTUS.

Chapter 2

YGAP: the Yeast Genome Annotation Pipeline

2.1 Abstract

Yeasts are a model system for exploring eukaryotic genome evolution. Next-generation sequencing technologies are poised to vastly increase the number of yeast genome sequences, both from resequencing projects (population studies) and from *de novo* sequencing projects (new species). However, the annotation of genomes presents a major bottleneck for *de novo* projects, because it still relies on a process that is largely manual. Here we present the Yeast Genome Annotation Pipeline (YGAP), an automated system designed specifically for new yeast genome sequences lacking transcriptome data. YGAP does automatic *de novo* annotation, exploiting homology and synteny information from other yeast species stored in the Yeast Gene Order Browser (YGOB) database. The basic premises underlying YGAP's approach are that data from other species already tells us what genes we should expect to find in any particular genomic region and that we should also expect that orthologous genes are likely to have similar intron/exon structures. Additionally, it is able to detect probable frameshift sequencing errors and can propose corrections for them. YGAP searches intelligently for introns, and detects tRNA genes and Ty-like elements. In tests on *Saccharomyces cerevisiae* and on the genomes of *Naumovozyma castellii* and *Tetrapisispora blattae* newly sequenced with Roche-454 technology, YGAP outperformed another popular annotation program (AUGUSTUS). For *S. cerevisiae* and *N. castellii*, 91-93% of YGAP's predicted gene structures were identical to those in previous manually curated gene sets. YGAP has been implemented as a webserver with a user-friendly interface at http://wolfe.gen.tcd.ie/annotation

2.2 Introduction

More genomes have been sequenced from yeast species than from any other group of eukaryotes. Yeasts provide an excellent system for exploring eukaryotic genome evolution by comparative genomics because their genomes are compact (9-20 Mb with 4700-6500 genes) with few introns, making them straightforward to sequence, but they still retain extensive synteny across deep phylogenetic distances (Liti and Louis 2005; Wolfe 2006; Dujon 2010). Moreover, there is a wealth of information about gene functions in *Saccharomyces cerevisiae*, probably the most extensively-studied model organism in the world (Christie et al. 2009).

Yeast comparative genomics has produced many insights into genome evolution, including the discovery of whole-genome duplication (WGD) (Wolfe and Shields 1997); development of methods for identifying conserved regulatory elements and RNA genes (Cliften et al. 2001; Kellis et al. 2003); exploration of changes in the genetic code (Butler et al. 2009); and detection of horizontal gene transfer and its functional consequences (Hall et al. 2005; Rolland et al. 2009). Furthermore, comparative genomics has played a major role in gene discovery and improving the quality of genome annotations. For example, a comparative analysis of four closely related *Saccharomyces* species (Kellis et al. 2003) led to a revision of the previous annotation of the *S. cerevisiae* genome: elimination of previously annotated ORFs, redefinition of start and stop codons, and discovery of new introns. A similar approach was conducted with the pathogenic basidiomycete yeast *Cryptococcus neoformans*, responsible for cryptococcal meningitis (Tenney et al. 2004).

The need for automated annotation has become urgent with the development of next-generation sequencing technologies, but annotating genomes remains a challenge and still relies on a process that includes many manual steps (Bryson et al. 2006; Souciet et al. 2009). Annotation can be viewed as consisting of two primary steps: inferring gene structures and making decisions about the orthology or paralogy relationships between these genes and genes in other species. Yeast genomes present an unusual set of circumstances at both of these steps. The first step, inferring gene structures, is very simple for most yeast genes because they are intronless. However,

accurate identification of the coordinates of the other ~4% of yeast genes that have introns is difficult unless cDNA information is available, particularly because many yeast introns are very close to the gene's start codon (Spingola et al. 1999; Miura et al. 2006). Some previous automated approaches to annotation of yeast genomes either ignored all introns (Kellis et al. 2004b), or used generic fungal gene models that resulted in the over-prediction of hundreds of nonexistent introns (Jeffries et al. 2007). The second step, classifying genes as orthologs or paralogs of genes in other species, is often ignored by automated approaches. They typically use BLAST (Altschul et al. 1990) to identify unidirectional or bidirectional best hits between genes in the new genome and a reference database, and then annotate genes as 'similar to' genes in other species, or as members of particular gene families, without an explicit statement about whether the authors consider the interspecies relationship to be an orthologous one. Decisions about orthology versus paralogy are important because, in general, orthologs tend to have conserved gene function whereas paralogs often diverge (Studer and Robinson-Rechavi 2009). For this reason, manual annotators and scientists working on specific genes usually want to identify orthologs between species, and these orthology decisions frequently make use of synteny information. In yeasts, orthology relationships are complicated by a WGD event in the common ancestor of several species, leading to a 2:1 synteny relationship between genomic regions in post-WGD and non-WGD species (Dietrich et al. 2004; Kellis et al. 2004a). Among all the automatic annotation tools currently available, only a few use synteny data (Vallenet et al. 2006; Stanke et al. 2008) and none consider WGD. Until recently, none had been developed specifically for yeast species (Schneider et al. 2010).

In 2005 our laboratory developed the Yeast Gene Order Browser (YGOB), which is a database and interface for comparative genomics for yeasts in the family Saccharomycetaceae (Byrne and Wolfe 2005). A major strength of YGOB is that it contains manually curated sets of orthologs (and WGD-derived paralogs in species that underwent WGD), which have been identified based on their conserved synteny relationships. We recently embarked on a project to sequence the genomes of multiple previously unstudied yeast species from the family Saccharomycetaceae by Roche-454 sequencing and *de novo* assembly. We reasoned that the information in YGOB could be used to automatically annotate the new yeast genomes with accuracy comparable to a manual annotation and here we present YGAP (Yeast Genome Annotation Pipeline),

the pipeline that carries out this task. The data input to YGAP are the entire YGOB database, the scaffold sequences from the new species, and (if available) its contigs and individual sequence reads. The output includes a set of annotation files, in a format compatible with YGOB, and a 'mini-YGOB' interface that allows gene order in the new species to be compared to others. To test our pipeline we used the genome of the extensively studied *S. cerevisiae* as well two of the new genomes from our sequencing project, *Naumovozyma castellii* and *Tetrapisispora blattae*.

2.3 Methods

2.3.1 Input data

After assembly, next-generation sequencing projects can produce three different types of output DNA sequence files: (i) a 'reads' file containing all the primary sequence reads; (ii) a 'contigs' file containing the contigs assembled from overlapping reads; and (iii) a 'scaffolds' file, typically made by concatenating together those contigs whose relative order and orientation is known, separated by runs of 'N' bases representing the estimated lengths of unsequenced gaps. For example, our genome project for *Namovozyma castellii* generated 1.4 million reads. The Celera assembler (Koren et al. 2010) assembled these reads into 3851 contigs, and arranged 43 of these contigs into 10 scaffolds that correspond to almost complete chromosomes (none of the unincorporated contigs is larger than 2.1 kb).

To run YGAP, the user must provide a scaffolds file from the new species. This is the only sequence file whose input is mandatory, but if contigs and reads files are also available YGAP can use these for optional steps. The user must also specify whether the new genome comes from a post-WGD or a non-WGD species (Byrne and Wolfe 2005).

YGAP also requires access to the YGOB database. This database (Byrne and Wolfe 2005) consists of previously annotated yeast genome sequences, and lists of the gene sets that comprise each of its ~9500 homology pillars. A pillar consists of a manually curated set of genes that are orthologs, or paralogs resulting from WGD. YGOB contains data from 11 species: *S. cerevisiae* (Goffeau et al. 1996), *S. bayanus* (Cliften et al. 2003; Kellis et al. 2003), *N. castellii* (formerly called *S. castellii*) (Cliften

et al. 2003), *Vanderwaltozyma polyspora* (Scannell et al. 2007b), *Candida glabrata* (Dujon et al. 2004), *Zygosaccharomyces rouxii* (Souciet et al. 2009), *Kluyveromyces lactis* (Dujon et al. 2004), *Eremothecium gossypii* (Dietrich et al. 2004),*Lachancea kluyveri* (Souciet et al. 2009), *L. thermotolerans* (Souciet et al. 2009), and *L. waltii* (Kellis et al. 2004a).

In the tests of YGAP described here, to avoid circular reasoning we omitted *S. cerevisiae* data from YGOB pillars when annotating the *S. cerevisiae* genome, and we omitted *N. castellii* data (from the draft sequence of the genome (Cliften et al. 2003)) when annotating our sequence of the *N. castellii* genome.

2.3.2 Checking the integrity of scaffolds

If the input data include primary 'paired-end' sequence reads that are expected to be close together in the genome, and also include contig data as well as scaffolds, YGAP will report on the consistency between the primary data and the scaffold structure. Using BLASTN, it maps each read from a pair onto a contig, provided that it has a unique hit in the genome. Pairs of reads that map to different contigs identify possible physical connections between contigs, which should correspond to the scaffold organization deduced by the assembly program. YGAP summarizes these data in the form of a matrix listing the number of read-pairs that support a connection between any two contigs, and lists the contigs in the order that they occur in the scaffolds. This allows the user to see the amount of support for any connection between two contigs in the scaffold structure, and the support for possible alternatives.

2.3.3 Locating genes

In an initial annotation step, tRNAscan-SE (Lowe and Eddy 1997) (run with default options) is used to detect and annotate tRNA genes. In the later steps of annotation, no protein-coding gene will be allowed to overlap with a tRNA gene; in this case, the gene will be tagged for manual correction.

Annotation of protein-coding genes is largely based on TBLASTN searches. We use every protein from every YGOB pillar as a query in a TBLASTN search against the genome (scaffolds file), and initially store all hits with E < 1e-05. For each pillar, we then identify the place in the genome where that pillar has its strongest hit, and other

places where it has weaker hits. That is, among the proteins encoded by the pillar (*P*) we identify the query protein (*Q*) that gives the lowest TBLASTN E-value (*E*) versus the genome and store the location of that hit. It is likely that query *Q* comes from the species that is most closely related to the new genome. We also store the location of weaker hits between *Q* and the genome, provided that the exponent of their E-values is lower than –30 and lower than *E*/2 (that is, if the strongest hit's E-value was 1e-100 we would retain other hits with E-values < 1e-50). For each stored location for query *Q*, we build a gene model as described later below.

Several YGOB pillars can match the same location in the new genome due to the existence of paralogous genes. Thus two or more pillars might hit identical or overlapping regions of the genome, and gene models would initially be constructed for both of them. After the TBLASTN searches are complete, we use synteny information to determine which pillar is the correct match for this genomic location. Specifically, if locus *L* in the new genome is hit by queries from two pillars *P1* and *P2*, we examine the regions of the new genome upstream and downstream of *L* and identify neighboring pillars (*Pleft* and *Pright*) that have been mapped unambiguously to these flanking regions. We then ask whether, in the Ancestral genome (Gordon et al. 2009) or in *S. cerevisiae*, *P1* or *P2* occurs in the interval between *Pleft* and *Pright*; if this is true for *P1* but not *P2*, we assign locus *L* to pillar *P1* and discard *P2* as a candidate for *L*. Note that this assignment is based on synteny, without regard to the TBLASTN E-values for *P1* and *P2*.

When a post-WGD species is being analyzed, the software allows two candidate genomic loci to be assigned to a single pillar, whereas a non-WGD species can only ever have one locus assigned to a pillar. Where a tandem duplication has occurred, only one of the duplicates is assigned to the pillar while the other is assigned to a separate pillar containing the second copy in any species in which the tandem duplication has occurred. If a locus is hit by several pillars, but none of them has conserved synteny, the location is annotated as a gene but is not assigned to any existing pillar.

2.3.4 Gene models

For each stored hit between a query protein from a pillar *P* and a genomic location *L*, a gene structure prediction is made, based on YGOB's information about intron/exon structures in pillar *P*. We choose a reference gene *R* from pillar *P* to use as a basis for the model, giving preference where possible to genes from the same species group (post-WGD or non-WGD) as the new genome. We also give preference to intron-containing gene models over gene models without introns, because we want to consider the possibility that the new gene might contain an intron if any of the existing annotated genes in the same pillar and species group contains an intron. For example, if the new genome comes from a post-WGD species, and some of the genes from post-WGD species in pillar *P* contain an intron, *R* is chosen to be the post-WGD intron-containing gene with the best TBLASTN E-value. If no such gene exists, the order of preference in choosing *R* from the set of genes in *P* is as follows: (i) the post-WGD gene with the best hit; (ii) the intron-containing gene with the best hit; (iii) the gene with the best hit. If more than one gene has the same BLAST E-value (such as 0.0), the gene with the best score is used. The mean lengths of the non-WGD genes and post-WGD genes in *P* are also calculated at this step. Once a reference gene *R* has been defined for locus *L*, we store the coordinates of the best TBLASTN hit between *R* and locus *L* regardless of its E-value.

The next step in creating a gene model at locus *L* depends on the number of HSPs generated between the reference gene *R* and the genome in the vicinity of locus *L*. If only one HSP is present, the endpoints of this HSP are used directly as seed coordinates to build a gene model. If two or more HSPs are detected on the same strand, we evaluate them in pairs in order of location (**Figure 2.1**) and we classify the relationship between each pair of consecutive HSPs as one of the following four situations, depending on their coordinates: (i) frameshift (the two HSPs are the result of a probable frameshift sequencing error); (ii) low similarity (the two HSPs and the region between them are all part of the same gene and can be merged without requiring a frameshift); (iii) intron (the two HSPs correspond to matches between individual exons and the genome sequence); (iv) gene duplication (all other instances of consecutive HSPs are assumed to be full or partial gene duplications, which is confirmed by their coordinates). In situations i-iii, the outer edges of the two HSPs are used as seeds for the gene model; in situation iv, two separate models are created.

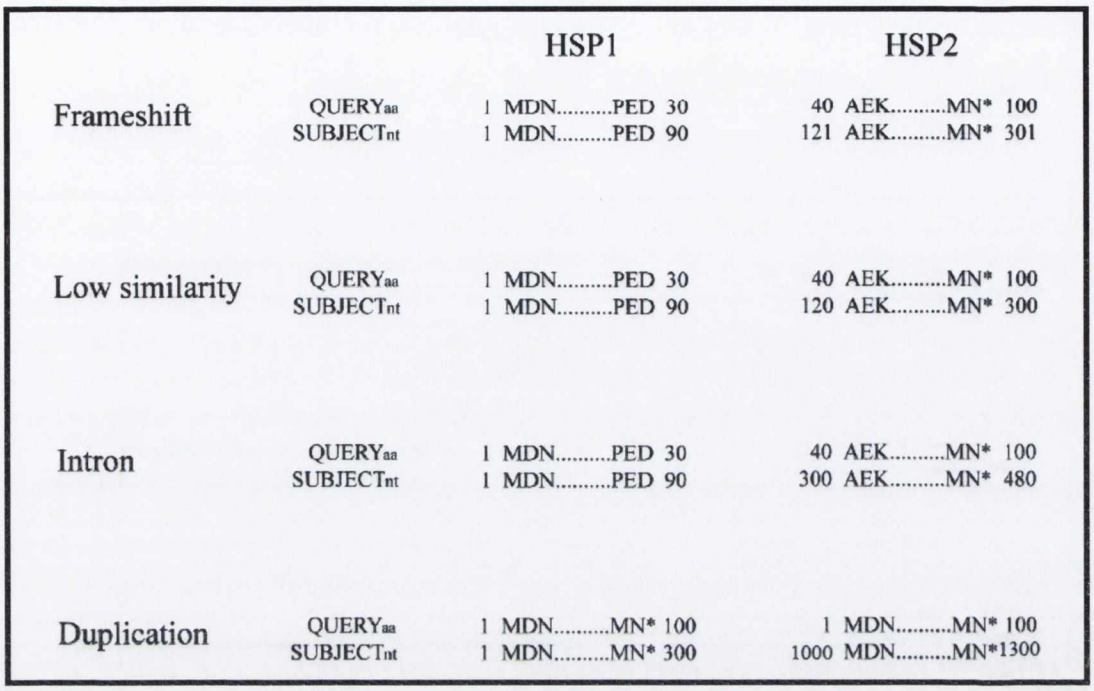


Figure 2.1 When two or more HSPs are present their coordinates are analysed in pairs based on consecutive order to define their relationship into one of four categories. (a) Frameshift: two consecutive HSPs are in different frames but the distance between them is similar in both the query and the subject. (b) Region of low similarity: two consecutive HSPs are in the same frame, and are separated by the same distance in both the query and the subject. (c) Intron: two consecutive HSPs for which query and subject coordinates are not conserved. This possibility is only considered if an existing gene from the same pillar and species group contains an intron. (d) Duplication: If all other possibilities have been excluded, two consecutive HSPs suggest a probable local duplication.

When these possibilities have been evaluated, the seed coordinates derived from the HSPs are then used to construct a gene model. We attempt to extend the seed coordinates upstream and downstream to find start and stop codons for the gene (**Figure 2.2**), by matching the HSP's location and frame to a map of all open reading frames in the genome (generated using the GetORF program from the EMBOSS package (Rice et al. 2000)). If we fail to find a suitable start or stop position by elongating the HSP in this way, we instead look for the requisite codon within the HSP. For instance, if no upstream start codon is found, we will trim the 5' end of the HSP by up to 45 nucleotides in order to find a suitable start codon. In the event that a suitable start codon is still lacking after this step, the gene is annotated with the seed coordinates and is

tagged for manual curation (indicating that the automated process could not construct a satisfactory gene model at this locus).

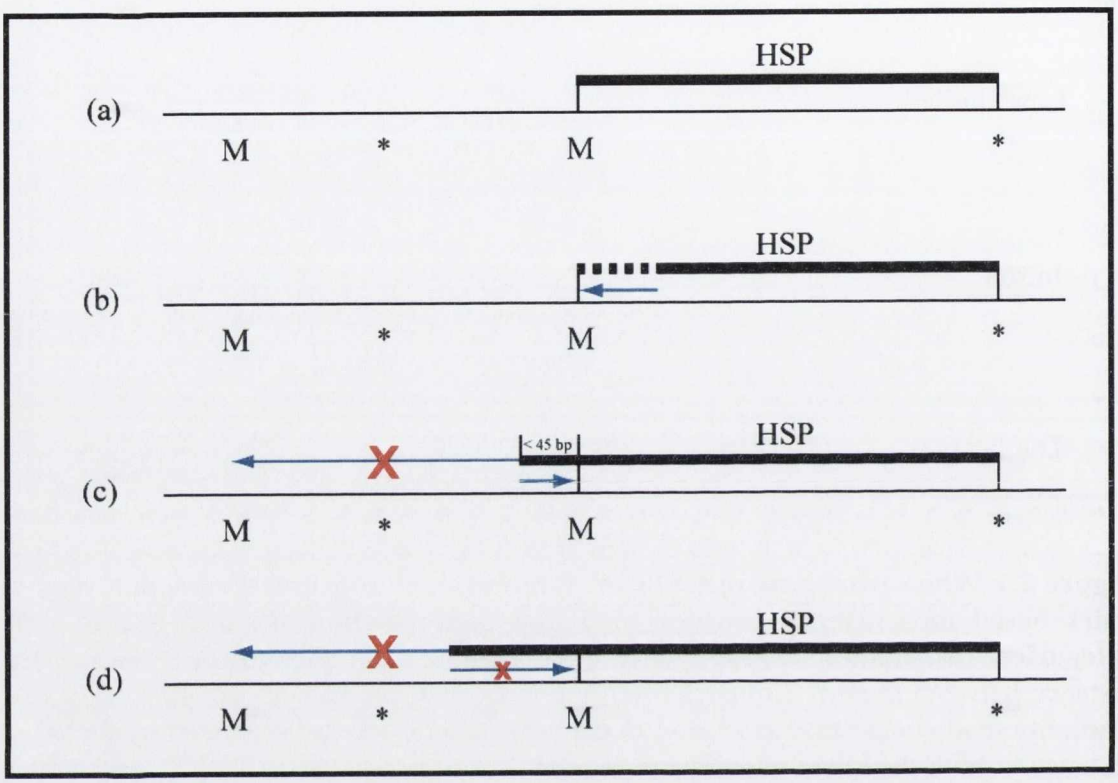


Figure 2.2 Method for defining start and stop codon coordinates. The BLAST HSP (black) is elongated or shrunk as follows: (a) If the HSP (or the upstream HSP, in the case where a pair of HSPs is being considered) begins with a leading methionine, no change is made to the starting coordinate. (b) If the HSP does not begin with a methionine, the ORF is extended upstream looking for the longest ORF starting with a methionine. (c) If during extension a stop codon (asterisk) is encountered before reaching a leading methionine, the software instead searches for a leading methionine within the first 45 nucleotides of the HSP. d) If no suitable starting methionine is found using these steps, the original coordinates of the HSP are kept and the gene is tagged for manual inspection. A similar procedure is followed to define the stop codon coordinates.

2.3.5 Intron annotation

Because introns in yeast genes are rare, we only consider the possibility that a gene model may require an intron if another gene pillar *P* already contains an annotated intron. This other gene must come from the same species group (that is, post-WGD or non-WGD) as the new genome; the user specifies whether the new genome is post-

WGD or non-WGD when launching YGAP. We use the HSPs generated in the TBLASTN search to test for the existence and location of the intron. If there are two HSPs, we search for a possible 5' splice site (GTATGT, GTCAGT, GTTCGT, GTACGT, GTAAGT, GCATGT, GTATGA, GTATGC), branchpoint site (ACTAAC, GCTAAC, ATTAAC) and 3' splice site (CAG, TAG) (Spingola et al. 1999). In many cases, however, the first exon of a gene is too small to generate a TBLASTN hit so there is only one HSP, corresponding to exon 2. In these cases we attempt to identify a suitable intron and a suitable start and stop codon, in order to make a protein similar to that of other species. As almost all intron-containing genes in YGOB have only one intron (2099 out of 2176 genes, totaled over all species), we search only for one intron per gene. Restricting the search to one intron per gene greatly simplifies the process, as without HSP data the number of combinations of exon features that could be generated to make a feasible multi-intron gene is too large to make an accurate prediction.

2.3.6 Frameshift correction

If the analysis of HSP pairs detects an apparent frameshift sequencing error (Figure 1) in the scaffold data, YGAP can try to correct the error automatically. This option can be enabled or disabled by the user when YGAP is launched. If enabled, the output from YGAP will include a modified version of the scaffolds file in which bases have been added, or more rarely removed, at particular sites in order to correct frameshifts. YGAP's output also includes lists of the genes in which frameshifts have been automatically corrected, and lists of genes in which a probable frameshift was detected but no automated correction was possible. The user can also choose to disable automatic frameshift correction, but still generate a list of genes in which probable frameshift errors have been detected.

The presence of a frameshift error usually results in two HSPs in different frames. YGAP provides the option of making two types of automatic correction: (i) If a file of primary sequence reads from the same species was provided as part of the input, we carry out a BLASTN search against the reads file using as a query the region from the scaffolds file corresponding to 50 bp upstream of the end of the first HSP to 50 bp downstream of the start of the second HSP. For BLASTN hits with E < 1e-30, we examine the match between the query and the read, and noting the indels. For each indel

seen in the search, we count how many reads contain it. We take the most common indel and test whether it would result in creation of an intact ORF. If it does, we make the corresponding change in the scaffold sequence and the frameshift has been fixed. If not, we do not make the change and instead we examine the next-most common indel. We continue this process until the frameshift has been fixed or until there are no more indels that were seen in at least 2 reads. This type of frameshift correction is optimal for reads obtained with the Roche 454 technology, because they are long enough to contain both the needed regions on the first and second HSP. (ii) If no reads file is available, one or two N nucleotides are inserted into the gene containing a frameshift, at the estimated site, in order to temporarily fix it.

2.3.7 SearchDOGS and large ORF steps

Two final steps of searching for protein-coding genes are carried out after the initial annotation has been completed. First, we run a version of SearchDOGS (OhEigeartaigh et al. 2011) to look for small, highly-divergent genes that can be recognized based on their conserved synteny to orthologs in other species (DOGS is an acronym for Database of Orthologous Genomic Segments). SearchDOGS does not employ any threshold for BLAST similarities, so it can find weak hits (E > 1e-5) that were missed by the TBLASTN method described above, provided that they are in a conserved genomic location and do not contain introns. Second, we use GetORF (Rice et al. 2000) to identify any large ORFs (≥150 amino acids, not overlapping with any other feature) that remain unannotated in the genome. Genes predicted by SearchDOGS and GetORF are included in the genome annotation and also listed separately in the YGAP output to allow them to be examined manually.

2.3.8 Retrotransposons

Retrotransposons (primarily Ty elements and similar elements in other species (Bleykasten-Grosshans and Neuveglise 2011)) pose a particular challenge to automated annotation because (i) they are mobile, so their locations are usually not conserved across species or even among different strains of the same species; (ii) they occur in multiple copies in most species and some copies can be highly similar in sequence; (iii) their repetitive nature tends to cause problems to sequence assembly software, so they often occur at the ends of contigs and their sequences are often incomplete; (iv) some

copies of these elements are inactive and their protein-coding regions are pseudogenes. In YGOB, annotated protein-coding genes that are parts of retrotransposons are flagged with a special 'Ty' label and displayed in a dark gray color. YGOB always leaves these genes as singletons and does not put them into pillars with genes from other species.

During protein-coding gene annotation in YGAP, if a YGOB gene carrying the 'Ty' label hits (E-value < 1e-5) a region of the new genome, that region will be flagged as Ty-like. YGAP will attempt to identify coordinates for the gene as described above, but the gene will be flagged as 'Ty' and will be left as a singleton. YGAP does not attempt to work out the detailed structure of retrotransposons.

2.3.9 Mini-YGOB browser

Pipeline results are presented in an individual webpage with a set of files listing the full annotation (including all the genes), tagged genes (one list per tag, such as genes tagged as having possible uncorrected frameshifts), as well as other files depending on what options the user has selected. Additionally, we provide the user with a browser interface to inspect the genome. This browser is a simplified version of YGOB, in which the only species displayed are the new genome, *S. cerevisiae* (as a post-WGD reference species), *A. gossypii* (as a non-WGD reference), and the inferred Ancestral genome (Gordon et al. 2009). This 'mini-YGOB' view allows the user to examine the structure of the annotated genome in the region around any gene of interest, for example to examine the context of singleton genes, and it provides an easy way to retrieve the predicted protein sequence of any gene of interest. The mini-YGOB interface also helps the user judge the quality of the genome assembly and annotation by showing the extent of colinearity between the new genome and the reference genomes, and by allowing the user to see the extent to which genes conserved among other species are missing (or remain unannotated) from the new genome. We anticipate that a user might run YGAP to produce an initial annotation of the genome, then use the mini-YGOB interface to look for structural problems in the assembly (such as telomere-to-telomere fusion artifacts, or to identify where small scaffolds may fit into the genome), and then perhaps modify the scaffolds file and re-run YGAP. The mini-YGOB view is only accessible to the user, and we intend it to be used for initial inspection of the genome and annotation,

rather than as a permanent database for the genome (new genomes can be added to YGOB proper when finalized).

2.4 Results

YGAP was designed to have a simple interface to our webserver. An upload screen (**Figure 2.3A**) allows the user to upload the sequence files (scaffolds, and possibly reads and contigs), to provide a prefix that will be used for gene names, to designate the species as post-WGD or non-WGD, and to select options related to automated frameshift correction. After the run is complete, the user receives an email directing him to a results page (**Figure 2.3B**). A sample results page, for *Tetrapisispora blattae*, can be viewed at http://wolfe.gen.tcd.ie/annotation/example.html. The results page includes links to several types of output files and gene lists, as well as a link to a "mini-YGOB" browser (**Figure 2.3C**).

2.4.1 Tests with S. cerevisiae

To test YGAP's performance we ran an automatic annotation of the genome of *S. cerevisiae*, which is very well studied and annotated. To do this, we retrieved the chromosomal DNA sequences of same version of the *S. cerevisiae* genome that is currently used in YGOB (strain S288c; based on sequence and annotation from SGD version of 25 August 2007, excluding genes annotated as 'dubious'). To avoid using the current *S. cerevisiae* gene set as a source of information, we removed all *S. cerevisiae* genes from the YGOB pillar set for this experiment. We then ran YGAP using the *S. cerevisiae* genome DNA as input, and with the frameshift correction option disabled. This setup replicates the simplest scenario that the pipeline may encounter, where no file of primary sequence reads is available but the scaffolds are expected to be highly accurate.

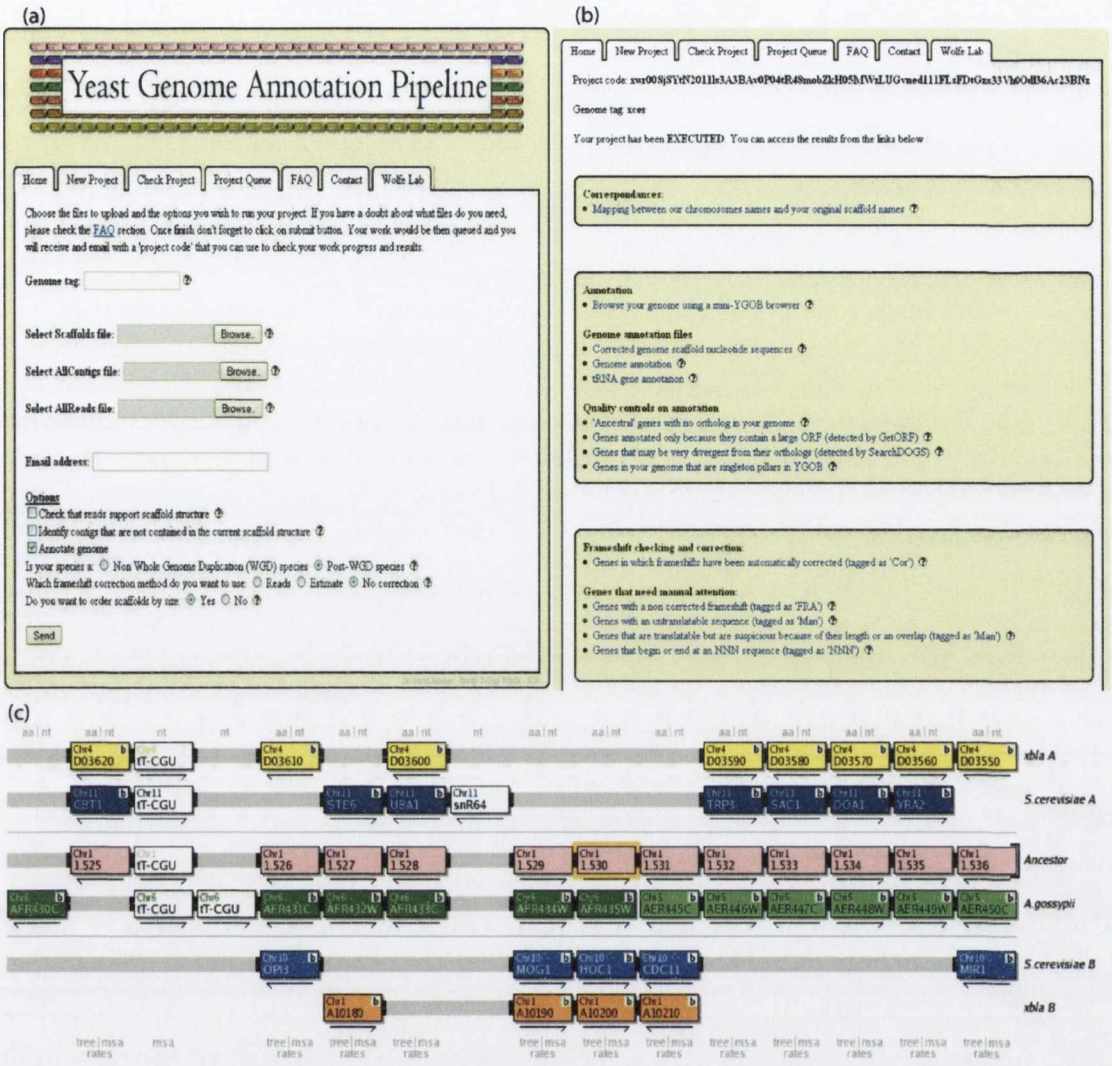


Figure 2.3 Screenshots from the YGAP webpage. (a) Upload screen. (b) Results page that includes links to several types of output files and gene lists. (c) Mini-YGOB browser showing the genomes of *E. gossypii* (non-WGD species, in green), *S. cerevisiae* (post-WGD species, in blue), the new annotated species (here, *T. blattae*, post-WGD species, in yellow/orange), and the Ancestral genome (in pink).

The whole annotation took approximately 5 hours on a 3 GHz processor with access to 4 GB of RAM and was then compared with the curated *S. cerevisiae* annotation in YGOB. For comparison, we did a similar annotation using AUGUSTUS (Stanke and Waack 2003) with default options set. AUGUSTUS is a widely used automatic annotation tool that has been specifically trained for *S. cerevisiae*. YGAP predicted 5659 potential genes in *S. cerevisiae*, compared with 5551 predicted by AUGUSTUS and 5604 curated genes in YGOB (including genes from Ty elements). To assess the accuracy of the gene models, we compared the predicted gene coordinates to

those in the YGOB database (**Table 2.1**). The analysis of coordinates (start codon position, stop codon position, and the coordinates of introns where present) shows that our pipeline predicted the structures of 5119 genes completely correctly, which is 181 more than AUGUSTUS and 91% of the genes in YGOB. However if we consider only start and stop codon positions and ignore intron structures, the gap between the YGAP and AUGUSTUS predictions is reduced to about 70 genes.

Table 2.1 Comparison of automatic reannotations of the *Saccharomyces cerevisiae* genome by YGAP and AUGUSTUS, to the reference annotation.

	Same Coordinates		False Negatives		Overlap		False Positives		Wrong Start or Stop coordinate	
Chromosome	YGAP	AUGUSTUS	YGAP	AUGUSTUS	YGAP	AUGUSTUS	YGAP	AUGUSTUS	YGAP	AUGUSTUS
Chr 1	75	76	7	5	1	1	2	3	8	9
Chr 2	354	346	3	13	2	2	5	6	28	26
Chr 3	135	126	1	9	6	0	6	4	9	16
Chr 4	668	646	8	24	2	1	15	16	47	54
Chr 5	236	232	1	6	7	0	6	10	18	24
Chr 6	101	106	2	8	1	1	8	5	14	3
Chr 7	457	446	1	9	6	1	7	14	37	45
Chr 8	237	235	2	7	2	1	9	8	22	20
Chr 9	180	176	0	7	5	1	4	4	20	21
Chr 10	312	300	0	9	4	0	6	5	25	32
Chr 11	284	275	1	9	5	0	1	0	18	24
Chr 12	445	414	3	14	5	2	10	10	24	48
Chr 13	405	387	2	10	7	0	5	9	22	39
Chr 14	353	337	0	11	4	3	3	4	25	31
Chr 15	465	451	3	14	4	0	5	12	34	42
Chr 16	412	385	10	17	4	0	7	7	25	49
Total	5119	4938	44	172	65	13	99	117	376	483

Columns show the numbers of genes in each category, when compared to genes in the reference YGOB annotation of the *S. cerevisiae* genome (which is based on *Saccharomyces* Genome Database annotation).

The numbers of false negative and false positive gene predictions by YGAP also compare favorably to AUGUSTUS. The numbers of false negatives (situations where the automated method fails to predict any gene model in a region of the genome where YGOB shows an annotated *S. cerevisiae* gene) for YGAP was four times lower than for AUGUSTUS (44 versus 172; **Table 2.1**). A second category of false negatives consists of "overlap" cases, where the automated programs failed to annotate a gene and instead extended a neighboring gene (usually by incorrect start codon assignment) so that it overlapped with the range of the unannotated gene. YGAP overlooked 65 genes for this reason, compared to 13 for AUGUSTUS. Combined, these two false negative categories amount to 109 genes for YGAP and 185 for AUGUSTUS. The numbers of false positives (gene models predicted in regions of the genome where no gene is present in

the YGOB annotation of *S. cerevisiae*) were more similar: 99 for YGAP and 117 for AUGUSTUS. Of the remaining gene predictions whose structures were incorrect, most had either a wrong start or stop coordinate (376 from YGAP and 483 from AUGUSTUS).

We also compared the performance of YGAP and AUGUSTUS in predicting the intron/exon structures of genes in *S. cerevisiae* (**Table 2.2**). In the *S. cerevisiae* genome there are 265 introns in the protein-coding regions of 256 genes (Christie et al. 2009). YGAP predicted a total of 146 introns, of which 2 were false positives (the gene actually has no intron). The main problem was that YGAP's false negative rate (122 true introns not predicted) is high. Of the 144 true-positive predictions from YGAP, the predicted intron coordinates were completely correct for 127 (87% of the predictions, or 47% of all introns studied). AUGUSTUS had a similarly high false negative rate, and predicted more introns in total due to a higher false positive rate (**Table 2.2**).

Table 2.2 Comparison of intron structure predictions in *S. cerevisiae* and *N. castellii* by YGAP and AUGUSTUS.

Species	Software	Predicted introns (a+b+c)	False positives (a)	Completely correct (b)	Real intron, incorrect coordinates (c)	False negatives (d)	Total introns studied (b+c+d)
S. cerevisiae	YGAP	146	2	127	17	122	266
	AUGUSTUS	221	90	87	44	121	252
N. castellii	YGAP	146	12	123	11	58	192
	AUGUSTUS	251	173	58	20	94	172

It should be noted that False Negatives in YGAP include not only those genes for which no intron was predicted by the software, but also those for which intron coordinates could not be defined and were tagged for manual curation. The total number of introns studied (rightmost column) differs between YGAP and AUGUSTUS because some genes were not predicted by both methods.

2.4.2 Automatic annotations of Naumovozyma castellii and Tetrapisispora blattae

We used YGAP to automatically annotate the genome of *Naumovozyma castelli*. This species has previously been called *Saccharomyces castellii* and *Naumovia castellii*. Its genome was originally sequenced, by Sanger sequencing to draft (3x) coverage by

Cliften et al. (2003) which resulted in hundreds of contigs. Genes in these contigs were annotated manually by our laboratory as part of the YGOB project (Byrne and Wolfe 2005), and we later added 18 genes using SearchDOGS (OhEigeartaigh et al. 2011). We refer to this annotation as the 'Scas' annotation (corresponding to the prefix of the gene names as annotated in YGOB). The total number of protein-coding genes in the Scas dataset is 5691.

We recently resequenced the same strain of *N. castellii* using the Roche-454 platform with 20x coverage, with a strategy designed to maximize the size of scaffolds. This genome was assembled without making use of the Sanger data. The resequenced genome comprises only 10 scaffolds, which compares reasonably well to a pulsed-field gel electrophoresis estimate that this species has 9 chromosomes (Petersen et al. 1999). We then used these *N. castellii* scaffolds as input to YGAP. As before, to avoid circular reasoning we ignored annotated *N. castellii* genes from the input YGOB pillars, but we included *S. cerevisiae* genes. For this run of YGAP, we included the sequence reads file and allowed automatic correction of frameshifts. The whole annotation took 6.25 hours on the same server, and 5682 protein-coding genes (including 19 from retrotransposons) were predicted (**Table 2.3**). We refer to this dataset as the 'Ncas' dataset (again corresponding to the prefix of the gene names). In this run, YGAP identified 184 positions in the genome as frameshift sites, and it automatically corrected 109 of these (81 nucleotide additions and 28 nucleotide deletions); it flagged the remaining 75 sites as probable frameshift sites that it was unable to correct. About 300 predicted genes were flagged as "Man", meaning the pipeline was not entirely sure with the prediction of the coordinates, and the user should check them manually.

Comparing the annotations shows that YGAP's predictions were identical to the manually predicted gene structures for 5260 genes (93% of the predicted genes), with 90 false positives and 40 false negatives (**Table 2.3**), while AUGUSTUS gets right 4701 genes (83%), with 182 false positives and 153 false negatives. As with the *S. cerevisiae* annotation, the numbers of false positives, false negatives and incorrect start/stop codons from YGAP were consistently lower than from AUGUSTUS. YGAP also outperformed in the prediction of intron coordinates, predicting far fewer false positive introns and getting the coordinates completely correct more often (**Table 2.2**).

We also used YGAP to annotate the genome of *Tetrapisispora blattae*, a post-WGD species for which no previous genomic data existed so there is no reference annotation to which we can directly compare YGAP's results. As input to YGAP we used the 8 large scaffolds and 373 contigs obtained from the Celera assembler, as well as 319,888 pairs of primary sequence reads. Automated frameshift correction and scaffold integrity checking were enabled in the YGAP run, which took 12.5 hours. The genome integrity checking steps identified two joins in the scaffold data that were not well supported by nonrepetitive paired sequence reads, so in the final version of the genome we broke these joins. The frameshift correction steps inserted a total of 398 nucleotides (including 194 As and 184 Ts) and deleted 48 from the 14.1 Mb genome, affecting the structures of 418 genes. YGAP predicted 5600 protein-coding genes in *T. blattae*, corresponding to 4534 loci in the ancestral genome (Gordon et al. 2009). There are 383 loci in the ancestral genome at which *T. blattae* has no annotated gene, and there are 830 annotated singletons in *T. blattae.*

Table 2.3 Comparison of automatic annotations of the *Naumovozyma castellii* genome by YGAP and AUGUSTUS, to the reference annotation.

	Same Coordinates		False Negatives		Overlap		False Positives		Wrong Start or Stop coordinate	
Scaffold	YGAP	AUGUSTUS	YGAP	AUGUSTUS	YGAP	AUGUSTUS	YGAP	AUGUSTUS	YGAP	AUGUSTUS
scf7180000013405	544	485	3	17	13	2	14	25	28	84
scf7180000013407	208	185	2	5	5	1	5	8	10	34
scf7180000013408	462	418	1	10	9	1	3	9	16	59
scf7180000013409	331	294	1	8	11	11	6	9	15	44
scf7180000013410	1427	1289	12	34	31	5	19	46	56	198
scf7180000013411	851	764	8	28	12	5	13	19	35	106
scf7180000013412	387	342	3	16	8	1	11	25	18	57
scf7180000013414	386	338	3	16	13	8	6	18	11	50
scf7180000013415a	290	250	4	12	1	3	3	8	17	47
scf7180000013415b	374	336	3	7	6	1	10	15	17	56
Total	5260	4701	40	153	109	38	90	182	223	735

Columns show the numbers of genes in each category, when compared to genes in the reference (Scas) manual annotation of *N. castellii* genes.

2.5 Discussion

Our aim was to develop a new bioinformatics pipeline for the automated annotation of yeast genomes, exploiting information from existing genomes to the greatest extent possible. The pipeline has been specifically designed for the *de novo* annotation of yeast genomes using the outputs from high-throughput sequencing methods, *i.e.* files of scaffolds, contigs, and reads. The full YGAP pipeline includes not only *de novo* genome annotation but also some error-checking tools including a verification that scaffold structure is well supported by the paired primary sequencing reads, and identification of any large contigs that were not included in the scaffolds. These steps are not essential in order to run the automatic annotation, but they can improve the quality of the sequenced genome and thus the quality of the annotation. YGAP is also unique in its ability to make use of a file of primary sequence reads to try to correct apparent frameshift errors in the assembly. However, these features require the user to upload the primary reads data, and at the moment this is only possible for Roche-454 or Sanger projects (the primary files from Illumina sequencing are simply too large to upload to our server). Nevertheless, we expect that these additional quality-control steps will eventually become unnecessary because they are largely dependent on the quality of the sequence and the assembly, so the need for them should decline as the quality of next-generation sequencing techniques improves. The basic annotation steps in YGAP require only a scaffolds file, which can come from any sequencing platform.

A key feature of YGAP is that it tries to use the orthologous genes from other species, identified by a synteny method, to make gene structure predictions. Some other previous automatic genome annotation tools have also been based on the identification of orthologs. Some of these are quite specific, such as Dogma (Wyman et al. 2004), an annotation tool designed to annotate organellar genomes, and MaGe (Vallenet et al. 2006) which annotates microbial genomes. Other automatic genome annotation tools can be more widely used and can annotate both prokaryotic and eukaryotic genomes. RAPYD (Schneider et al. 2010) is an annotation platform that has been developed specifically for yeast species and uses AUGUSTUS (Stanke and Waack 2003) as its main tool for gene prediction.

We tested YGAP using the *S. cerevisiae* annotation in YGOB as a 'gold standard' for reference and compared the results to predictions made by AUGUSTUS using its *S. cerevisiae* model. The results showed that YGAP was able to correctly predict more than 90% of *S. cerevisiae* gene structures correctly. Importantly, the use of multi-species annotations in YGOB together with synteny information resulted in a significant reduction in the numbers of both false-positive and false-negative gene predictions. Moreover, the genome annotations and the orthology and paralogy relationships between the genes present in YGOB have been manually curated, so it is unlikely that using them in our annotation pipeline will propagate errors in the newly automated annotations. A manual check showed that most spurious YGAP gene annotations (false positives) correspond to annotated pseudogene features. YGAP's failure to detect certain genes (false negatives) was due to: (i) the genes being species-specific gene gains, (ii) the genes being located in subtelomeric regions where rapid gene family expansion has resulted in multiple gene duplications, and (iii) in post-WGD species, occasional failure to annotate both copies of a gene retained in duplicate after WGD if both copies are located on the same scaffold.

YGAP's performance on intron-containing genes is less impressive, resulting in correct prediction of both the presence of an intron and the locations of its boundaries only about half the time, but it nonetheless outperforms AUGUSTUS and makes few false-positive predictions of introns. One cause of the poor performance on introns may be inaccurate annotation of intron-containing loci in the existing genome annotations in YGOB. It should also be noted that YGAP will only predict a maximum of one intron per gene, whereas in fact a small number of genes (9 in *S. cerevisiae*) are known to have two introns in their coding regions. One solution for the user would be to also have RNA-seq data of his new genome, and then perform an alignment of the assembled RNA-seq sequences with the scaffold sequences. With this method he will be able to see the exact location of introns in his new genome.

In conclusion, YGAP has been able to correctly annotate 90% and 93% of the genes in *Saccharomyces cerevisiae* and *Naumovozyma castellii* respectively. All the results are provided to the user for manual inspection in different lists of tagged genes. These include a list of genes in which frameshifts have been automatically corrected; a list of those in which a frameshift probably exists but was not automatically corrected; a

list of genes whose DNA sequences are not properly translatable (for example, due to the presence of an unannotated intron); and list of genes that extend into regions of scaffolds that contain continuous runs of N nucleotides, making identification of start/stop codons impossible. YGAP has proven very effective in the annotation of new genomes from Saccharomycetaceae species and we are currently using it to annotate the genomes of seven species sequenced using Roche-454 technology.

The newly sequenced and automatically annotated genome sequences were used in a study on yeast sex chromosome evolution (Gordon et al. 2011), detailed in **Appendix I**. In this work, I automatically annotated seven yeast species, and I helped with the manual curation of the annotations using Artemis (Carver et al. 2008). All of these species belong to the Saccharomycetaceae family. Despite their broad evolutionary range and significant individual variations in each lineage, the members of this family are known to have a high degree of conserved synteny (Souciet et al 2009). However, when genomes of Saccharomycetaceae yeasts were compared to yeast genomes from other families, it was shown that the gene order was no longer well conserved (Seoighe et al. 2000; Fischer et al. 2006). If the user wants to annotate a yeast species that does not belong to the Saccharomycetaceae family with YGAP, the automatic annotation will mainly be based on sequence similarity, because the gene order will not be conserved.

Moreover, a special in-house version of YGAP was built, adapted to work with the Candida Gene Order Browser (CGOB) (Fitzpatrick et al. 2010) instead of YGOB. This version of YGAP was used in a collaborative project with Prof. Geraldine Butler's laboratory at University College Dublin to automatically annotate the genome sequence of *Candida orthopsilosis*, a species closely related to *Candida parapsilosis* (whose genome was used as a reference for the conservation of the synteny) (Riccombeni et al. 2012).

Chapter 3

Study of the singleton genes in newly sequenced Saccharomycetaceae species

3.1 Abstract

Gene gain is one of the key mechanisms in genome evolution. Genes can be gained by duplication, by horizontal transfer, or by *de novo* evolution. Species-specific genes (singletons) are important for species' adaptation to variable environments. I have classified the singleton genes from seven newly sequenced yeast species automatically annotated by YGAP, depending on their inferred origins. Singletons in these species represent 7% of the genes on average. Most of them seem to have arisen from duplication events, even if the original gene is not always identifiable. Some of them have become amplified, such as a 14-member family of singletons in *Tetrapisispora blattae*. Horizontal gene transfers are very rare in our dataset, except in *Torulaspora delbrueckii*. Orphan genes are genes that lack homologs in other lineages and can represent up to 46% of the singletons. Some orphans belong to orphan gene families that can have up to 20 members.

Singleton: In this chapter, we use the word "singleton" to refer to genes that were not automatically assigned to an existing YGOB pillar by our pipeline YGAP. Those genes have no syntenic orthologs in the other yeast species present in YGOB. Some of them can be lineage-specific.

3.2 Introduction

Orphan genes have been studied since the earliest genome sequencing projects: when chromosome III of *S. cerevisiae* – the first eukaryotic chromosome ever sequenced – was analysed, half of its protein-coding ORFs had no known homologs in any organism (Oliver et al. 1992; Dujon 1996). Today, many other yeast genomes, some from species closely related to *S. cerevisiae*, have been sequenced, and less than 2% of the proteome of the baker's yeast is regarded as orphan (Ekman and Elofsson 2010).

Two scenarios have been suggested to explain the emergence of orphan genes (Tautz and Domazet-Loso 2011): (i) the duplication-divergence scenario, where a newly duplicated gene will go through a phase of fast adaptive evolution. During this time it will lose all similarity to its progenitor. In some cases a rearrangement or transposon insertion can be combined with the initial duplication, which will modify the gene even more (Long et al. 2003). (ii) the *de novo* scenario, where a functional gene can arise from a non-coding sequence based on the cryptic presence of functional sites. Examples of orphan genes gained by both of these two mechanisms have been documented in yeast, *Drosophila* and primates (Levine et al. 2006; Cai et al. 2008; Knowles and McLysaght 2009; Toll-Riera et al. 2009).

In this chapter, I examine the genes automatically annotated as singletons by YGAP in the seven newly-sequenced species of the YGOB dataset. Singletons are genes that were not automatically assigned to an existing YGOB pillar during the annotation process. Those genes have no syntenic orthologs in other species present in YGOB. Singletons can be orphan genes, if they have no homologs in any other species. They can also be single duplicated genes whose progenitor is easily identifiable by sequence similarity, like the genes *ADH1* and *ADH2* in *S. cerevisiae*, which differ by only 24 amino acids (Thomson et al. 2005). Some may not have a clear progenitor gene, but contain a specific protein domain and can be linked to a protein superfamily. Those genes can have originated from a duplication event and diverged quickly, which can explain the low sequence similarity between the mother and daughter genes except at the location of the conserved domain. Singletons can also be gained by Horizontal Gene Transfer (HGT). HGT is important in prokaryotes but rarer in yeast. Several cases of HGT have been documented in *S. cerevisiae* and *E. gossypii* (Hall et al. 2005). To be fixed in a population, a gene gained by HGT should provide a particular benefit, like an adaptation to a specific condition. For example, in *S. cerevisiae*, the gain by HGT of the gene *BDS1* that encodes a multifunctional sulfatase means that this species can live in places where organic sulphur is predominant (Hall et al. 2005).

In this study, I have classified the singletons automatically annotated in the seven newly sequenced yeast species into several categories, depending on their probable origin. I present several interesting examples of genes gained by HGT or by

multiple duplications in some species, and what kind of benefit they could have provided. Finally, I discuss the orphan genes I have found.

3.3 Methods

All of the new genomes were annotated with YGAP, the automatic annotation pipeline presented in Chapter 2. Some of them are currently undergoing additional manual editing using the sequence editor Artemis (Carver et al. 2008).

Singletons were classified into one of the five following categories: (i) **Duplication**, when the singleton was formed by an identifiable species-specific gene duplication; (ii) **Other species**, when the singleton has no paralogs in its own genome, but it has nonsyntenic homologs in other species. It may therefore be the result of ancient gene duplication and differential loss, or horizontal gene transfer from a closely related donor, or gene transposition; (iii) **Domain**, when the singleton has no clear homologs but contains an identifiable protein domain; (iv) **Horizontal gene transfer**, when the singleton is the probable product of horizontal transfer from a donor outside the fungal family Saccharomycetaceae; and (v) **Orphan**, when the singleton has no recognizable homologs or protein domains. To achieve this classification, a BLASTP against all the proteins in the YGOB database is initially done for all the singletons. If a singleton from a species X has a strong hit (more than one-third of the E-value of its hit to itself) with a gene from the same species that was annotated and put into an existing YGOB pillar by our annotation pipeline, the singleton is classified in the "Duplication" category. If no such hit exists, but if the singleton shares sequence similarity with proteins in other species, it will go into the "Other Species" category. If the BLASTP search results in no significant hits, a BLASTP of the singleton against the NCBI nr (non-redundant) protein database is done. If the singleton matches proteins in other species that are not in the YGOB database, it is classified as "HGT". If no hit is found in the nr database, a search for a particular protein domain is done against several databases: Pfam (Finn et al. 2010), ProDom (Servant et al. 2002) and CDD (Marchler-Bauer et al. 2011). If the singleton contains one or several known domains, it is classified in the "Domain" category. If the singleton is not classified in any of the preceding categories, it is considered as an "Orphan".

Phylogenetic trees were made by the Maximum Likelihood method using PhyML from the Seaview 4 package (Gouy et al. 2010).

3.4 Results

We have now automatically annotated a total of seven species with YGAP. Most of these species are currently under additional manual curation, for example to correct the remaining frameshifts or introns that were not detected during the annotation process. We have currently annotated 5257 genes in *N. castellii*, 5159 genes in *N. dairenensis*, 5011 genes in *T. phaffii*, 5280 genes in *T. blattae*, 5098 genes in *K. africana*, 5000 genes in *K. naganishii* and 5077 genes in *T. delbrueckii*. These numbers are going to change slightly with the manual correction in progress.

We were interested in studying the singletons in those seven species, and classified them into five categories as defined in Methods (**Table 3.1**). The number of singletons annotated is quite similar (around 350 or about 7% of all annotated genes) among the species, except for *T. delbrueckii* (only 206 or 4%) and in *T. blattae* (almost 600 or 11%). These values can be compared to the significantly lower singleton frequencies of 3.2% (*E. gossypii*) to 7.8% (*V. polyspora*) among the other species in YGOB, which have been more extensively manually curated. The number of singletons in the new genomes is likely to be overestimated, due to errors resulting from next-generation sequencing (frameshifts) or failure of the automated annotation process to assign a gene to the appropriate pillar. The number of singletons in a species will also depend to some extent on whether or not another closely related genome is present in YGOB; for example there are 125 nonsingleton pillars that contain only a *S. cerevisiae* gene and a *S. bayanus* gene. The most frequent categories of singletons are those in the "Other Species" (27-50%), "Orphan" (17-46%) and "Duplication" (15-31%) categories (**Table 3.1**). It is notable that a high proportion of *T. delbrueckii* singletons (50%) have paralogs in other species, and that *K. naganishii* – a species that has only been isolated from decaying leaves and soil in Japan – has a high frequency of orphans. In the sections below, I discuss some of the more interesting examples of singletons that we identified.

Table 3.1. Classification of the singletons annotated in the newly sequenced species.

	N. cas	*N. dai*	*T. pha*	*T. bla*	*K. afr*	*K. nag*	*T. del*
Duplication	108 (28%)	71 (20%)	60 (19%)	167 (28%)	112 (31%)	84 (25%)	31 (15%)
Other species	105 (27%)	108 (31%)	141 (44%)	250 (42%)	125 (34%)	87 (26%)	93 (45%)
Domain	30 (8%)	45 (14%)	1 (<1%)	7 (1%)	18 (5%)	12 (3%)	25 (12%)
HGT	2 (1%)	0 (0%)	6 (2%)	0 (0%)	4 (1%)	0 (0%)	21 (10%)
Orphan	137 (36%)	120 (35%)	113 (35%)	172 (29%)	105 (29%)	156 (46%)	36 (18%)
Total	382	344	321	596	364	339	206

N.cas: *N. castellii*; *N. dai*: *N. dairenensis*; *T. pha*: *T. phaffii*; *T. bla*: *T. blattae*; *K. afr*: *K. africana*; *K. nag*: *K. naganishii*; *T. del*: *T. delbrueckii*.

3.4.1 Horizontal gene transfers into *T. delbrueckii*

We identified three groups of singleton genes in *T. delbrueckii* that are apparent products of horizontal transfer.

3.4.1.1 Multiple transposase-related genes

In the genome of *T. delbrueckii* we discovered a family of nine predicted genes that appear to code for transposases. These nine genes are scattered over nine separated locations, two of them subtelomeric, on seven scaffolds in *T. delbrueckii*. They have a single strong BLASTP hit to only one other gene present in YGOB: *AGL264W*, from *E. gossypii*. This *E. gossypii* gene codes for a protein related to bacterial transposases from the IS605 family (called IS605-TnpB). Hall et al. (2005) showed that *E. gossypii* had gained *AGL264W* from a probable bacterial donor, a species close to *Helicobacter pylori*. The IS605 transposable element family was first described in *H. pylori* (Hook-Nikanne et al. 1998). The transposases from this family are the smallest known DNA transposases. They contain two ORFs (OrfA and OrfB) that have high sequence similarity to other transposase genes of other bacterial IS elements. They do not have inverted sequences at their ends, which are usually a characteristic of many prokaryotic and eukaryotic transposons; and they always insert just 3′ of a specific four or five nucleotide sequence, unlike many DNA transposons which integrate randomly (Barabas et al. 2008).

The genome of the wine yeast *S. cerevisiae* strain EC1118 contains a pseudogene (EC1118_1F14_0067g) that is very similar to *AGL264W*, but there is no similar gene or pseudogene in the reference *S. cerevisiae* strain S288c. A BLASTP search with those nine *T. delbrueckii* proteins against the nr database showed that a similar gene (annotated as QA23_5166) is also present in another *Saccharomyces cerevisiae* strain called Lalvin QA23 that is used in white wine production (Borneman et al. 2011). No other genomes in YGOB contain genes related to this transposase family.

Using the package Seaview 4 I aligned and built a phylogenetic tree (**Figure 3.1**) from of the amino sequences of the predicted nine *T. delbrueckii* proteins, the *AGL164W* protein, the inferred translation of EC1118 pseudogene, the Lalvin protein and the IS605 transposase (TnpB) from *H. pylori*. I also used Pfam to search for the domains present in these sequences. The sequence lengths of the nine predicted *T. delbrueckii* proteins range from 168 amino acids to 495 amino acids. The three longest sequences (XDEL0G01240, XDEL0C00200 and XDEL0F03030) as well as *AGL164W*, the pseudogene from EC1118 and the *H. pylori* sequence contain three Pfam domains. All the other sequences seem to be truncated. Interestingly, a BLASTX search using the intergenic DNA regions between the six incomplete *T. delbrueckii* sequences and their neighboring genes, against the YGOB protein database, showed the missing parts of the genes scattered in different reading frames. It is unlikely that these frameshift events are due to sequencing errors, because six of the nine genes of the same family are affected, and a BLASTN search against the primary *T. delbrueckii* sequence reads did not identify any reads able to correct the frameshift events. Moreover, in *S. cerevisiae* the annotated locations of incomplete Lalvin gene QA23_5166 and the almost complete pseudogene EC1118_1F14_0067g are also flanked by similar pseudogene-like fragments of the transposase gene. It is likely that in *T. delbrueckii*, six copies of the transposase gene turned into pseudogenes, and the three remaining intact sequences could still be active. It is also possible that the *S. cerevisiae* gene QA23_5166 is in fact a pseudogene too. This search on the neighboring regions of all the transposase sequences also showed that no other genes seem to have been transferred along with the transposase.

The phylogenetic tree (**Figure 3.1**), which is rooted using the bacterial sequence, has several interesting properties. First, the two *S. cerevisiae* sequences (which cluster together) lie within the diversity of a larger clade of the nine *T. delbrueckii* genes. The two *S. cerevisiae* sequences cluster specifically with one *T. delbrueckii* family member, XDEL0D00170. This topology indicates unambiguously that the transposase family has amplified by recurrent gene duplications in *T. delbrueckii*, and that one branch of this *T. delbrueckii* family is the donor of a HGT event into the two *S. cerevisiae* wine strains.

How did such a horizontal transfer occur? Lalvin QA23 and EC1118 are commercial strains of *S. cerevisiae*. The former is used primarily for white wine production; the latter is also used in the production of white wines, as well as Champagne, cider, and sometimes red wines. *T. delbrueckii* is naturally found in fermented milk and rice, on grapes and in grape must (Renault et al. 2009). This species is mostly used in brewing German-style wheat beers. It has been shown that *T. delbrueckii* had a positive impact on the flavor of alcoholic beverages, because it produces low amount of undesirable compounds (Ciani et al. 2006; Renault et al. 2009). It also presents high fermentation purity, and was proposed as a way of minimizing acetic acid production in wine under standard or high-sugar conditions, in mixed culture with *S. cerevisiae* (Bely et al. 2008; Renault et al. 2009). It is therefore possible that *T. delbrueckii* has been in contact with *S. cerevisiae* wine strains, and that one member of the transposase-like family in *T. delbrueckii* was transferred by HGT to the genome of a common ancestor of the *S. cerevisiae* wine strains.

A second feature in **Figure 3.1** is that the *E. gossypii* gene *AGL264W* forms an outgroup to the entire clade of *T. delbrueckii* genes. The amount of sequence divergence between the transposase-like genes in these two non-WGD yeasts is surprisingly large (average amino acid sequence identity 36%) by comparison to the divergence of both yeasts from *Helicobacter pylori*. This observation and the fact that the site of the *AGL264W* gene is not syntenic with any of the *T. delbrueckii* copies suggest the possibility that two different bacterial donors were the source of the *E. gossypii* and *T. delbrueckii* genes. Such a conclusion is also supported by the fact that the two species are not closely related. If the genes in *E. gossypii* and *T. delbrueckii* have been inherited vertically from a common ancestor, then all the species that diverged between *E.*

gossypii and *T. delbrueckii* must have repeatedly lost the gene, which is highly improbable.

A further inference from the phylogenetic tree is that at least some members of the transposase family are likely to be active in *T. delbrueckii*. This is the simplest explanation for the observed mobility of the genes, both within *T. delbrueckii* and from *T. delbrueckii* to *S. cerevisiae*. It will be of interest to test this hypothesis in the laboratory. The most probable candidates for active copies are the fully intact copies XDEL0G01240, XDEL0C00200 and XDEL0F03030.

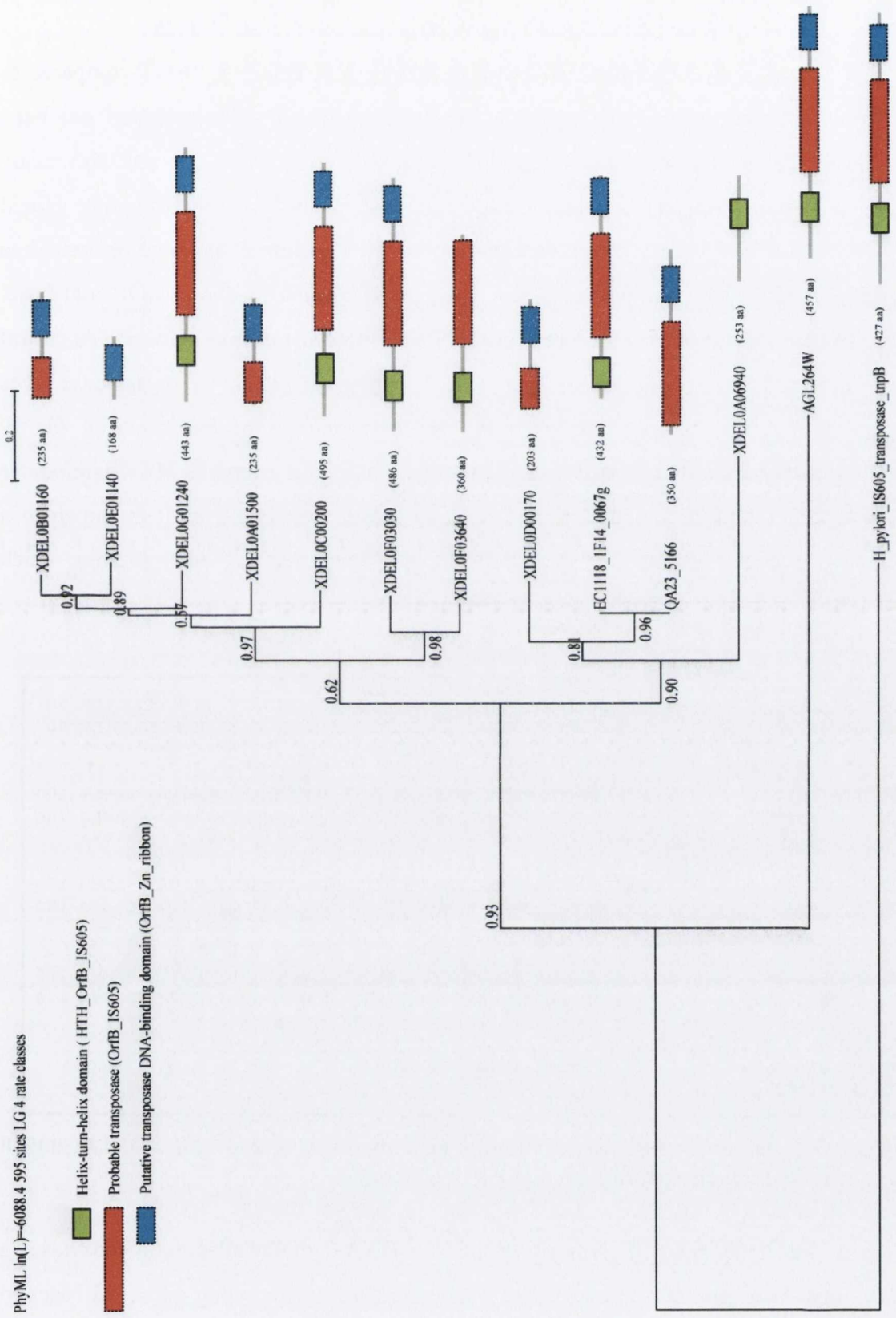


Figure 3.1: Phylogenetic tree of the nine predicted *T. delbrueckii* genes and the similar genes or pseudogenes found in other yeast species. The sequence from *H. pylori* is used as an outgroup.

3.4.1.2 3-methyl-adenine DNA glycosylase gene in *T. delbrueckii*

Interestingly, *T. delbrueckii* contains many genes gained by HGT compared to the other studied species (**Table 3.2**). This yeast is found worldwide and has been collected from different fruits, vegetables and soil, so it is in contact with many different types of organisms, which can potentially explain the number of genes gained by HGT and the diversity of the potential donors. For example, the gene annotated as XDEL0C00540 has no homologs in the yeasts present in the YGOB database. XDEL0C00540 appears as an insert in the *T. delbrueckii* genome, between Ancestral loci Anc_8.27 and Anc_8.28. However, a BLASTP search against the nr database gives many strong hits to bacterial genomes, the best ones from the bacterial family Sphingobacteriaceae (frequently found in soil or bog), for example *Mucilaginibacter paludis* (Pankratov et al. 2007). The sequence is well conserved, as it is shown in a protein sequence alignment between XDEL0C00540 and the *M. paludis* protein, which is the best hit from the BLASTP search (**Figure 3.2**).

```
XDEL0C00540             MKLSETFYQNTNVTAVARSLLGKYLFTRINGQLTGGYIVETEAYNGIIDK 50
ZP_07745263.1           MKLPVDFYTHGDVLAITRNLLGKYLFTKIDGMVTGGYIVEAEAYNGVIDK 50
                        ***.  ** : :* *::*.********:*:* :*******:*****:***

XDEL0C00540             ASHAYGNRLTPRTKIMFEDGGVAYVYLCYGIHEMLNVVTSVSGQPHAVLI 100
ZP_07745263.1           ASHAYGNRRTPRTETMFMHGGVSYVYLCYGIHEMFNIVTSGEGKPHAVLI 100
                        ******** ****: ** .***:***********:*:*** .*:******

XDEL0C00540             RAVNPTEGIDIMMTRRNMAVIKPTITAGPGSVCKALGISRKINGISLQSN 150
ZP_07745263.1           RAIVPTDGIEEMLLRRDMPMLKPNITAGPGSVAKALGISRKINGFDLQGD 150
                        **: **:**: *: **:*.::**.********.**********:.**.:

XDEL0C00540             VLWIEDQGLNIPDDQVAVVPRIGIAYAGEDASLPYRFYVKDNVYVSKRVT 200
ZP_07745263.1           VIWIEDRGLSFSNDEIKAVPRVGVSYAAEDALLPYRLFVKGNPYVSKPNK 200
                        *:****:**.:.:*:: .***:*::**.*** ****::**.* ****  .

XDEL0C00540             KRVTKRVTKRIKEIGDTIGT 220
ZP_07745263.1           --------------------
```

Figure 3.2. Amino acid sequence alignment between *T. delbrueckii* XDEL0C00540 and *Mucilaginibacter paludis protein* ZP_07745263.1.

This gene from *M. paludis* encodes a DNA-3-methyladenine glycosylase, a DNA repair enzyme that initiates base excision repair by removing damaged bases to create abasic sites that are subsequently repaired (Lau et al. 2000). Interestingly, the *S. cerevisiae* gene *MAG1* (*YER142C*) also codes for a 3-methyladenine DNA glycosylase, and that has a syntenic ortholog in *T. delbrueckii* (annotated as XDEL0A10030). Even though they are predicted to have the same enzymatic activity, these two *T. delbrueckii*

proteins are unrelated in sequence and belong to different superfamilies. XDEL0A10030 and *MAG1* belongs to the ENDO3c or HhH-GPD superfamily of base excision DNA repair proteins. This superfamily contains a diverse range of structurally related DNA repair proteins, and is named after its hallmark helix-hairpin-helix and Gly/Pro rich loop followed by a conserved aspartate (Bruner et al. 2000). However, the DNA-3-methyladenine glycosylase encoded by the singleton XDEL0C00540 belongs to another family – alkyladenine DNA glycosylase (AAG) – that lacks the helix-hairpin-helix active site motif associated with other DNA glycosylases and is structurally distinct from them. The AAG family is ubiquitous in bacteria, but *T. delbrueckii* XDEL0C00540 is the only example of such a gene in a eukaryotic genome.

Table 3.2. List of genes gained by HGT in the seven newly annotated species

	Gene name	Best BLASTP hit against the nr protein database
T. del	XDEL0A00110	gb|AEH63371.1| short-chain dehydrogenase/reductase SDR [Zymomonas mobilis subsp. mobilis ATCC 10988]
	XDEL0B03240	XP_001549329.1 BC1G_11878 [Botryotinia fuckeliana B05.10]
	XDEL0B03250	GENE ID: 5999386 AOR_1_400024 | siderochrome-iron transporter [Aspergillus oryzae RIB40]
	XDEL0C00150	gb|ADW01250.1| GNAT family acetyltransferase [Lactobacillus phage LF1]
	XDEL0C00540	ZP_07745263.1| DNA-3-methyladenine glycosylase [Mucilaginibacter paludis DSM 18603]
	XDEL0D04410	GENE ID: 5124391 PGUG_05541 | hypothetical protein [Pichia guilliermondii ATCC 6260]
	XDEL0D06720	GLRG_06579 [Glomerella graminicola M1.001]
	XDEL0E00100	YP_001452115.1 hypothetical protein CKO_00523 [Citrobacter koseri ATCC BAA-895]
	XDEL0E00130	NP_595025.1|hydroxyacid dehydrogenase (predicted) [Schizosaccharomyces pombe 972h-]
	XDEL0C07090	similar to XDEL0E00130 (shorter)
	XDEL0F04790	XP_002490681.1 hypothetical protein [Pichia pastoris GS115] + XPHA0F00100, XPHA0F00110 and XPHA0H00100
	XDEL0F04900	CAY79383.1 EC1118_1F14_0078p [Saccharomyces cerevisiae EC1118], EGA80284.1 QA23_5165 [Saccharomyces cerevisiae Lalvin QA23]
	XDEL0G00190	YP_095107.1 hypothetical protein lpg1074 [Legionella pneumophila subsp. pneumophila str. Philadelphia 1],

	XDEL0C00200	gb|EGA80285.1 hypothetical protein QA23_5166 [Saccharomyces cerevisiae Lalvin QA23]
	XDEL0A01500	gb|EGA80285.1 hypothetical protein QA23_5166 [Saccharomyces cerevisiae Lalvin QA23]
	XDEL0B01160	gb|EGA80285.1 hypothetical protein QA23_5166 [Saccharomyces cerevisiae Lalvin QA23]
	XDEL0G01240	gb|EGA80285.1 hypothetical protein QA23_5166 [Saccharomyces cerevisiae Lalvin QA23]
	XDEL0D00170	gb|EGA80285.1 hypothetical protein QA23_5166 [Saccharomyces cerevisiae Lalvin QA23]
	XDEL0E01140	gb|EGA80285.1 hypothetical protein QA23_5166 [Saccharomyces cerevisiae Lalvin QA23]
	XDEL0F03640	gb|EGA80285.1 hypothetical protein QA23_5166 [Saccharomyces cerevisiae Lalvin QA23]
	XDEL0A06940	gb|EGA80285.1 hypothetical protein QA23_5166 [Saccharomyces cerevisiae Lalvin QA23]
	XDEL0F03030	gb|EGA80285.1 hypothetical protein QA23_5166 [Saccharomyces cerevisiae Lalvin QA23]
N. cas	NCAS0A09100	XP_001400751.2 2-amino-3-carboxymuconate-6-semialdehyde decarboxylase [Aspergillus niger CBS 513.88]
	NCAS0J00150	XP_001274208.1 Zn-dependent hydrolases of the beta-lactamase fold, putative [Aspergillus clavatus NRRL 1]
K. afr	XAFR0G03000	PICST_31085 | hypothetical protein [Pichia stipitis CBS 6054]
	XAFR0J00100	CAQ82464.1 arginase/agmatinase/formimionoglutamate hydrolase [Photorhabdus asymbiotica]
	XAFR0K01490	gb|EEX42832.1| phenolic acid decarboxylase [Vibrio furnissii CIP 102972]
	XAFR0K01500	CBA28397.1 Probable phenolic acid decarboxylase [Cronobacter turicensis z3032]
T. pha	XPHA0A03540	XP_462365.2| DEHA2G18986p [Debaryomyces hansenii CBS767]
	XPHA0L00660	XP_462365.2| DEHA2G18986p [Debaryomyces hansenii CBS767]
	XPHA0F03680	XP_462365.2| DEHA2G18986p [Debaryomyces hansenii CBS767]
	XPHA0F00100	XP_002490681.1 hypothetical protein [Pichia pastoris GS115] + XDEL0F04790
	XPHA0F00110	XP_002490681.1 hypothetical protein [Pichia pastoris GS115] + XDEL0F04790

XPHA0H00100 XP_002490681.1 hypothetical protein [Pichia pastoris GS115] + XDEL0F04790

Genes that have their names colored belong to the same family. For example, there are two gene families specific to *T. delbrueckii* (one with the two green genes, one with the nine blue genes). But in this species, there is also an orange gene very similar to a gene family in *T. phaffii* (the three orange genes).

3.4.1.3 A *T. delbrueckii* gene related to lignin degradation

XDEL0D06720 is another example of genes gained by HGT. In this case, the gene is closest to the chromosome's telomere. XDEL0D06720 has no apparent homolog in YGOB, but a BLASTP search against the NCBI nr database reveals many hits in species of filamentous ascomycete fungi (Pezizomycotina). The strongest hit is to a protein from the genome of *Glomerella graminicola*, a plant pathogen that causes anthracnose or canker (general terms for a large number of different plant diseases) in many cereal species. Figure 3.3 shows a protein sequence alignment between XDEL0D06720 and the gene from *G. graminicola* GLRG_06579, indicating very strong sequence conservation (76% identity).

```
XDEL0D06720 MSATSN-SKPIVFYDIAMRPPVEKNCCSPNPWKTRQALNFKGLRYSTSLVGLPDIPKVRE 59
GLRG_06579  MPTTPTPSSPIVFYDIAQRPPVEKTCCAPNPWKTRLALNFKGLPYSTSWVPLPDISKVRR 60
            *.:*.. *.******** ******.**:******* ******* **** * ****.***.

XDEL0D06720 GLNVPPCRKFPDGNDFFTLPIIEDPATNSLVGDSFDIAVYLQKTYPNAGVGDLFPSQKLD 119
GLRG_06579  SLKVSPCRKFADGSDFFTLPIIEDPATGSLVGDSFDIAVYLNETYPDSGAGDLFPPQALD 120
            .*:*.*****.**.*************.*************::***::*.*****.* **

XDEL0D06720 YVFSHDLVPAVPLSECNETDFPEYARFNMNVDAAFTVHVQLAIQCFPLDPATAETSKAEF 179
GLRG_06579  YTYEHDGI-LVPLSSCRDSAFPEYARFNMNVDAAFSMHVQLTVDGFPFDPATAETTKAEF 179
            *.:.** :  ****.*.:: ***************::****::: **:*******:****

XDEL0D06720 VRRAGATRWEDFALVGEQREKVKDSLRDTLGGLAKVFLRDTSGPFLLGTRASYADIIVGG 239
GLRG_06579  VRRAGVSSFEDFALVGERREQMKDSLRDTLGGLAKLFLRDPGGPFLLGTRASYADLIVGA 239
            *****.: :********:**::*************:****..*************:***.

XDEL0D06720 WLHMMNVTLPESEWRELRSWHEGLFGQLYDALEVFAEVK 278
GLRG_06579  WLRMMRATLPESEWEEVRTWHTGLFGQLHDALEAYAKVK 278
            **:**..*******.*:*:** ******:****.:*:**
```

Figure 3.3. Sequence alignment of XDEL0D06720 and GLRG_06579 (from *G. graminicola*).

The functions of the homologs of XDEL0D06720 in filamentous fungi are unknown. However, a search for known protein domains shows that these proteins

belong to the glutathione-S-transferase superfamily (containing GST-N and GST-C domains) and have a specific relationship to a subfamily called the "beta etherase LigE" subfamily. LigE is a bacterial GST-like protein that catalyzes the cleavage of the beta-aryl ether linkages, which is the most important process in the biological degradation of lignin (Masai et al. 2003). Lignin degradation is accomplished by the cooperative actions of fungal and bacterial enzyme systems: the degradation of native lignin is initiated by the attack of enzymes secreted by basidiomycete fungi such as *Phanerochaete* (lignin peroxidase, manganese peroxidase, and laccase); then the resulting low-molecular-weight lignin is further degraded and mineralized by bacteria (Masai et al. 2003). The function of the LigE-like gene in Pezizomycotina is unknown. Otsuka et al. (2003) determined that a novel fungus (strain name 2BW-1) found in soil was able to catalyze the cleavage of the beta-aryl ether linkages of high-molecular mass lignin. To determine the taxonomic position of the fungus they sequenced its 18S rDNA and found that train 2BW-1 is an ascomycete, probably a member of the genus *Chaetomium* (Otsuka et al. 2003). Some species of *Chaetomium* are able to grow on wood chips and decompose wood via the degradation of cellulose; however, the degradation system of lignin or lignin related compounds have not been studied in these species. Interestingly, one of the best hits in our BLASTP search is a gene from the genome of *Chaetomium thermophilum* (E-value: 1e-122; 62% amino acid sequence identity).

The function of the LigE-like gene in *T. delbrueckii* is also unknown, but it is clear that this gene originated by HGT of a Pezizomycota gene into the *T. delbrueckii* genome. There are no reports of lignin-degradation by *T. delbrueckii*.

3.4.2 Multiple duplications in *Tetrapisispora blattae*

We saw previously that *T. blattae* has more singletons than all the other studied species (**Table 3.1**). One of the reasons for this high number is the presence of large duplicated gene families, and some of these are discussed below. This yeast has been isolated from the intestines of a cockroach (*Blatta orientalis*). Whether or not *T. blattae* is an intestinal symbiont of the oriental cockroach is not known, as there have been no further reports of its isolation from that or any other habitat.

3.4.2.1 Duplicated genes of the *HSP30* family

We discovered in the *T. blattae* genome a family of twelve genes related to the small heat shock protein Hsp30. Two of them (XBLA0D00670 and XBLA0B02770) belong to the same YGOB pillar (meaning they are WGD-paralogs of each other). This pillar contains the ancestral gene Anc_8.794 and genes from most yeast species but none from *S. cerevisiae*. The other ten genes (XBLA0D00680, XBLA0D00690, XBLA0B01690, XBLA0B01700, XBLA0B01710, XBLA0B01720, XBLA0B01730, XBLA0B06820, XBLA0B06880 and XBLA0B06890) are singletons (meaning they do not have syntenic orthologs in the species present in YGOB, they are species-specific duplicated genes). Most of them are arranged in tandem repeats (gene numbers 10 units apart indicate consecutive genes on the chromosome). Moreover, two other species also have duplicated singletons from this family: there are two duplicated singletons related to HSP30 in *N. castellii,* and one in *N. dairenensis*. The sequence similarity among the twelve genes is quite well conserved, as shown in the protein alignment in **Figure 3.4**.

```
XBLA0B01720      -MSSTGDFVELVKRSGNRAVHVNPPHGVDIHLTEHGSDWLWSVYSVYGVLAIMYACFFFY 59
XBLA0B06820      -MS--GEFVELVKKSGNRAVEVNPPTGLDVHLTKSGSDWLWTVYSIFGVLAVLYACFFFY 57
XBLA0B01690      MSSANGGFIEFIKKAGNRAVHVNPPHGVDIHLTESGSDWLWSVYSVFGVLSIMYAVFFFY 60
XBLA0B01730      -MSANGEFVELLKR-GNRAVHVNPPHGLDIHLTESGSDWLWSVYSVFGVLSIMYACFFFY 58
XBLA0B01710      -MTATSDFVEIVKRSGNRAVHVNPPHGLDIHLTEHGSDWLWSVYSVFGVLSIMYAVFFFY 59
XBLA0B06880      -MSSDTEYLELIKRSGNRAVHVNPPHGLDIHLTESGSDWLWSVYSVFGVLAIMYACFFFY 59
XBLA0B06890      -MSTTGDFIELVKRTGNRAVKVNAPHGLDIHLTESGSDWLWSVYSVYGVLAIMYACFFFY 59
XBLA0B01700      -MSVNGGFVEIAKRSGNRAVHVNPPHGLDIHLTEHGSDWLWSVYSVFGVLTLMYACFFFY 59
XBLA0D00680      ----------------------------------------------------MYCVFFFF 8
XBLA0D00690      ----MSQEFIELVRRGNRAVHVNPPKDLDVHLTEHGSDWLWSVYSVFGVLAVMYATFFFY 56
XBLA0D00670      -----MSYIDLIKRGGNQAVRVNPPHMLDIHLTEHGSDWLWSVYSVFGVLSVLYAIFFFY 55
XBLA0B02770      MSSLENGYVNLFKR-GNRAVHVNPPTGTDIHLTESGSDWLWTVFSVFGFLGGIYAIMFFY 59
                                                                 :*. :**:

XBLA0B01720      QQFKSGS-QLVRYTMAGPFLISLFQFFQYYTYASNLGWTAIQAEFKGEDISDPVTDLSPQ 118
XBLA0B06820      QEMKNGS-QLVRYTMAAPFLISLFEFFQYYTYASNLGWTGIQAEFKGEDVSDPVSGESPN 116
XBLA0B01690      QEYTNGT-QLIRYTMAAPFLISLFEFFQYFTYASNLGWTSIQAEFRSEDVSRPVTDLYPN 119
XBLA0B01730      QEMKGGS-QLIRYTMAGPFLISIFEFFQYYTYASNLGWTNIQAEFRGEDVSHPVTDLYPN 117
XBLA0B01710      QEYRNGS-QLIKYTMAGPFLISIFEFFQYYTYASNLGWTNIQAEFRGEDVSHPVTDLYPN 118
XBLA0B06880      QEMKNGS-QLVRYTMAGPFLISLFEFFQYFTYASNLGWTAIQAEFRSEDVSHSATNMYPN 118
XBLA0B06890      QEFKNGR-QLTKFALAGPFLISLFEFFQYYTYASNLGWTDIQAEFRSEDVSHSVTNEYPN 118
XBLA0B01700      QEYNHGS-QLVRFAMAGPFLISLFQFFQYFTYASNLGWTDIQAEFRSEDVSHSITDEYPN 118
XBLA0D00680      QEYYDGT-QLIRYAMAGPFLISLFQFFQYFTYASNLGWTNIQAEFRSEDVSRPVTDEYPN 67
XBLA0D00690      QQLKGGR-QLIRYTLAGPFLISIFQFFQYFTYASNLGWTNIQAEFRGEDVSRPVTNMYPN 115
XBLA0D00670      QEMKSGS-QLIRYTMAAPFLISIIQFFQYFTYASNLGWCIIQAEFRGEDVSQSVTNMYPN 114
XBLA0B02770      VENKKEYNKLQRYTIAGPFIICIFESFAYYTYASNLGWGGIQAEFRGEDVSHSITNKYPN 119
                  :      :* ::::*.**:*.::: * *:******** *****:.**:* . :.  *:

XBLA0B01720      VRQIFYAKYVSWFLCWPIQLFMFELAGFTIKNSVEFYS-----TWEMIHSLLVQIIGFEY 173
XBLA0B06820      IRQIFYAKYIAWFLCWPIQLFMVELAGVTINHQGEPLS-----VWDIMHSLIVQVIGFEF 171
XBLA0B01690      VRQIFYAKYVAWFLCWPIQLFLVELAGITINHSGEVFS-----TWELIHSLLVQIIGFEF 174
XBLA0B01730      VRQIFYAKYVAWFLCWPIQLFLLELAGITINHSGEVFS-----TWELIHSLLVQMIGYEY 172
XBLA0B01710      VRQIFYAKYVAWFLCWPIQLFLMELAGITINHSGEVFS-----TWELIHSLIVQIIGYEY 173
XBLA0B06880      MRQIFYAKYVAWFLCWPIQLFMLEMAGVTINHTAQVFS-----TWEMIHSLLVQIIGFEY 173
XBLA0B06890      VRQIFYAKYVAWFLCWPIQLFMFELAGSTINHTGEFFS-----TWEMIHSLLVQIIGFEY 173
XBLA0B01700      VRQIFYAKYVAWFLCWPIQLFLLDVAAVTINKEVQYFS-----TWEMIHQLFVQIIGFEY 173
XBLA0D00680      VRQIFYAKYVAWFLCWPIQIFMLELAGVTVR-EGEFVT-----TWEMIHSLIVQIISYEY 121
XBLA0D00690      VRQIFYAKYVAWFLCWPIQIFLLELAGITIDHQGEVLS-----CWALIHSILVQIIGYEY 170
XBLA0D00670      NRQIFYAKYIAWFLCWPIQLFLVEMTAFTVNRNNDSVPTATYGAWGMIHSLLVQIIGFEF 174
XBLA0B02770      VRQVFYCKFIGWVLSWPVLYFLNEMASLCVH-AGENID---FFFWDIAHSLLVQILGTTY 175
                  **:**.*::.*.*.**:  *: :::.  :    :         * : *.::**::.  :

XBLA0B01720      WVISLLVGSLIRSSYKWGYWTMGAFVMLVVLGIQFKRQVFDLKLRGVSLGMYCTAIIVIL 233
XBLA0B06820      WIIALLVGALIRSSYKWGYWTMGVFVSLVVLGIQFKRQVFDLRTRGLSLAMYFVAIIVIL 231
XBLA0B01690      WVISLLVGSLIHSTYKWGYWTMGIFVSFCILGVQAKRQLFDLKLRGVSLIIYVLAVICIL 234
XBLA0B01730      WVISLLVGALIHSTYKWGYWVMGIFVSFCVLGVQFKRQIFDLKIRGVSLIVYLLSIICTL 232
XBLA0B01710      WVISLLVGALIHSTYKWGYWVMGVFVSLFILGIQIKRQVFDLKVRGVSLVVYLLSIICTL 233
XBLA0B06880      WVISLLVGSLIHSTYKWGYWVMGIFVAFVVLGIQAKRQIGDLQIRGVSLGMYIVAVICIL 233
XBLA0B06890      WVISLLVGSLIHSTYKWGYWVMGIFVLFVVLGIQFKRQIFDLKIRGVSLAMYCTAIIVIL 233
XBLA0B01700      WIISLLVGALIHSTYKWGYWVTGIFVALIVLGIQFKRQVFDLKIRGISLGMYCTAIIVIL 233
XBLA0D00680      WIISLLVGSLIHSTYKWGYWTMGIFVLFVVMGVQFKRQVFELKIRGISLWMYLTSMIIVL 181
XBLA0D00690      WIIALLVGSLIHSTYKWGYWVMGVFVLLCTLGIQCKRQVFELKIRGFSLAMYLTAIIVTL 230
XBLA0D00670      WVICLLVGALIHSTYKWGYWVMGIFVLLCVLGIQIKRQIFELKLRGVTLCVFFIAVICVL 234
XBLA0B02770      WIVCLLVGILIHSTYKWGYWVMGVFVLLTIEGIKLKRQIVDLKIRGFQLCVFLTHTIIVL 235
                 *::.**** **:*:******. * ** :   *:: ***: :*: **. * ::    *  *
```

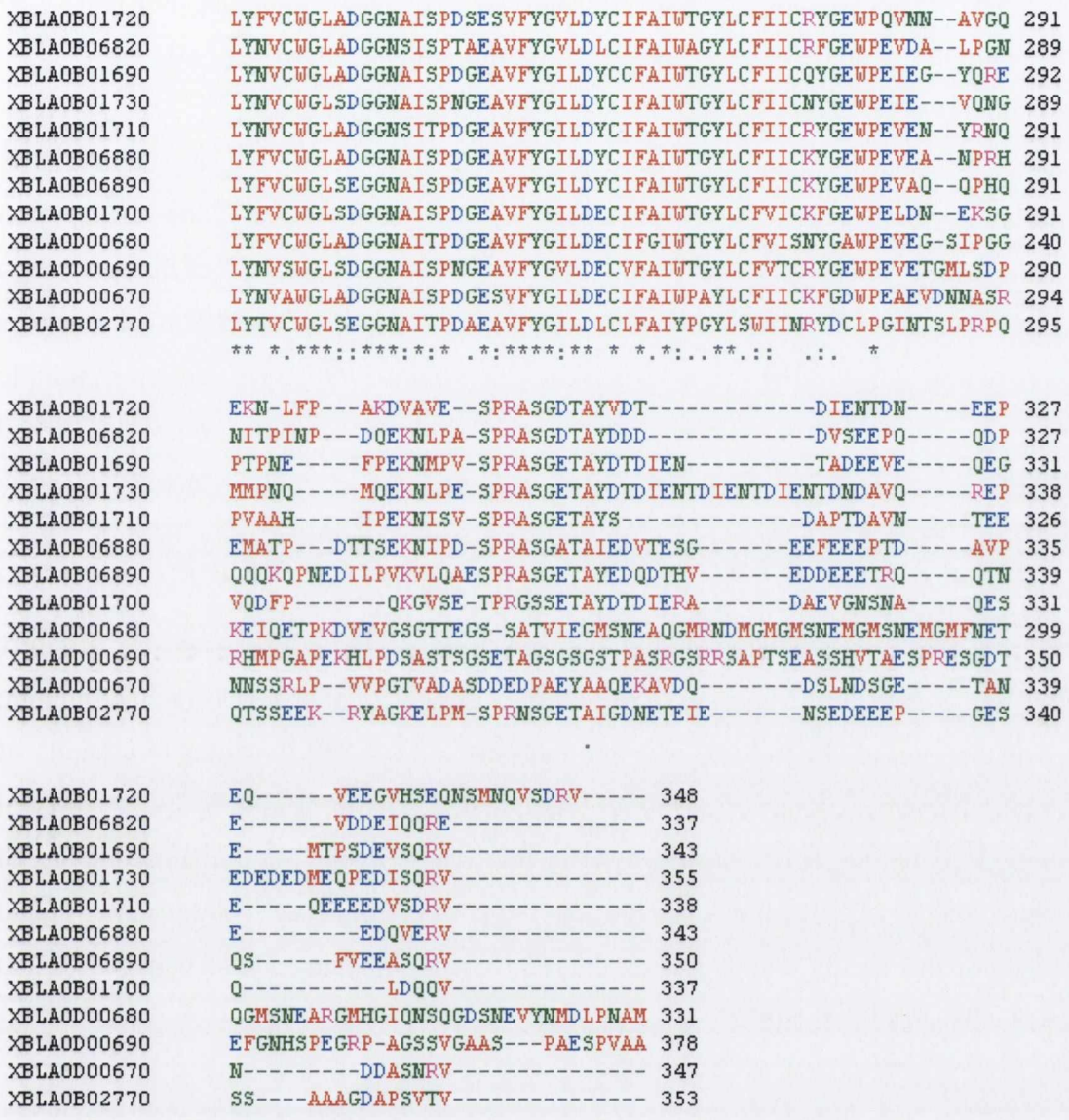


Figure 3.4. Amino acid sequence alignment of the twelve *T. blattae* genes related to the HSP30 family.

There are neither *S. cerevisiae* nor *S. bayanus* genes in the pillar containing the ancestral gene Anc_8.794. However, these two species each contain a gene (*HSP30/YCR021W* and *Sbay_675.56*) that is highly similar to the Anc_8.794 pillar but located elsewhere in the genome (between Anc_1.437 and Anc_1.451), hence corresponding to a site of rearrangement in the *Saccharomyces* species. These two genes are likely to be orthologous to the genes in the pillar Anc_8.794. *S. cerevisiae* Hsp30 is a hydrophobic plasma membrane localized, stress-responsive protein that

negatively regulates the major proton pump Pma1 (Piper et al. 1997); it is induced by heat shock, ethanol treatment, weak organic acid, glucose limitation, and entry into stationary phase (Panaretou and Piper 1992; Regnacq and Boucherie 1993).

One might hypothesize that amplification of the *HSP30* gene family in *T. blattae* might have occurred in response to a similar amplification of the *PMA* proton pump genes. In fact, *T. blattae* contains only two *PMA* genes. There does however seem to be a weak correlation between the numbers of *PMA* genes and *HSP30* genes across species (**Table 3.3**): all the species with more than one *HSP30* gene also have multiple *PMA* genes. However, the factor(s) that led to the dramatic amplification of the HSP30 family in *T. blattae* remain unexplained, and we cannot even be certain that the amplification is statistically significant (see discussion of the PIR family, below).

A search for known domains showed that the Hsp30 proteins are distantly related to members of the bacterial opsin family. The bacterial opsins are retinal-binding proteins that provide light-dependent ion transport and sensory functions to a family of halophilic bacteria (Blanck et al. 1989; Oesterhelt and Tittor 1989). They are integral membrane proteins predicted to have seven transmembrane (TM) domains, the last of which contains the attachment point for retinal (a conserved lysine). The fungal members of this family do not contain the retinal binding lysine and so cannot function as opsins (Blanck et al. 1989).

Table 3.3. Numbers of *HSP30* genes and *PMA1*-like genes among Saccharomycetaceae species (from YGOB).

Species	PMA1-like genes*	HSP30 genes†
T. blattae	2 (2)	12 (2)
T. phaffii	1 (1)	0 (0)
V. polyspora	1 (1)	0 (0)
N. dairenensis	2 (1)	3 (2)
N. castellii	2 (1)	4 (2)
K. naganishii	2 (2)	1 (1)
K. africana	2 (1)	2 (2)
C. glabrata	1 (1)	1 (1)
S. cerevisiae	2 (1)	1 (0)

S. bayanus	2 (1)	1 (0)
Z. rouxii	1 (1)	1 (1)
T. delbrueckii	1 (1)	1 (1)
K. lactis	1 (1)	0 (0)
E. gossypii	1 (1)	0 (0)
L. kluyveri	1 (1)	1 (1)
L. thermotolerans	1 (1)	1 (1)
L. waltii	1 (1)	1 (1)

* Numbers in parentheses show the number of genes from each species in the ancestral pillar Anc_4.115.
† Numbers in parentheses show the number of genes from each species in the ancestral pillar Anc_8.794.

3.4.2.2 Duplicated genes of the PIR family in *T. blattae*

In *S. cerevisiae*, the PIR family is a family of proteins that contain a sequence of 18-19 amino acid residues repeated tandemly seven to ten times (Toh-e et al. 1993). The genes in this family in *S. cerevisiae* are *PIR1* (*YKL164C*), *PIR2* (*HSP150* or *YJL159W*), *PIR3* (*YKL163W*), *PIR4* (*CIS3* or *YJL158C*) and *YJL160C*. These genes are arranged as two tandem repeats: one on chromosome X containing three genes, and one on chromosome XI containing two genes. These two arrays are paralogous products of the WGD (**Figure 3.5**) and correspond to an ancestral tandem array from Anc_1.187 to Anc_1.189. All the yeast species in YGOB have arrays of *PIR* genes, and these are also found in the species from the CTG clade.

T. blattae also has genes annotated into those three YGOB pillars. However, contrary to the other species in YGOB, *T. blattae* has an additional 13 *PIR* gene singletons scattered around its genome (**Table 3.4**). Analysis of the locations of these genes shows that eight of them coincide with sites of structural rearrangement in the *T. blattae* genome. Six *PIR* genes are located within small groups of ~5 genes that seem to be additions to the *T. blattae* genome at sites of translocations between chromosomes. Two others coincide with endpoints of local rearrangements, most likely inversions. Of the five *PIR* genes not associated with larger rearrangements, four were added to the *T. blattae* genome as simple insertions of one gene between two neighboring ancestral genes, and one was gained at a telomeric location.

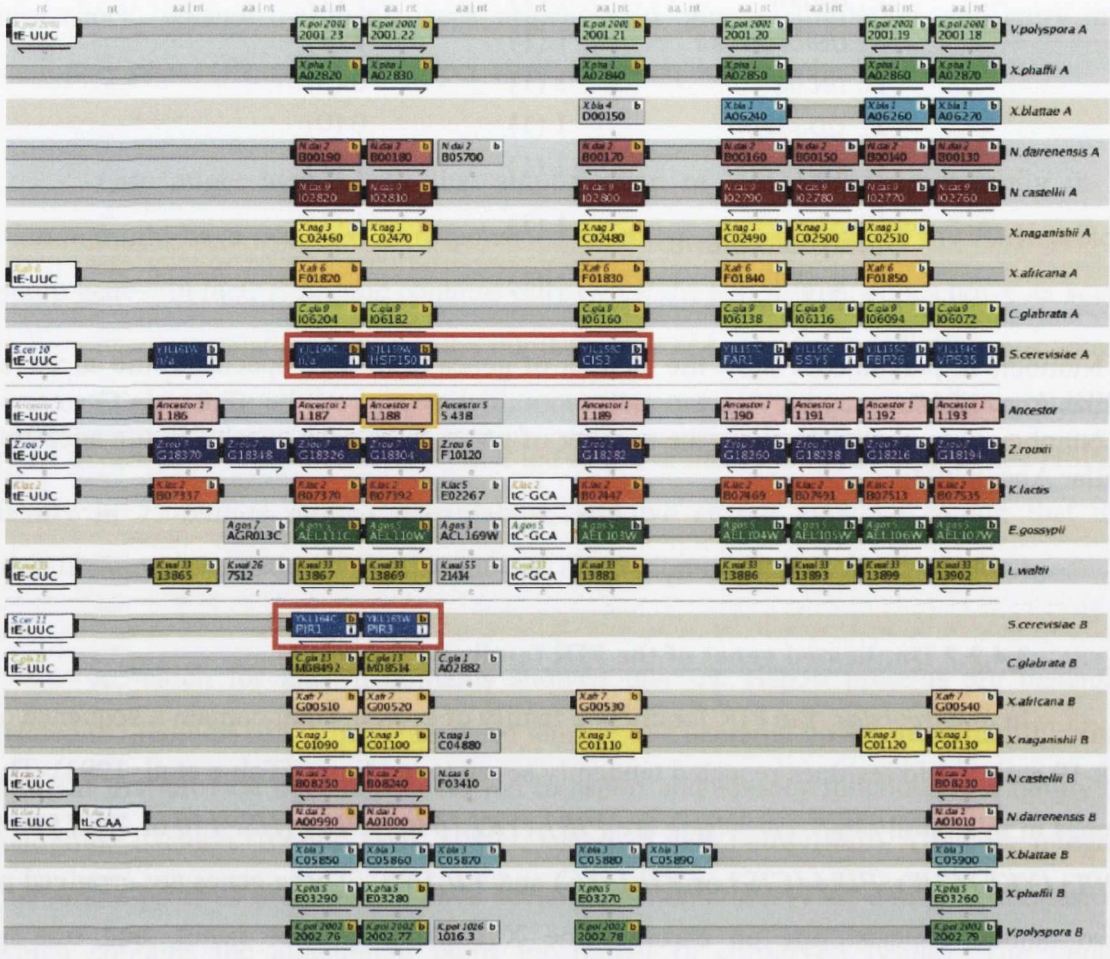


Figure 3.5. Visualisation of the *PIR* genes in YGOB. The five *S. cerevisiae PIR* genes are arranged as two tandem repeats (framed in red) on chromosomes X (upper repeat) and XI (lower repeat).

The PIR and HSP30 families have both expanded substantially in *T. blattae*. It is tempting to speculate that their amplification is the result of natural selection, either for increased sequence diversity within each family, or simply for increased dosage of protein expression. However, neither of these families shows evidence of positive selection on their sequences, which rules out the first possibility. Using the program yn00 from the PAML package (Yang 1997) with all possible pairs of PIR sequences, and all possible pairs of HSP30 sequences, showed that the Ka/Ks ratios were less than 1 for all comparisons (maximum values were 0.5478 for PIR, and 0.2321 for HSP30 families). These calculations were made on improved sequence annotations to minimise sequencing errors. It is not easy to design a test of the second hypothesis. One would need to show that the degree of amplification seen in the PIR and HSP30 families of *T.*

blattae lies significantly outside the normal range of gene family amplifications seen among other gene families in the whole genomes of this set of species. As far as I know no previous studies have developed standard methods for this type of statistical quantification of family amplification. We did not attempt to develop such a method because, even if the amplifications in *T. blattae* could be shown to be statistically significant, the selective forces causing their amplification would remain unidentified. It is however worth noting that both the *PIR* and the *HSP30* genes are known to be stress-indicible in *S. cerevisiae*.

Table 3.4. Locations of 16 *PIR* gene homologs in *T. blattae*.

Gene	Anc pillar	Location
XBLA0A00870		Translocation site between Anc_2.606 and Anc_7.171
XBLA0A01190		Translocation site between Anc_7.158 and Anc_1.384
XBLA0A04270		Translocation site between Anc_8.370 and Anc_7.17
XBLA0A06230		Translocation site between Anc_6.221 and Anc_1.190+
XBLA0B00410		Translocation site between Anc_2.48 and Anc_1.58
XBLA0B07280		Singleton insertion between Anc_2.346 and Anc_2.347
XBLA0C05850	Anc_1.187	
XBLA0C05860	Anc_1.188	
XBLA0C05880	Anc_1.189	
XBLA0D00150		Telomeric†
XBLA0F00670		Local rearrangement site between Anc_3.195 and Anc_3.216
XBLA0G01080		Local rearrangement site between Anc_6.10 and Anc_6.21
XBLA0H02860		Singleton insertion between Anc_5.217 and Anc_5.218
XBLA0H03100		Translocation site between Anc_1.246 and Anc_8.373
*XBLA0I01410/1420**		Singleton insertion between Anc_2.440 and Anc_2.441
XBLA0J01310		Singleton insertion between Anc_7.269 and Anc_7.272

* Probable frameshift sequencing error in the analyzed version of the *T. blattae* genome sequence.
† Erroneously assigned to Anc_1.189 in current release of YGOB.
+ Should be assigned to one of the ancestral pillars Anc_1.187/188/189

3.4.3 Orphan genes in the new species

The "orphan" category accounts for a large proportion of the singletons in most species (Table 3.1). Orphans are genes that have no syntenic orthologs in the other yeast species present in YGOB, and have neither homologs in any other species present in the NCBI nr database nor known protein domains. For example, in *K. naganishii*, 46% of

singletons are classified as orphans. In *N. castellii*, among the 167 orphan genes, 68 of them (41%) belong to a family that contains at least two members. However, in *T. blattae*, only 14% of the orphan genes are part of a family. Small families (from two to five genes) are more frequent, but larger families (up to twenty members in *N. castellii*) are also found (**Figure 3.6**).

```
NDAI0K03020     --------------------MDLVTVNGLKLRKPFSESRRKSNIWWFLPNYFSN-----N 35
NDAI0B00560     --------------------MDLVTVNGLKLRKPFSESRRKSNIWWFLPNYFSN-----N 35
NDAI0I00120     --------------------MDFVTVNGLKLRRPFSESRRKNNIWWFLSNYFSN-----N 35
NDAI0B00510     ------------------------------------------------------------
NDAI0K00130     --------------------MEFAFINGLKLRKPFPESRIENKIWGFLSNCFSN-----S 35
NDAI0D03670     MTRDQEDVGGKEVGTPFFSLQKSTTQQGIPLRRPFGK-RNICQRWEVLKGLVSG-----S 54
NDAI0J00210     ----------MKNNYNYQEKMNSNTSQGLVLREKAAKN-MVLRIYTQLKRKIP------- 42
NDAI0J00770     ----------MKNNYNYQEKTGSNTSQGLMLREKAAKN-MVLRIYTQLKRKIP------- 42
NDAI0J00400     ------------------------------------------------------------
NDAI0C06320     --------------------MVSDPNDGLLLREKASK----MSIYTRLKQKVP------- 29
NDAI0J00540     --------------------MSSNTNRCLMVRKKALPMNIFRYEYTKLKEKLS------- 33
NDAI0K02490     MTFCKPIIYSSPHGLPLIEYHEKFKQPSKEKKKAYGTDSLRKKIFNFFTKNIPSEEREKT 60


NDAI0K03020     STSTANTFSVKQDNWILDEPTTFALDNEPSSHWIFKECIMKYFLTDTGRSFP-------- 87
NDAI0B00560     STSTANTFSVKQESWILDEPTTFALDNEPSSHWIFKECIMKYFLTDTGRSFP-------- 87
NDAI0I00120     STPAPTSYSAKGDNWILDEPTTFALDYGEWSNTIFKEYIMKYFLIDKGRNFL-------- 87
NDAI0B00510     ------------------------------------MKYFLTDTGRSFP-------- 13
NDAI0K00130     STPSTNTYSVKQESWILDEPTTFALDNRVSSSWIFKECIMKYFFTDKGRSFP-------- 87
NDAI0D03670     TNITQEANMLTDNDWIMDEPFTFGISNDYSLAVIYRDYIMKHYLVSPEYFVN-------- 106
NDAI0J00210     ------FCKKPVKYWILDEPNSFKVCP-YELGFIFKYDVVRHYLKSNTPFYSSVSNTEYK 95
NDAI0J00770     ------FCKKPVKYWILDEPNSFKVCP-YELGFIFKYDVVRHYLKSNTPFYSSVSNTEYK 95
NDAI0J00400     ------------------------------------------------------------
NDAI0C06320     ------FCKKPVKYWIIAEPNSFKGCP-IRLCTLFKCYIVHNYLKSDTPLYSSYSNTDYE 82
NDAI0J00540     ------FCKSEV-YWILDEPNSFKLCP-REICYLFKYCLMYHYLKSSTPFYSSVFNTEYR 85
NDAI0K02490     VTKKALLENPGGKDFILDDPFIFILSQGKKISSLYRNTFTKESCKRYEAPLLC------- 113


NDAI0K03020     --------GTVVEAPFHKLMLKGGNFTLLVLGPYLFEEEGFYSCNIEEPGKELLKTKFDF 139
NDAI0B00560     --------GTVVEAPFHKLMLKGGNPALLVLEPYLFEEEGFYSCNIEEPGKELLKTKFDF 139
NDAI0I00120     --------ETVVETPFHKLMLKGGAPTLLVLGPNLYEEQAFYSCDIEEAGKDLLCSEYTL 139
NDAI0B00510     --------GPVVEAPFHKLMLKGGNPTLLVLGPNLYKEQAFYSCDIEETGKELLNIKYEI 65
NDAI0K00130     --------NTMLETPFHKQMLKGDTSTLLVLGPNLFKEEGFYSCDIEEAGKELIKTKYHF 139
NDAI0D03670     --------HMGSDFPIDDIIREGNQHTLVVVGNRLEIQTAYLSCDIDEKGKHGLN-DCPF 157
NDAI0J00210     PRFPFHRKRSKMYDPQYFLKNGISYNCYMLVGPDMGNRTIYFSTAISEDGMLELNISLSL 155
NDAI0J00770     PRFPFHRKRSKMYDSQYFLKRGRFYKCYMLVGPDMGNRTIYFSTAISEDGMLRFNNLLTV 155
NDAI0J00400     --------------------------MLVGPDMGNRMIYFSTAISEDGMLELNISLSL 32
NDAI0C06320     ----FYKRMSRLDDPESYFRNIFLSMSYMMVGRDSDQRKFYFSTTITEDAKAELTKELST 138
NDAI0J00540     SLFSFYRRKSQLYDPHYYLRDDHSYWAYLVRGPDTGNGKLYFSTTISEEGIDGFNKSVSG 145
NDAI0K02490     --------DSNYEKKLDQYLDQYKKPILLIFGLNFNSKAIFLPTSLDKMSEDNFKMTNQS 165
                                          ::          : .  : : .   :


NDAI0K03020     EYPELDSLLCRIHIYTDVQYEGYSYGASYKASENSMKPTK-MLFLPPSQYSSIPLQAIFY 198
NDAI0B00560     EYPELDSLLCRIHIYTDVQYEGYSYEASYKASENSIKPDK-ILFLPPSRYSNIPLQAVFY 198
NDAI0I00120     LGTVLQVFIRHIHYHTNVQYEGYSYKGSYKISERPVKPNR-LLYLPPSQYSNIPLQAIFY 198
NDAI0B00510     GNLELNGFLRYIHYRANVQYEGYCYKGSYKIHEHPMKPNK-ILYLPPSQYSSIPLQAIFY 124
NDAI0K00130     SAVDFWGLLQSIHLLTNVQYEGYCYKECCKLSRTLGKPTK-MLFLPPSQYSSIPLQAIFY 198
NDAI0D03670     FVPNVKDLLVSMERVQAIIYKSDNYKKLQEMPFSPFQKSHSFDFLPPSPYSGLPIQAIFF 217
NDAI0J00210     YKIPTSNSDIFANLNIHNGDWERCN--DMPRLADLLNENE-VGFLPPSDFAPVPCQVLFY 212
NDAI0J00770     YKIPTSNSDIFSNLNIYNGDWERCN--DMPDLADLLNENE-VGFLPPSDFAPVPCQVLFY 212
NDAI0J00400     YKIPTSNSDIFANLNIHNGDWERCN--DMPRLPELLSENE-VGFLPPSDFAPVPCQVLFY 89
NDAI0C06320     YAPTPDISSAFFDVELS-GRCCRYPPTGVLPLLDLLQNND-IAFLPPSDYGSVPCEVFFY 196
NDAI0J00540     YKNEPP-PLGFIDLNIFNGEIKASE--DMPSLPAILGDND-IEFLPPSDFGPTPCEVLFY 201
NDAI0K02490     GTNISDMINAAITYSELVQFQGSERDYRFEIIRSLGVQSNQFEFLPPSEYSNFPLPIIFL 225
                                                  . :**** :.  *   .*
```

```
NDAIOK03020     GFSKEDAIIYRDLLVVYRIRSIRRGYRPTIKNVLNVFFDPPYEIYVIP----SPHGSDST 254
NDAIOB00560     GFSPEDAIIYRDLLIVYRIRAIRRGYKPTMDKVLSVFFDPPYEIYVIP----SPHGNDSK 254
NDAIOI00120     GFSQEDAIIYRDLLIVYRIRAIRRGYKPTMDKVLSVFFDPPYEIYVIP----SPHGDNSK 254
NDAIOB00510     GFSPEDAIIYRDLLIVYRIRAIRRGYKPTMDKVLSVFFDPPYEIYVIP----SPHGNDSK 180
NDAIOK00130     GFSKEDAIIYRDLLIVYRIRAVRRGYKPTMTNVLNVFFDPPYEIYVIP----SMHGSDSK 254
NDAIOD03670     HCTELEAETYRELLVVYRNRVIRQHDPPTVERVMRLFLNPPYEIYVRPSFI-TGKDQESA 276
NDAIOJ00210     HFQESDAIIYRDLFVVYRNRAIKRGYMPNFPSILEVFLDPPYEIFIMNVDDNGNILQDKK 272
NDAIOJ00770     HFQESDAIIYRDLFVVYRNRAIKRGYMPSFPSILMVFLDPPYEIFIMNVDDNGNILQDKK 272
NDAIOJ00400     HFQESDAIIYRDLFVVYRNRAIKRGYMPSFPSILMVFLDPPYEIFIMNVDDNGNILQDRK 149
NDAIOC06320     HFREPEAIVYRDLFMVYRSRAVKRGYSPGFHNILKVFHDPPYEIFITNVDENGKILEDRK 256
NDAIOJ00540     HFKKSDAIVYRDLFIVYRNRAIKRGYLPNFDGILKVLLDPPYEIFITDIAIKGEPVNGRK 261
NDAIOK02490     QCHLRHAIIYRDFLIVCRNKVMREHREPSFENILKIFNDPPYEITMELS-----PNEGIE 280
                     .*  **:::: * : ::.   * .  :: :: :***** :          .

NDAIOK03020     QEYIEVDHCEDVTDEERQTYQA-YFMLKMKYMLGDNLEYST--------FSELLESRNHD 305
NDAIOB00560     QENMEVDHCEDVTEEKRRTHQA-YFMLKMKYMLGDKLKYST--------FSELLENTNHA 305
NDAIOI00120     QEYIEVNHCEDVTEEERQTYQA-YFMLKMKYMLGDKLEYST--------FSEVLENTNHA 305
NDAIOB00510     QEYMEVDHCEDVTKEERQTYQA-YFMLKMRYMLGDKLEYST--------FSELLENTNHA 231
NDAIOK00130     EEYIEVNHSEDVTEEERQTYQA-YFMLKMKYMLGDKLNYST--------FSELLEDTNHA 305
NDAIOD03670     DEFSHQEHSSDVTCEERKTYQY-YTALKGKNWLDMIKREKE--------LEETFLDACGL 327
NDAIOJ00210     EEYNEVYHDCDVTWAERHCYQK-YFVCKVNQMIEGSSRYFT--------LSYLQDHLERN 323
NDAIOJ00770     EEYNEVYHDCDVTWAERHCYTKNYFVCKVNQMIEGSSRYFT--------LSYLQDHLERN 324
NDAIOJ00400     EEYNEVYHEHDVTWAERHCYQK-YFVCKVNQMIEGSNSYFT--------LSYLQDHLERN 200
NDAIOC06320     ERYNEVYNERDVTWAERHCYQK-YTVCKVNRMRKWSSSYFT--------LSFLQDHLGRN 307
NDAIOJ00540     KKYSEVYHDRDVTWAERHCYQR-YFVCKFNQMRKGSSSYFT--------LSFLQEHLEGN 312
NDAIOK02490     PKFYEWERERDITWAERHCYQQ-YFLLRAKLMSYDRVCFPDRIEPREEGLEEIQKEDDSF 339
                 .  .  .  *:*  :*: :   *   : .                    :.

NDAIOK03020     HSESVYLEESNHFESQV---------------------- 322
NDAIOB00560     YFESIFLEESNDFES------------------------ 320
NDAIOI00120     YCESVHLEEV----------------------------- 315
NDAIOB00510     YCESVHLEEV----------------------------- 241
NDAIOK00130     YSESVYLEESNPFESQV---------------------- 322
NDAIOD03670     RL------------------------------------- 329
NDAIOJ00210     PVFHRKGKTGYKITTPFPEFT------------------ 344
NDAIOJ00770     PVFHRKGKTGYKITTPFPEFTY----------------- 346
NDAIOJ00400     PVFHRKGKTGYKITTPFPEFT------------------ 221
NDAIOC06320     PVFHRKGKNRYKIMARSTESS------------------ 328
NDAIOJ00540     PVLRRKNKATYIVSTSCH--------------------- 330
NDAIOK02490     YVQKKGGEHPIMDVYGKYVPNNLENQATMNYHKKGYKL 377
```

Figure 3.6. Amino acid sequence alignment of orphan genes belonging to the same gene family in *N. dairenensis*.

As stated in the introduction (part **3.2**), two scenarios have been suggested to explain the emergence of orphan genes: the duplication-divergence scenario, and the *de novo* scenario. All the orphans in the **Table 3.1** are species-specific genes. We have found very few orphan genes that are for example clade-specific genes (genes found only in species from 1 of the 12 clades of the Saccharomycetaceae family). This fact could be explained by the duplication/divergence scenario: Consider a gene family that is evolving rapidly by duplication and divergence within an ancestral species A, and species A then splits into descendant species B and C. It is unlikely that the rapid evolution will suddenly stop (but if it does, each family member in B will have a single

ortholog in C). It is more likely that the process will continue, and will gradually make the families appear more and more species-specific, and less and less clade-specific. The few clade-specific orphan genes we found share a syntenic location, but they usually lack much sequence similarity (less than 25%). It is therefore not easy to determine whether they are clade-specific orphan genes, or independent species-specific orphan singletons. However, there are some exceptions: for example, NCAS0B07970 and NDAI0B05350 are orphan genes specific to the genus *Naumovozyma*. They share the same syntenic location (between Anc_1.226 and Anc_1.227) and have 53% amino acid sequence identity.

Chapter 4

Evolution of the location of the ribosomal DNA array in Saccharomycotina

4.1 Abstract

In this chapter we study how the location of the large ribosomal DNA (rDNA) array has changed during yeast evolution. The rDNA is a huge locus of roughly 1 million basepairs, made up of numerous tandem copies of ~10 kb unit containing the rRNA genes. We show here that it has moved within the genome several times. We identify an ancestral rDNA location at an internal (non-telomeric) site on Ancestral chromosome 8 that is still used in *S. cerevisiae*, some other post-WGD species, and some non-WGD species. We show, however, that the rDNA has completely disappeared from the ancestral site on multiple independent occasions. In four separate lineages, the rDNA moved from the ancestral site to a subtelomere. In *L. thermotolerans* (and probably also *L. waltii*), rDNA disappeared from the ancestral site and integrated into a new non-telomeric site that previously contained only a tRNA gene. Remarkably, two of the sites of rDNA arrays (the ancestral one and the derived one in *L. thermotolerans*) are right beside genes (*HMO1* and *CDC14*) coding for proteins that maintain the specialized chromatin structure of rDNA. This observation suggests that the location of the rDNA array is determined by natural selection. The mechanism by which this large and complex locus moves around the genome is unknown but may involve the formation of extrachromosomal circles.

4.2 Introduction

Eukaryotes have four types of ribosomal genes specifying 5S, 5.8S, 18S and 25S rRNAs. In most eukaryotes, the 5S gene is not part of the 18S-5.8S-25S tandem rDNA array (**Figure 1.4**) (Hillis and Dixon 1991). However, in fungi, the 5S gene is usually present within the rDNA array (Rubin and Sulston 1973; Ganley and Kobayashi 2007). In yeast species from the Saccharomycetaceae family and the CTG clade, the 5S gene is

usually transcribed in the opposite orientation relative to the three other rDNA genes (Bergeron and Drouin 2008).

rDNA arrays are most of the time organized in tandem head-to-tail repeats. These tandem units can be located at a single place on one chromosome, or a species can have several rDNA arrays on different chromosomes (though in yeast species, the maximum number of arrays we observed in a species was three). The rDNA arrays can have an internal location on a chromosome, like the *S. cerevisiae* array on chromosome XII (Johnston et al. 1997), or they can be found in subtelomeric regions, as in *Y. lipolytica* and *C. glabrata* (Dujon et al. 2004). The array of complete tandem rDNA units is often flanked by an incomplete or rearranged partial unit (Johnston et al. 1997).

In this chapter I compare the locations of the rDNA arrays in yeast species from the Saccharomycetaceae family and from the CTG clade. I try to identify the possible evolutionary mechanisms that led to the various current locations of the rDNA arrays in these species. I discuss whether particular locations can be favoured by natural selection.

4.3 Methods

For most species whose genome sequences have been published, the locations of rRNA genes were taken directly from the existing genome sequence annotations. In some cases we used BLASTN searches with *S. cerevisiae* 18S, 5.8S, 25S and 5S genes to identify rRNA loci in another species.

For *S. bayanus*, whose genome sequence is fragmented into many contigs (Kellis et al 2003; Cliften et al. 2003), we inferred that the rDNA is probably at the same location as in *S. cerevisiae* because a contig (n° 528) matches chromosome XII of *S. cerevisiae* and terminates at the tQ-UUG gene which in *S. cerevisiae* is beside the rDNA. A recent Illumina resequencing project for three species from the genus *Saccharomyces*, including *S. bayanus*, also concluded that the rDNA loci in all three species is syntenic with the *S. cerevisiae* locus, even though a gap persists in the assembly of this region in the three species (Scannell et al. 2011). A similar situation

exists for *Lachancea (Kluyveromyces) waltii*, whose draft genome sequence consists of numerous contigs connected into scaffolds (Kellis et al. 2004) and is discussed below.

For the seven species newly sequenced in our laboratory, we found that in all cases the initial shotgun assembly of the genome (by the Celera Assembler) did not include a complete rDNA unit in the scaffolds or contigs. Instead, BLASTN searches showed that their rDNAs were dispersed onto several hundred short contigs (< 1 kb) that the Celera Assembler had not merged into larger contigs or integrated into scaffolds. We are unsure of the reason for this assembly failure, but it may be related to the large number of rDNA-derived reads from each species or the tandem repeating nature of the array. For each species, we used BLASTN with the *S. cerevisiae* rDNA unit as a query to identify all the short Celera contigs containing their rDNA, and then used these short contigs as input to the CAP3 assembler (Huang and Madan 1999) to obtain longer rDNA contigs. For most species the CAP3 output from this step consisted of two contigs: a large contig containing the 18S, 5.8S and 25S genes, and a small contig around the 5S gene, but not the complete ETS/IGS region joining them (**Figure 1.4**) because this region has low similarity between species. To obtain the sequence of a complete rDNA repeating unit from each species, we extended the initial 18S-5.8S-25S and 5S CAP3 contigs by 'walking' (successively identifying and incorporating unused Celera contigs that overlapped with their ends) until they merged into a single circular consensus unit.

We then incorporated these consensus rDNA units into the scaffold sequences for each species by identifying places in the genome where scaffold sequence overlapped with the rDNA consensus. In general, these overlaps involved matches between an end of a scaffold and an arbitrary point in the rDNA repeating unit. We extended the scaffold by extending the match to add an incomplete rDNA unit, followed by one complete unit to make annotation simpler. In cases where two scaffolds in a species ended in rDNAs with opposite orientations relative to the scaffold, we assumed that the two scaffolds were part of the same chromosome and merged them at the rDNA.

Synteny relationships were identified by using YGOB for the family Saccharomycetaceae (Byrne and Wolfe 2005; Kurtzman 2011) and CGOB (Fitzpatrick

et al. 2010) for the CTG clade species. We used tRNAscan-SE (Lowe and Eddy 1997) to search for the presence of tRNA genes in the vicinity of the rDNA array.

4.4 Results and Discussion

4.4.1 An ancestral rDNA location in Saccharomycetaceae

We found that nine of the 15 studied Saccharomycetaceae species – six post-WGD and three non-WGD – share a syntenic location for their rDNA arrays, which can therefore be inferred to be an ancestral rDNA location predating the WGD (**Figure 4.1**). Compared to the Ancestral yeast genome, the location of the ancestral rDNA array is between protein-coding genes Anc_8.371 and Anc_8.372. Remarkably, Anc_8.372 corresponds to the *S. cerevisiae* gene *HMO1* (*YDR174W*), which codes for an HMG-domain DNA-binding protein that is present in the chromatin of rDNA – specifically in the subset of rDNA units that are actively being transcribed (Merz et al. 2008). Hmo1 appears to be present in these rDNA units in place of histones. It is an rDNA-binding component of the yeast Pol I transcription system, the polymerase whose only function is to transcribe the rDNA (Gadal et al. 2002; Merz et al. 2008). HMO1 may play a role similar to that proposed for the mammalian counterpart UBF (Kasahara et al. 2007): both of them associates at many locations throughout the rRNA gene locus (Mais et al. 2005; Hall et al. 2006); both apparently recruits the whole Pol I machinery to their binding sites integrated at ectopic sites and contribute to the establishment of an "open" chromatin conformation at active rRNA genes (Mais et al. 2005; Hall et al. 2006). However, binding of HMO1 is Pol I dependent (contrary to UBF) (Kasahara et al. 2007). Lastly, HMO1 appears to be recruited to the rDNA array after removal of nucleosomes by elongating Pol I, and thus may play a role in the maintenance of but not the establishment of open chromatin in the rRNA gene cluster (Kasahara et al. 2007). The *HMO1* gene was ancestrally beside the rDNA array and is still beside it in several species (Neuveglise et al. 2000), but in *S. cerevisiae* after the WGD the rDNA ended up on chromosome XII while *HMO1* ended up on the sister region of chromosome IV (**Figure 4.2**).

A conserved tRNA gene (*tQ-UUG*) is present beside or close to the rDNA array in five of the species (two non-WGD and three post-WGD) whose rDNA is still at the

ancestral position. The ancestral location of *tQ-UUG* can be inferred to be between protein-coding genes Anc_8.368 and Anc_8.369, which is three genes away from the ancestral rDNA. In some post-WGD species (*S. cerevisiae* and *Kaz. africana*), *tQ-UUG* and the rDNA have become neighbors because the three intervening genes have been lost from the rDNA chromosome and instead are on the non-rDNA containing chromosome. In *S. cerevisiae* these intervening genes are now on chromosome IV beside *HMO1* (Anc_8.369 = *HSP42/YDR171W*; Anc_8.370 = *SUP35/YDR172W*; Anc_8.371 = *ARG82/YDR173C*).

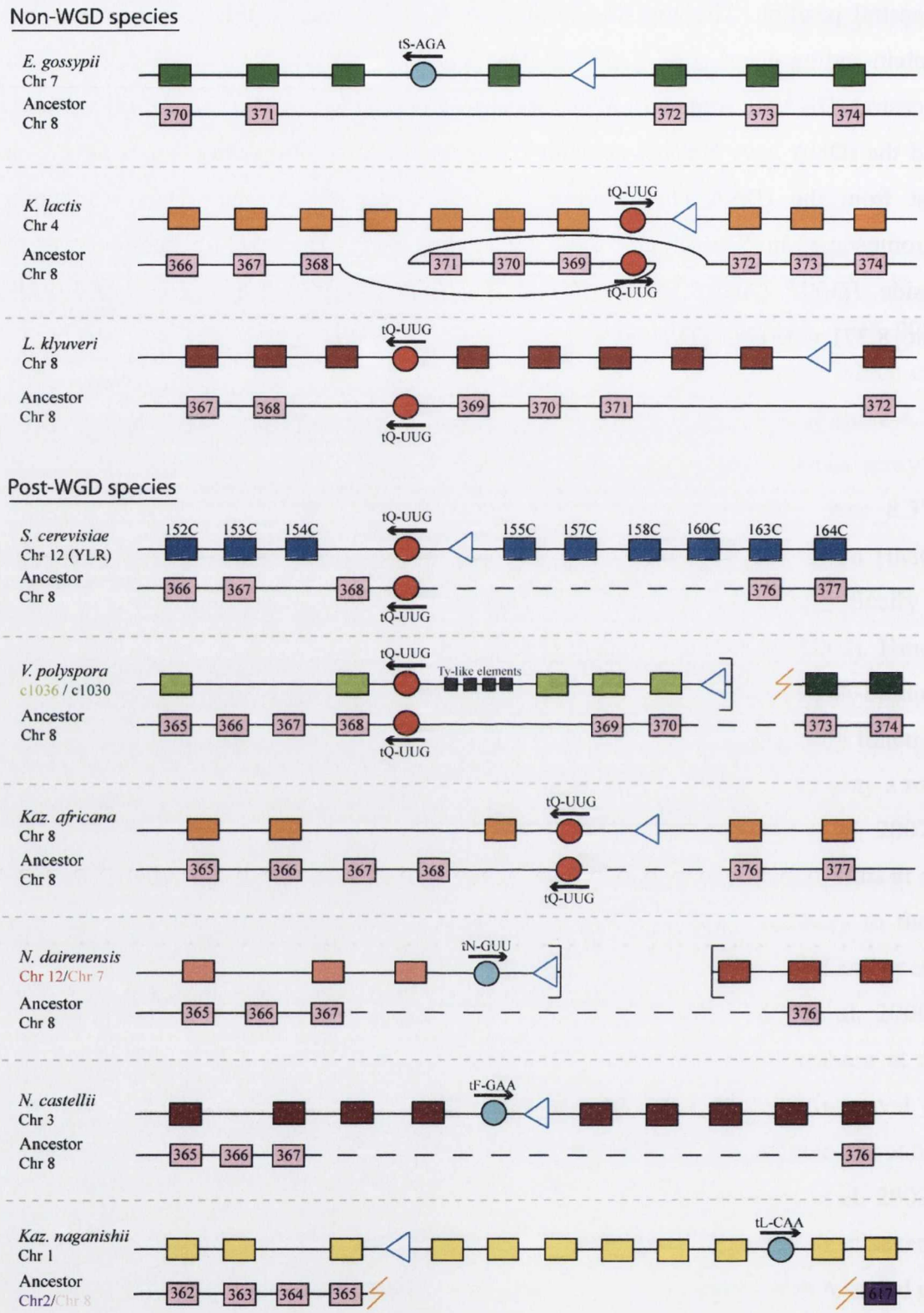


Figure 4.1. An ancestral location of the rDNA array. This location is shared by nine species. Protein-coding genes are represented by boxes; for each species, the corresponding ancestral genes on Anc_8 are shown and numbered. rDNA arrays are represented by white triangles. tRNA genes are represented by circles (red for the conserved *tQ-UUG* gene, blue for other tRNAs). Orange bolts represent a break in the synteny.

In the four other species whose rDNA is at the ancestral site, various different single tRNA genes are present in the vicinity of the rDNA array (**Figure 4.1**). For *N. castellii* and *N. dairenensis*, however, the *tQ-UUG* gene is conserved but located on the sister region paralogous to the rDNA-containing region (not shown in **Figure 4.1**). Among the species whose rDNA array is not located at the ancestral site, *tQ-UUG* is nevertheless conserved between Anc_8.368 and Anc_8.369 location in most of them (*Z. rouxii*, *T. delbrueckii*, *C. glabrata* and *T. blattae*), but in *L. thermotolerans* and *L. waltii* a different tRNA gene (*tT-AGU*) is present at this site (visible in **Figure 4.4**).

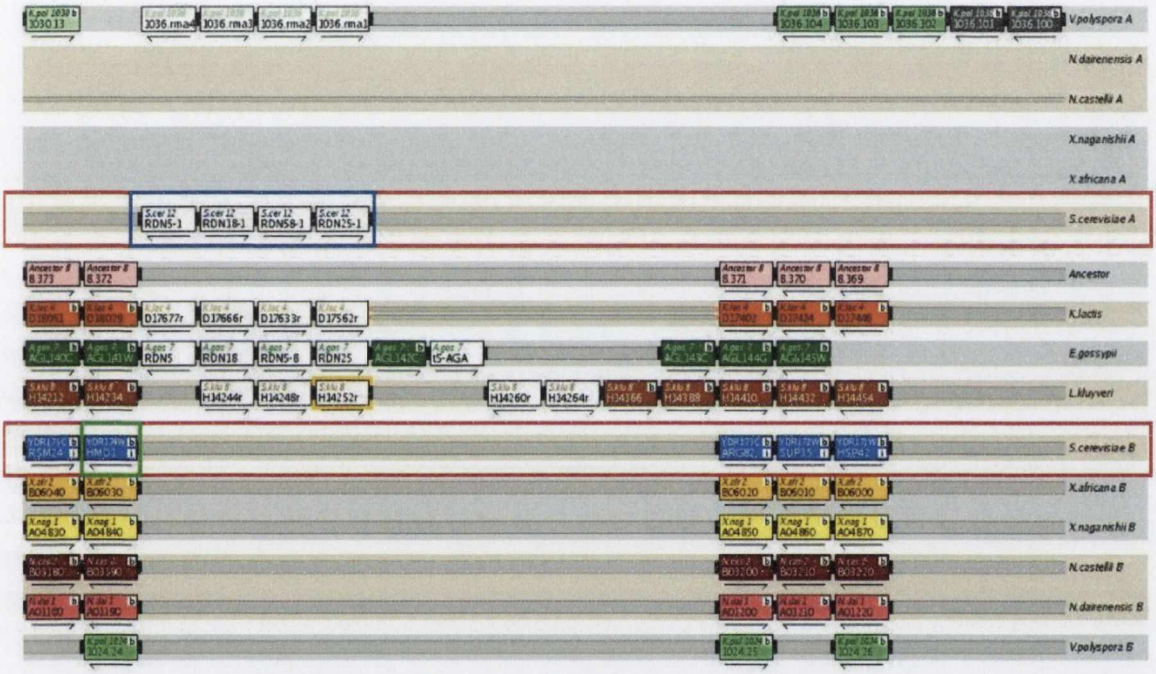


Figure 4.2. Screenshot from YGOB showing the location of the rDNA array and the protein-coding gene HMO1 in *S. cerevisiae*. The sister regions of *S. cerevisiae* chromosomes XII and IV are framed in red. In these regions, the rDNA array is framed in blue, and the HMO1 gene in green. These two elements are present at the ancestral location. In *E. gossypii*, *K. lactis* and *L. kluyveri* (all non-duplicated species, their rDNA array locations located on the figure between the two *cerevisiae* sister regions), the HMO1 gene is located just beside the rDNA array.

In seven of the nine species that retain rDNA at the ancestral location, the gene order around the ancestral location is conserved with no break in synteny. The exceptions are *K. lactis* and *Kaz. naganishii*. In *K. lactis* we observe a small inversion that only concerns three genes and the tRNA gene. However, in *Kaz. naganishii* we can see a break in the synteny: one side of the rDNA array (the left side in **Figure 4.1**) is the ancestral location, but on the other side there are several singleton genes (mostly for

hexose transporters), a tRNA gene *tL-CAA*, followed by a gene orthologous to Anc_2.617 (=*S. cerevisiae STV1*, a subunit of vacuolar ATPase). The simplest way to explain this break in the synteny is by two successive events, as shown on **Figure 4.3**: a reciprocal translocation between the pink and the purple chromosomes (Anc_2 and Anc_8), close to the rDNA array; and then a translocation event involving a telomere.

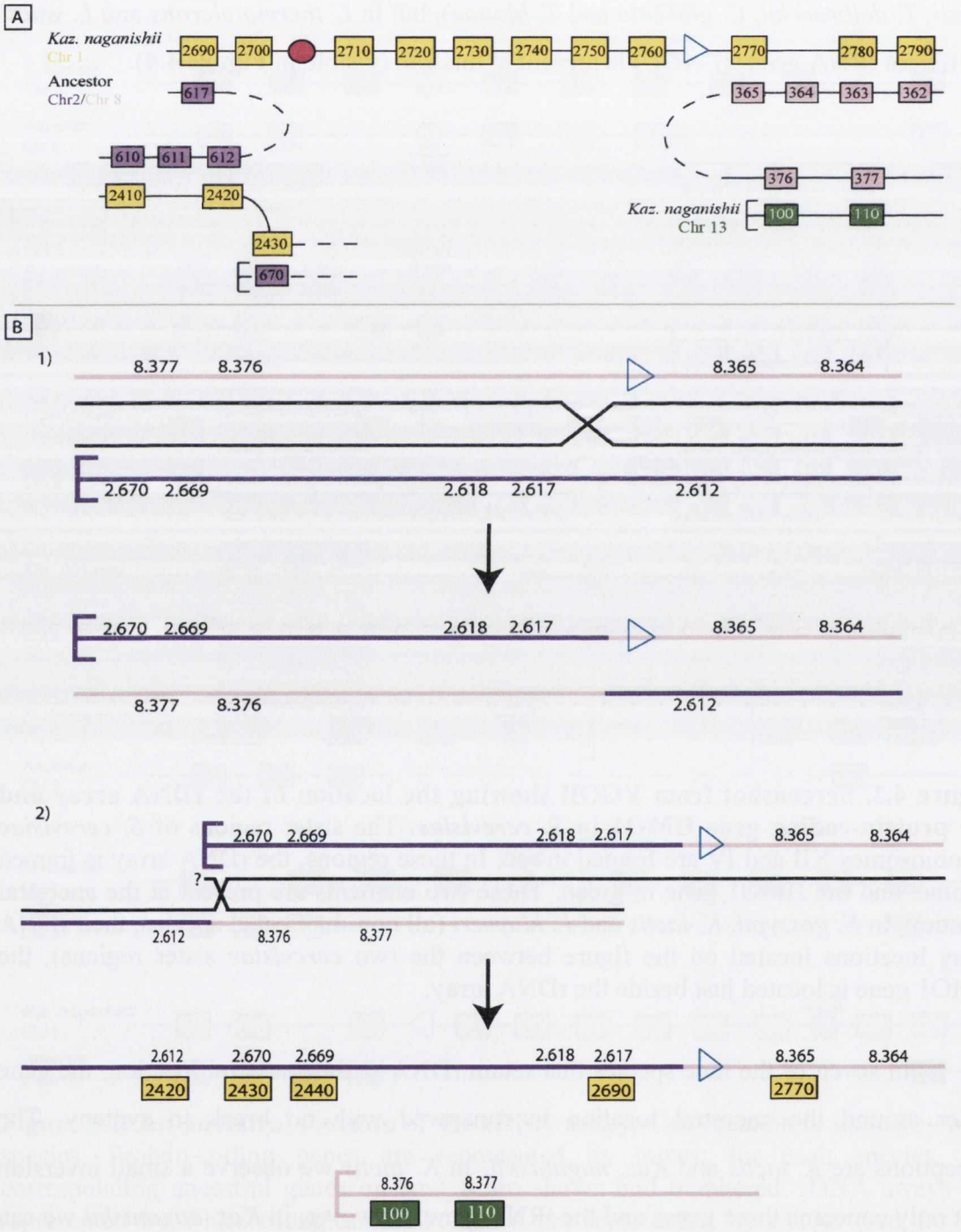


Figure 4.3: Inferred rearrangement events forming the current location of the rDNA array in *K. naganishii*. (A) The current organization of the array and its flanking genes. (B) Two translocations can explain the reorganization of synteny in this

region: 1) a translocation between Ancestral chromosome 8 (in pink, just beside the rDNA array (blue triangle)) and Ancestral chromosome 2 (in purple, somewhere between genes 2.617 and 2.612); 2) another translocation between the chromosomes involved in the previous translocation. In the last picture, only the *Kaz. naganishii* genes from (A) with an homologous gene in the ancestral genome are shown (in yellow and green).

What does the ancestral rDNA locus look like in species that no longer have an rDNA array at this position? These 'empty' ancestral rDNA loci are compared in **Figure 4.4**. In *T. delbrueckii* the rDNA is at a telomere and the intergenic sequence at the ancestral site between its Anc_8.371 (*ARG82*) and Anc_8.372 (*HMO1*) genes is only 170 bp with no sequence similarity to anything. This observation suggests that a mechanism must exist for clean deletion of large regions without altering the neighboring regions. In *T. delbrueckii*, the size of the telomeric rDNA cluster is estimated to be 980 kb (Maleszka and Clark-Walker 1993), so it probably had a similar size when present at the ancestral location. In *L. thermotolerans* the ancestral site, devoid of rDNA, contains two genes: one for a *DAL5*-like transmembrane transporter that is shared with *L. waltii* (**Figure 4.4**), and one that is similar to the *S. cerevisiae* zinc ion transporter gene *ZRT2*. An orthologous *ZRT2* gene is present close to the rDNA arrays at this ancestral site in *E. gossypii* and *L. kluyveri* (**Figure 4.1**). Thus the *ZRT2* gene became linked to the rDNA in this clade of non-WGD species, and then later the rDNA moved away in *L. thermotolerans* leaving *ZRT2* behind. In *Z. rouxii*, the rDNA has been lost from the interval between Anc_8.371 and Anc_8.372 and the loss coincides with a species-specific rearrangement between this point and two other chromosomal sites, one of which is telomeric (**Figure 4.4**).

Among the post-WGD species, too, there are 'empty' sites in the chromosomal regions that are paralogous to the regions that retained rDNA at the ancestral site. These sites can be seen in *S. cerevisiae* (chromosome IV region near *HMO1*), *C. glabrata* and *T. blattae* in the upper part of **Figure 4.4**, and possibly also *T. phaffii* (lower part of Figure 4.3; the YGOB tracking for this species is uncertain). Interestingly, the genome of the lager brewing yeast *S. pastorianus* (Nakao et al. 2009), which is an interspecies hybrid between *S. cerevisiae* and *S. eubayanus* (Libkind et al. 2011), may reveal an intermediate state in the process of rDNA array loss after polyploidization. This species has two nuclear sub-genomes (one from each parental species), with two types of

chromosome XII, one closely related to *S. cerevisiae* (called Sc-type), and one closely related to *S. eubayanus* (called Sb-type). The length of the rDNA region on the Sc-type chromosome XII was estimated to be >350 kb, whereas the one on the Sb-type chromosome XII was only 18 kb, indicating a massive collapse of the rDNA region on the chromosome originated from *S. eubayanus* (Nakao et al. 2009).

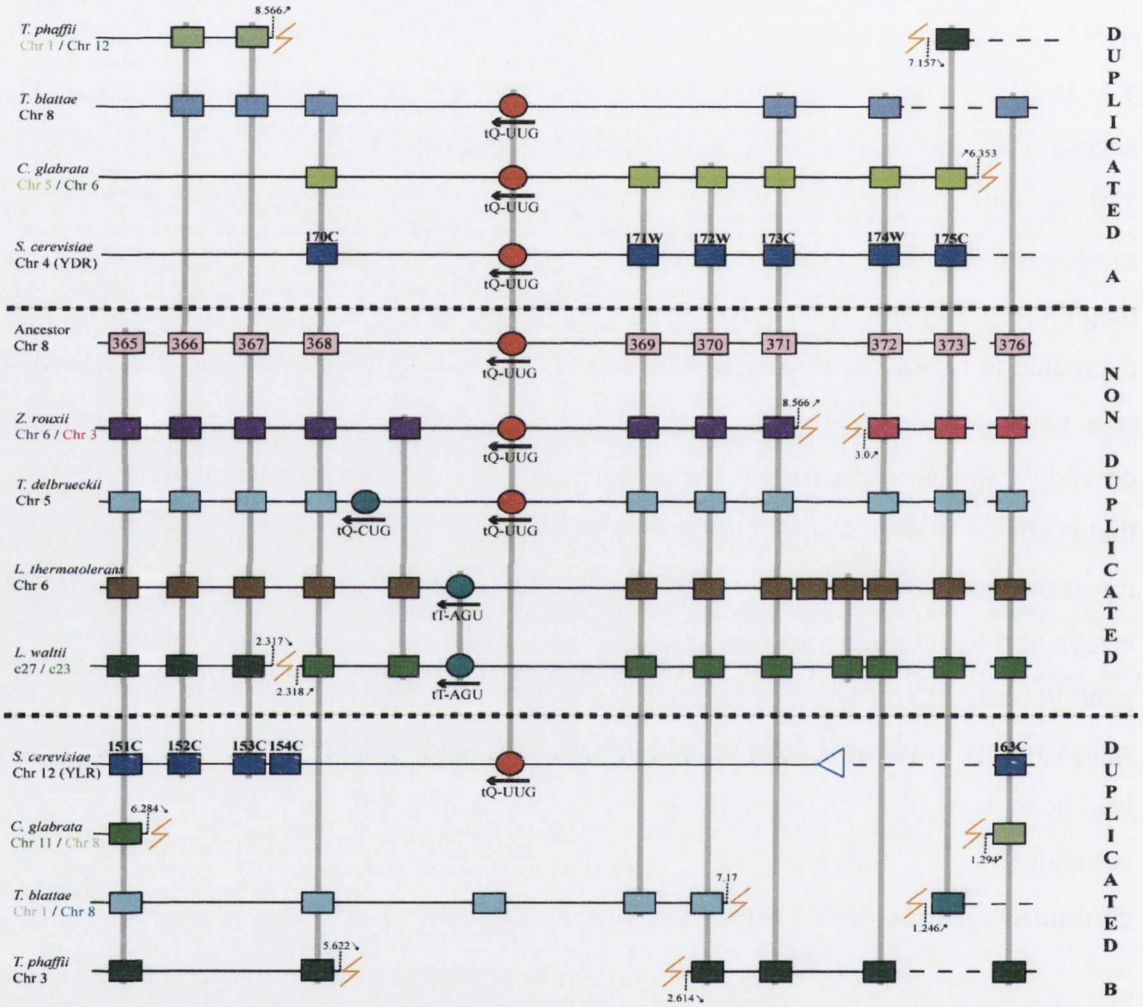


Figure 4.4. Organization of the 'empty' ancestral rDNA site in species where the rDNA genes are elsewhere in the genome. Genes are represented by boxes. tRNA elements are represented by circles (red for conserved tQ-UUG genes, blue for other tRNAs). Orange bolts represent a break in the synteny. The genome of *S. cerevisiae* is included to show the location of the ancestral rDNA (triangle). The dashed lines separate the homologous regions A and B of WGD species and the unique region of non-WGD species.

4.4.2 Species with subtelomeric rDNA arrays

Five species (*Tor. delbrueckii*, *Z. rouxii*, *T. blattae*, *T. phaffii* and *C. glabrata*) have their rDNA arrays located in subtelomeric regions. As described below, the subtelomeric locations in these different species correspond to different points in the Ancestral genome, so there is no evidence that any of them is an ancient conserved location for rDNA. However, this is a weak argument because it is quite likely that the telomeric rDNA arrays can easily be exchanged between chromosomes, muddling their linkage to Ancestral genes. In *C. glabrata*, the rDNA is at the end of chromosome 12 (the closest gene in the Ancestral genome is Anc_8.2), and in *T. delbrueckii* at the end of chromosome 7 (Anc_7.1). In *T. phaffii*, the array is at the end of the chromosome 9 (Anc_8.877), and a partial 5S gene is found at the end of the chromosome 1 (Anc_3.581). Similarly, *Z. rouxii* has a complete subtelomeric rDNA array on chromosome 5 (Anc_3.581) and a partial 25S gene at the telomere of chromosome 6 (Anc_6.1). In *T. blattae*, we assigned two rDNA arrays to subtelomeric regions of chromosomes 4 (Anc_2.83) and 5 (Anc_4.389). The decision to assign rDNA loci to both chromosomes was somewhat arbitrary because the two chromosomes share a large and almost identical DNA sequence duplication that extends approximately 20 kb further into the chromosomes (towards the centromere) beyond the rDNA genes. Our initial (Celera Assembler) scaffolds of the *T. blattae* genome ended in this repeat, so when we manually added the rDNA consensus unit sequence into the genome (as described in the Methods) we had a choice of two apparently identical sites and decided that the most probable real situation is that the subtelomeres of both chromosomes have an rDNA array.

In the five yeast species with subtelomeric rDNA arrays, only the rDNA genes of *Z. rouxii* (at their two different locations) and *T. delbrueckii* have tRNA genes in their vicinity. These tRNAs genes are not of the same type: *tK-UUU* (chromosome 4) and *tI-AAU* (chromosome 5) for *Z. rouxii*, and *tV-AAC* for *T. delbrueckii*.

Telomere length is quite dynamic, balanced between erosion (due to the inability of the DNA polymerase to fully replicate this region) and gain of sequences by telomerase-mediated extension (Eckert-Boulet and Lisby 2010). It has been shown that in the absence of telomerase, telomeres erode on average 3-5 bp per mitotic cell cycle in

S. cerevisiae (Eckert-Boulet and Lisby 2010) . The shortening of telomeres at successive cell divisions has a role in cellular senescence. rDNA is also a lifespan determinant, as reviewed it in **Section 1.4**. The different mechanisms and roles of both telomeres and rDNA in senescence have been discussed and compared (Kobayashi 2011).

To prevent telomere ends from being recognized by the DNA repair machinery and processed as they were double-strand breaks, telomeres are capped by specific proteins including Rap1, the Sir2-Sir3-Sir4 complex and the CST (Cdc13-Stn1-Ten1) complex (Wellinger 2010). This prevents end-to-end fusion of chromosomes. The capping proteins also regulate telomere length and telomerase activity (Eckert-Boulet and Lisby 2010). I previously discussed the fact that the homogenization of the rDNA arrays in yeasts could occur by homologous recombination between different tandem repeats. But some of the proteins involved in telomere capping inhibit the homologous recombination at telomeres unless telomeres are short (Negrini et al. 2007). However, this telomere capping only affects the GC-rich tandem repeats of a degenerate TG_{1-3} sequence at the end of the chromosome, the same region that is acted upon by telomerase which can synthesize de novo TG-rich repeat sequences. So sequence homogenization of the rDNA arrays by homologous recombination of tandem repeats is possible even if the arrays are located in a telomeric region.

4.4.3 Evolutionary movement of rDNA in L. thermotolerans

In *L. thermotolerans*, the rDNA array has been cleanly deleted from the ancestral location (**Figure 4.4**) and instead has been inserted at a different non-telomeric location (**Figure 4.5**). The site of the array insertion matches Ancestral chromosome 1, with gene Anc_1.349 on one side and gene Anc_1.372 on the other side. A tRNA gene (*tY-GUA*) is located beside the rDNA array in *L. thermotolerans* and is conserved in most other species.

The genome sequence of *L. thermotolerans* is complete (Souciet et al. 2009) whereas that of *L. waltii* is composed of many short contigs (Kellis et al. 2004a). The two species are closely related and the limited data available for *L. waltii* suggest that its rDNA array is probably located at the same place as in *L. thermotolerans*. In the left part of **Figure 4.5** we can see that *L. waltii* contig n°0 ends at the *tY-GUA* gene. In the

right part, beside *L. waltii* gene 16287 there are five Ty element genes (not shown in **Figure 4.5**) after which is the end of the contig n°34. It is probable that the rDNA array lies between these two contigs in *L. waltii*. In contrast, the third species in the genus *Lachancea*, *L. kluyveri*, has only the *tY-GUA* gene at this location (**Figure 4.5**).

The fact that the genes flanking the *L. thermotolerans* rDNA – Anc_1.349 and Anc_1.372 – are not consecutive in the Ancestral genome indicates that a rearrangement has occurred. However, we can determine that this rearrangement (an inversion) was not caused by the insertion of the rDNA genes but occurred before the insertion, because *E. gossypii* and *L. kluyveri* also show the same rearrangement even though their rDNA arrays are located at the ancestral location. In *K. lactis* (the other non-WGD species in this clade), gene order in the region has been affected by an interchromosomal translocation (involving *K. lactis* chromosomes 1 and 2), that is probably a species-specific event that occurred after a shared inversion in the common ancestor of these five non-WGD species.

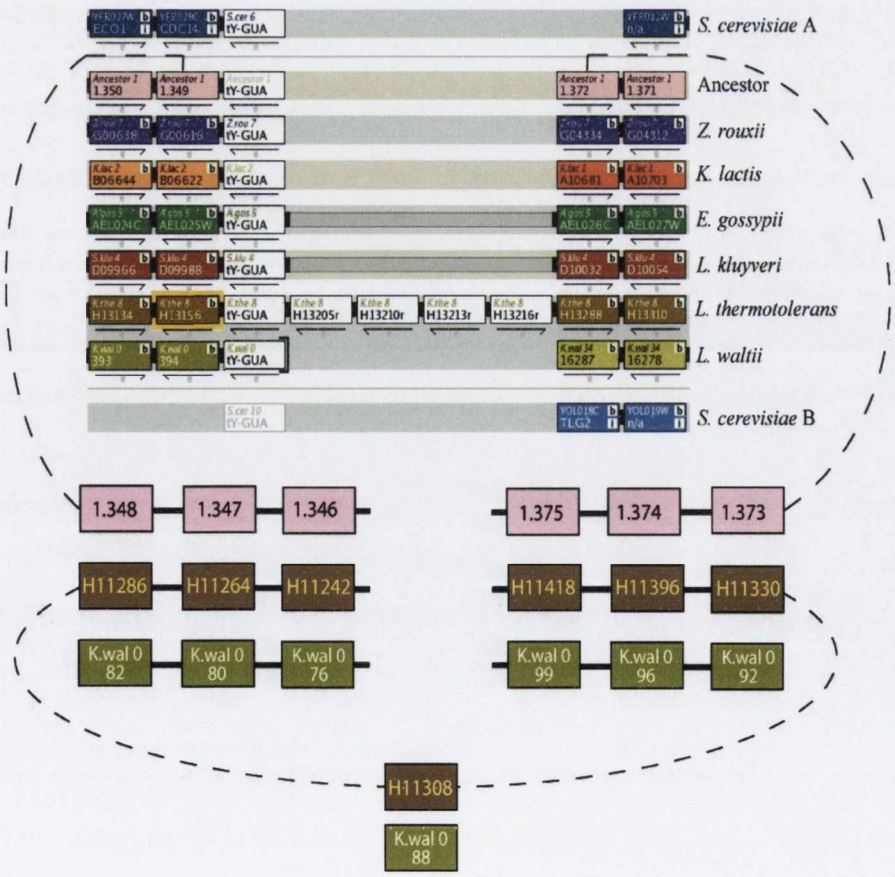


Figure 4.5. The rDNA array location in *L. thermotolerans*. This figure shows the location of the rDNA genes (H13205r/25S, H13210r/5.8S, H13213r/18S and

H13216r/5S) in *L. thermotolerans* and probably in *L. waltii*, and the inversion prior the insertion of the array, which also concerns *L. kluyveri* and *E. gossypii*.

Interestingly, the new site to which rDNA moved (probably in the common ancestor of *L. thermotolerans* and *L. waltii*) is again located beside a gene with an important function in rDNA function. This gene is *CDC14* (Anc_1.349) which in *S. cerevisiae* codes for a protein phosphatase that is located in the nucleolus and is required for exit from mitosis. The rDNA array is the last part of the genome at which the replicated chromosomes separate. In metaphase, cohesin destruction is not sufficient for segregation of the long arm of chromosome XII that contains the rDNA (Sullivan et al. 2004). Separase, at the same time as cleaving cohesin, activates the *CDC14* phosphatase (Sullivan and Uhlmann 2003). *CDC14* and the condensin complex will then condensate and shorten this particular region of chromosome XII sufficiently for segregation (D'Amours et al. 2004; Sullivan et al. 2004). Furthermore, it was shown that *CDC14* inhibits transcription of rDNA by Pol I during anaphase (Clemente-Blanco et al. 2009). Transcription inhibition is necessary for complete chromosome disjunction, because rRNA transcripts block condensin binding to rDNA, which is essential for the compaction of the locus.

To summarize the situation for the family Saccharomycetaceae, we have identified a genomic site (between Anc_8.371 and Anc_8.372) that can be inferred by parsimony to have been the location of the rDNA in the common ancestor of the family and at the moment of the WGD. This ancestral location is retained in at least nine species. However, losses of the entire rDNA array from this site have occurred several times independently. First, in the common ancestor of *L. thermotolerans* and *L. waltii*, the array moved to a site corresponding to Ancestral chromosome 1. Second, after WGD, one complete copy of the array was lost in the common ancestor of all post-WGD species, from a site that maps to chromosome IV of *S. cerevisiae*. Third, the array appears to have moved from its ancestral location to subtelomeric locations on several occasions during the evolution of family Saccharomycetaceae. Although it is not possible to use synteny data to infer whether subtelomeric rDNA arrays in different species are orthologous, the phylogenetic distribution of species with subtelomeric arrays indicates that at least four independent movements of the rDNA from the ancestral site to a subtelomeric site must have occurred (**Table 4.1**): once in the

common ancestor of *Tor. delbrueckii* and *Z. rouxii*; once in the *C. glabrata* lineage; and once each on the lineages leading to *T. phaffii* and *T. blattae* (see tree in **Figure 3** of **Appendix I**). Whether the relocation event seen in *L. thermotolerans* was actually the result of a two-step process, with relocation from the ancestral location to a telomere being later followed by relocation from a telomere to an internal chromosomal site, is open to debate.

Table 4.1. Summary of rDNA locations in species of the family Saccharomycetaceae.

Species	Location of the rDNA array	Closest ancestral gene	Nearby tRNA gene (if exists)	At the ancestral site
Saccharomyces cerevisiae	Ancestral	Anc_8.368	tQ-UUG	rDNA array
Saccharomyces bayanus	Ancestral	Anc_8.368	tQ-UUG	rDNA array
Candida glabrata	Telomeric	Anc_8.2	-	Rearranged
Kazachstania naganishii	Ancestral	Anc_8.368	tL-CAA	rDNA array
Kazachstania africana	Ancestral	Anc_8.368	tQ-UUG	rDNA array
Naumovia castellii	Ancestral	Anc_8.368	tF-GAA	rDNA array
Naumovia dairenensis	Ancestral	Anc_8.368	tN-GUU	rDNA array
Tetrapisispora blattae	Telomeric	(5)Anc_4.389/(4)2.83	-	Empty
Tetrapisispora phaffii	Telomeric	Anc_8.877	-	Rearranged
Vanderwaltozyma polyspora	Ancestral	Anc_8.368	tQ-UUG	rDNA array
Zygosaccharomyces rouxii	Telomeric	Anc_3.581	tK-UUU	Empty
Torulaspora delbrueckii	Telomeric	Anc_7.1	tV-AAC	Empty
Lachancea thermotolerans	Internal	Anc_1.349/1.372	tY-GUA	Empty
Lachancea waltii	Internal?	Anc_1.349/1.372	tY-GUA	Empty
Lachancea kluyveri	Ancestral	Anc_8.368	tQ-UUG	rDNA array
Kluyveromyces lactis	Ancestral	Anc_8.368	tQ-UUG	rDNA array
Eremothecium gossypii	Ancestral	Anc_8.368	tS-AGA	rDNA array

The different colors define the twelve clades. Numbers in parentheses represent chromosome numbers in cases where the full rDNA array is found in several locations in a genome.

4.4.4 rDNA locations in the CTG clade

We saw that the yeast species in YGOB retain extensive synteny with each other, even across relatively large evolutionary distances. Using CGOB, we can also see that the species from the CTG clade retain considerable synteny. However, there is little conservation of gene order between the species in YGOB and those in CGOB.

Among the species in the CTG clade, there is again some conservation and some variation in rDNA array location. For example, *C. albicans*, *C. dubliniensis* and *C. tropicalis* share the same location for their rDNA arrays (**Figure 4.6A**); so also do *C.*

parapsilosis and *L. elongisporus* (**Figure 4.6B**). We can see that the genes near the rDNA arrays in the latter two species are orthologs of genes located on the chromosome R (the rDNA-containing chromosome) in *C. albicans* and *C. dublinensis*; however, these genes are not close to the rDNA. The *C. dubliniensis* gene names in Figure 4.5 give an indication of the scale because in this species the genes are named sequentially along the chromosome, increasing in steps of 10. In *C. dubliniensis* itself the rDNA array is located between genes 34350 and 34360. However, in *C. parapsilosis* and *L. elongisporus*, the closest genes to the rDNA array are orthologs of *C. dubliniensis* genes 34450 and 33560 (**Figure 4.6B**), genes that are on chromosome R but 11 and 80 genes away from the rDNA, respectively. We cannot say that all five species share the same rDNA location. As with the Saccharomycetaceae, a tRNA gene (*tG-GCC*) is found close to the rDNA array in three of the five species.

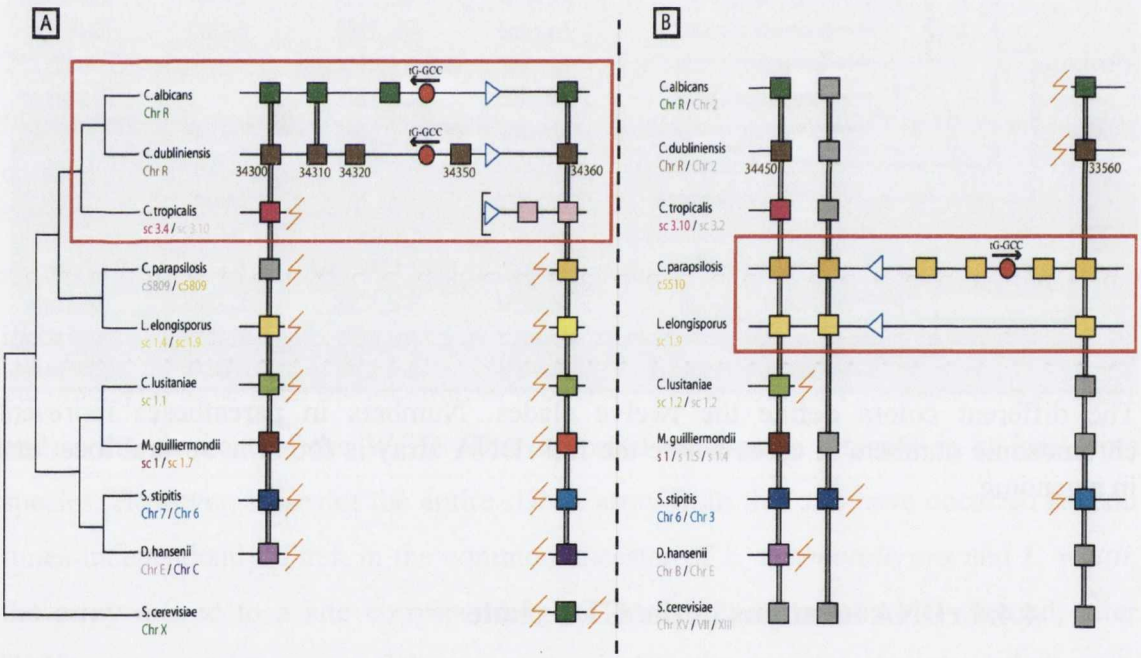


Figure 4.6. The location of the rDNA genes in several species from the CTG clade. Boxes represent genes, circles represent tRNA genes, and triangles represent rDNA arrays. Orange bolts represent a break in the synteny. Genes drawn in vertical pillars are orthologs.

The four other species in CGOB all seem to have different locations for their rDNA arrays. In *C. lusitaniae*, there are three rDNA arrays, all of which are subtelomeric (at one end of supercontig c8, and both ends of supercontig c3). There are

no tRNA genes in the vicinity of any of the three rDNA arrays. The closest genes to the arrays on supercontig c3 are orthologs of *C. dubliniensis* genes 30010 and 30250, located on the chromosome R but far from the rDNA. The closest gene to the rDNA array on supercontig c8 is orthologous to a *C. dubliniensis* gene on chromosome 5. Moreover, a solo 5S gene is found in a non-telomeric position on c3. In *S. stipitis*, the rDNA array is located at an internal site on chromosome 3. Synteny of the genes located upstream and downstream of the array is conserved in three other species (*D. hansenii*, *C. lusitaniae* and *M. guillermondii*) as shown in **Figure 4.7A**. There is a tRNA gene (*tA-TGC*) close to the rDNA array. The closest gene to the array is ortholog of the *C. dubliniensis* gene 33370, again located on chromosome R but not close to the rDNA array. In *M. guillermondii*, the rDNA array is located in an area full of singletons and genes that have orthologs only in *D. hansenii* (**Figure 4.7B**). Finally, we do not have information about the location of the rDNA genes in *D. hansenii*: when we do a BLASTN search of the four rDNA genes against its genome sequence, we obtain no hits.

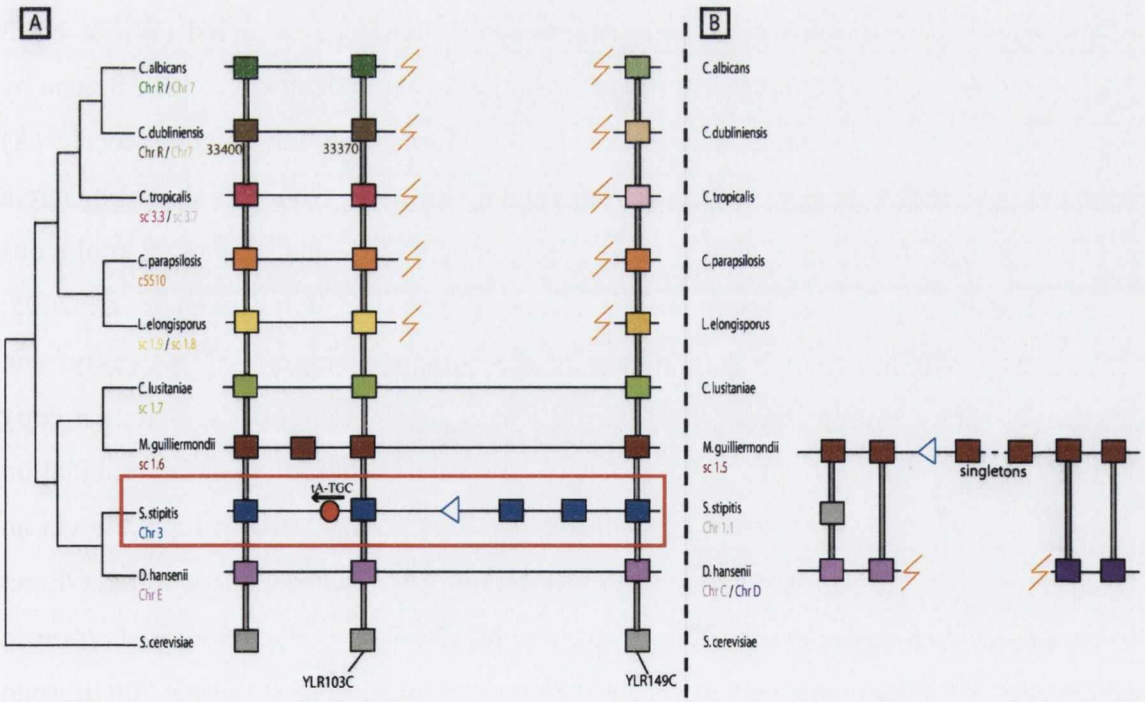


Figure 4.7. The location of the rDNA genes in *S. stipitis* (A) and *M. guillermondi* (B). Boxes represent genes, circles represent tRNA elements, and triangles represent rDNA arrays. Orange bolts represent a break in the synteny. Genes drawn in vertical pillars are orthologs.

To conclude, the process of rDNA location evolution in the CTG clade appears to be similar to that in the Saccharomycetaceae, with evidence for both internal and subtelomeric locations of the array and an apparent recent relocation of the array to an internal site in *S. stipis*. The functions of the genes flanking the *S. stipitis* insertion site do not have any obvious roles in rDNA maintenance in the same way that *HMO1* and *CDC14* do; they are *CDC45/YLR103W* coding for a DNA replication initiation factor, and *YLR149C* (unknown function).

There are no known mechanisms to explain how the whole rDNA array can become relocated in a genome. However, possible mechanisms have been proposed to explain the removal of the 5S rRNA gene from the array and its linkage to multigene families in different species (Drouin and de Sa 1995). It has been suggested that retrotransposons may have mediated the relocation of single 5S rDNA gene copies and their linkage to the rDNA array in angiosperms (Garcia et al. 2009) or are responsible for 5S gene duplications (Drouin 2000). However, concerted translocation of several copies at a time by transposable elements for any gene has not been shown yet, so it is unlikely that hundreds of 5S copies were relocated this way, although the combination of the transposable activity with other mechanisms cannot be excluded (Wicke et al. 2011). Another possible mechanism is the mediation of the movement of the 5S gene by the insertion of extrachromosomal covalently closed circular DNA (ecccDNA) containing 5S rRNA gene sequences (Drouin and de Sa 1995; Cohen et al. 2008; Garcia et al. 2009). These circular elements are thought to result from homologous recombination between adjacent repeats (Degroote et al. 1990; Cohen et al. 2005; Cohen et al. 2008), a process responsible for the homogenization of rDNA arrays and other multigene families. Interestingly, in old *S. cerevisiae* mother cells, a circular copy of the rDNA array pops out of the genomic array due to homologous recombination (Johnson et al. 1999). These extrachromosomal rDNA circles (ERCs) replicate via an origin present in the rDNA sequence and proliferate. ERCs were thought to be a direct cause of aging (Johnson et al. 1999). Ganley et al. (2009) established a yeast strain in which ERCs did not duplicate and accumulate in the mother cell. However, the lifespan of this strain was shortened (Ganley et al. 2009). The direct cause of lifespan shortening was rDNA instability, which is known to increase the number of ERCs (Kaeberlein et al. 1999).

rDNA arrays are often located in pericentromeric or subtelomeric regions (Eichler and Sankoff 2003; Averbeck and Eickbush 2005). A study on the dynamics of rDNA clusters within the genus *Mus* has reached several conclusions (Cazaux et al. 2011): rDNA arrays are almost exclusively located in pericentromeric regions; there is a strong association between rDNA sites and centromeres, maybe due to their shared constraint of concerted evolution; rDNA clusters coincide with breakpoints in the genus *Mus*; and a substantial rate of rDNA cluster change (insertion, deletion) also occurred in the absence of chromosomal rearrangements. The 'point' structure of Saccharomycetaceae centromeres means that their mode of evolution is likely to be very different from mammalian centromeres, but the evolutionary processes at telomeres may be more similar. Telomeres and subtelomeres are regions of great plasticity, with dynamics that allow for the amplification or variation in the number of repeated sequences and telomeric genes (Torres-Machorro et al. 2010). Subtelomeres are hotspots of interchromosomal recombination and segmental duplication (Linardopoulou et al. 2005; Brown et al. 2010). However, it is not known whether subtelomeric rDNA plays a role in the maintenance of the characteristic telomeric structure or whether it uses this particular structure to regulate its expression and to maintain the sequence and copy number (Torres-Machorro et al. 2010).

Chapter 5

Conclusions

The need for automated genome annotation methods has been important since the development of next-generation sequencing technologies. The Yeast Genome Annotation Pipeline (YGAP) presented in **Chapter 2** has been specifically designed for the *de novo* annotation of yeast genomes using the outputs from high-throughput sequencing methods. YGAP has proven very effective in the annotation of six new genomes from Saccharomycetaceae species, and the reannotation of *N. castellii*. The analysis of singleton genes in these seven automatically annotated genomes showed species-specific features (**Chapter 3**). For example, *Torulaspora delbrueckii* contains many genes gained by horizontal gene transfer compared to the other studied species; and we have observed large duplicated gene families in *Tetrapisispora blattae*. These gained genes must provide some benefits to the species to be kept, and it will be of interest to study their alleged functions in the lab. Orphan genes (genes that lack homologs in other lineages) can represent almost half of all the singletons in some species. Some of them belong to orphan gene families that can have up to 20 members, even if smaller families (two to five members) are more often found. The addition of the six newly sequenced genomes in the YGOB database improved our knowledge on the Saccharomycetaceae family. Our laboratory has used these newly sequenced genomes in a study on yeast sex chromosome evolution (**Appendix I**). They were also used in a study on how the location of the rDNA array has changed during yeast evolution (**Chapter 4**). We showed the existence of an ancestral location at an internal site in Saccharomycetaceae species, and that the rDNA array was relocated to telomeric or internal locations in some species. Based on our observations, we suggested that the location of the rDNA array is determined by natural selection. The mechanism responsible for the relocation of the array is still unknown.

References

Ahearn, D. G. (1998). Yeasts pathogenic for humans. The Yeasts, A Taxonomy. C. P. Kurtzman and J. W. Fell. Amsterdam, Elsevier: 9-12.

Altschul, S. F., et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Andalis, A. A., et al. (2004). "Defects arising from whole-genome duplications in *Saccharomyces cerevisiae*." Genetics **167**(3): 1109-21.

Ann Chai, L. Y., et al. (2010). "*Candida tropicalis* in human disease." Crit Rev Microbiol **36**(4): 282-98.

Aucott, J. N., et al. (1990). "Invasive infection with *Saccharomyces cerevisiae*: report of three cases and review." Rev Infect Dis **12**(3): 406-11.

Averbeck, K. T. and T. H. Eickbush (2005). "Monitoring the mode and tempo of concerted evolution in the *Drosophila melanogaster* rDNA locus." Genetics **171**(4): 1837-46.

Baldauf, S. L., et al. (2000). "A kingdom-level phylogeny of eukaryotes based on combined protein data." Science **290**(5493): 972-7.

Barabas, O., et al. (2008). "Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection." Cell **132**(2): 208-20.

Barnett, J. A. (2008). "A history of research on yeasts 12: medical yeasts part 1, *Candida albicans*." Yeast **25**(6): 385-417.

Barth, G. and C. Gaillardin (1997). "Physiology and genetics of the dimorphic fungus *Yarrowia lipolytica*." FEMS Microbiol Rev **19**(4): 219-37.

Bayha, K. M., et al. (2010). "Evolutionary relationships among scyphozoan jellyfish families based on complete taxon sampling and phylogenetic analyses of 18S and 28S ribosomal DNA." Integr Comp Biol **50**(3): 436-55.

Bely, M., et al. (2008). "Impact of mixed *Torulaspora delbrueckii-Saccharomyces cerevisiae* culture on high-sugar fermentation." Int J Food Microbiol **122**(3): 312-20.

Bennett, R. J. (2009). "A *Candida*-based view of fungal sex and pathogenesis." Genome Biol **10**(7): 230.

Bergeron, J. and G. Drouin (2008). "The evolution of 5S ribosomal RNA genes linked to the rDNA units of fungal species." Curr Genet **54**(3): 123-31.

Bhat, P. J. and T. V. Murthy (2001). "Transcriptional control of the GAL/MEL regulon of yeast *Saccharomyces cerevisiae*: mechanism of galactose-mediated signal transduction." Mol Microbiol **40**(5): 1059-66.

Birney, E., et al. (2004). "GeneWise and Genomewise." Genome Res **14**(5): 988-95.

Blanck, A., et al. (1989). "Primary structure of sensory rhodopsin I, a prokaryotic photoreceptor." Embo J **8**(13): 3963-71.

Bleykasten-Grosshans, C. and C. Neuveglise (2011). "Transposable elements in yeasts." C R Biol **334**(8-9): 679-86.

Bon, E., et al. (2003). "Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns." Nucleic Acids Res **31**(4): 1121-35.

Borneman, A. R., et al. (2011). "Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*." PLoS Genet **7**(2): e1001287.

Boutet, E., et al. (2007). "UniProtKB/Swiss-Prot." Methods Mol Biol **406**: 89-112.
Bowman, S. M. and S. J. Free (2006). "The structure and synthesis of the fungal cell wall." Bioessays **28**(8): 799-808.
Boyer, F., et al. (2005). "Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data." Bioinformatics **21**(23): 4209-15.
Brosch, R., et al. (2000). "Comparative genomics of the leprosy and tubercle bacilli." Res Microbiol **151**(2): 135-42.
Brosius, J. (1991). "Retroposons--seeds of evolution." Science **251**(4995): 753.
Brown, C. A., et al. (2010). "Rapid expansion and functional divergence of subtelomeric gene families in yeasts." Curr Biol **20**(10): 895-903.
Brown, D. D. and K. Sugimoto (1974). "The structure and evolution of ribosomal and 5S DNAs in Xenopus laevis and Xenopus mulleri." Cold Spring Harb Symp Quant Biol **38**: 501-5.
Brown, D. D., et al. (1972). "A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes." J Mol Biol **63**(1): 57-73.
Bruner, S. D., et al. (2000). "Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA." Nature **403**(6772): 859-66.
Bruns, T. (2006). "Evolutionary biology: a kingdom revised." Nature **443**(7113): 758-61.
Bryson, K., et al. (2006). "AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system." Nucleic Acids Res **34**(12): 3533-45.
Butler, G., et al. (2004). "Evolution of the MAT locus and its Ho endonuclease in yeast species." Proc Natl Acad Sci U S A **101**(6): 1632-7.
Butler, G., et al. (2009). "Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes." Nature **459**(7247): 657-62.
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**(10): 1456-61.
Caccio, S. M., et al. (2010). "Identification of *Giardia* species and *Giardia duodenalis* assemblages by sequence analysis of the 5.8S rDNA gene and internal transcribed spacers." Parasitology **137**(6): 919-25.
Cai, J., et al. (2008). "De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*." Genetics **179**(1): 487-96.
Carranza, S., et al. (1996). "Evidence that two types of 18S rDNA coexist in the genome of *Dugesia (Schmidtea) mediterranea* (Platyhelminthes, Turbellaria, Tricladida)." Mol Biol Evol **13**(6): 824-32.
Carver, T., et al. (2008). "Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database." Bioinformatics **24**(23): 2672-6.
Cavalier-Smith, T. and E. E. Chao (1995). "The Opalozoan *Apusomonas* is Related to the Common Ancestor of Animals, Fungi, and Choanoflagellates." Proceedings: Biological Sciences **261**(1360): 1-6.
Cazaux, B., et al. (2011). "Are ribosomal DNA clusters rearrangement hotspots?: a case study in the genus *Mus* (Rodentia, Muridae)." BMC Evol Biol **11**: 124.
Chappell, A. S. and V. Lundblad (2004). "Structural elements required for association of the *Saccharomyces cerevisiae* telomerase RNA with the Est2 reverse transcriptase." Mol Cell Biol **24**(17): 7720-36.

Christie, K. R., et al. (2009). "Functional annotations for the *Saccharomyces cerevisiae* genome: the knowns and the known unknowns." Trends Microbiol **17**(7): 286-94.

Ciani, M., et al. (2006). "Fermentation behaviour and metabolic interactions of multistarter wine yeast fermentations." Int J Food Microbiol **108**(2): 239-45.

Clemente-Blanco, A., et al. (2009). "Cdc14 inhibits transcription by RNA polymerase I during anaphase." Nature **458**(7235): 219-22.

Cliften, P., et al. (2003). "Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting." Science **301**(5629): 71-6.

Cliften, P. F., et al. (2001). "Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis." Genome Res **11**(7): 1175-86.

Cohen, S., et al. (2005). "Evidence for rolling circle replication of tandem genes in *Drosophila*." Nucleic Acids Res **33**(14): 4519-26.

Cohen, S., et al. (2008). "Extrachromosomal circular DNA derived from tandemly repeated genomic sequences in plants." Plant J **53**(6): 1027-34.

Cole, S. T. (1998). "Comparative mycobacterial genomics." Curr Opin Microbiol **1**(5): 567-71.

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.

D'Amours, D., et al. (2004). "Cdc14 and condensin control the dissolution of cohesin-independent chromosome linkages at repeated DNA." Cell **117**(4): 455-69.

Dammann, R., et al. (1993). "Chromatin structures and transcription of rDNA in yeast *Saccharomyces cerevisiae*." Nucleic Acids Res **21**(10): 2331-8.

Degroote, F., et al. (1990). "Nucleotide sequence of circular DNA molecules homologous to the 240 bp tandem repeats of the intergenic spacer of *Drosophila melanogaster* ribosomal DNA." Biol Cell **69**(1): 69-70.

Demain, A. L., et al. (1998). The industrial and agricultural significance of yeasts. The Yeasts, A Taxonomy. C. P. Kurtzman and J. W. Fell. Amsterdam, Elsevier**:** 13-19.

Dietrich, F. S., et al. (2004). "The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome." Science **304**(5668): 304-7.

Domergue, R., et al. (2005). "Nicotinic acid limitation regulates silencing of *Candida* adhesins during UTI." Science **308**(5723): 866-70.

Dover, G. (1982). "Molecular drive: a cohesive mode of species evolution." Nature **299**(5879): 111-7.

Drouin, G. (2000). "Expressed retrotransposed 5S rRNA genes in the mouse and rat genomes." Genome **43**(1): 213-5.

Drouin, G. and M. M. de Sa (1995). "The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families." Mol Biol Evol **12**(3): 481-93.

Dujon, B. (1996). "The yeast genome project: what did we learn?" Trends Genet **12**(7): 263-70.

Dujon, B. (2006). "Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution." Trends Genet **22**(7): 375-87.

Dujon, B. (2010). "Yeast evolutionary genomics." Nat Rev Genet **11**(7): 512-24.

Dujon, B., et al. (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.

Durbin, R., et al. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press.

Durrens, P., et al. (2008). "Fusion and fission of genes define a metric between fungal genomes." PLoS Comput Biol **4**(10): e1000200.

Eckert-Boulet, N. and M. Lisby (2010). "Regulation of homologous recombination at telomeres in budding yeast." FEBS Lett **584**(17): 3696-702.
Eichler, E. E. and D. Sankoff (2003). "Structural dynamics of eukaryotic chromosome evolution." Science **301**(5634): 793-7.
Ekman, D. and A. Elofsson (2010). "Identifying and quantifying orphan protein sequences in fungi." J Mol Biol **396**(2): 396-405.
Fink, G. R. (1987). "Pseudogenes in yeast?" Cell **49**(1): 5-6.
Finn, R. D., et al. (2010). "The Pfam protein families database." Nucleic Acids Res **38**(Database issue): D211-22.
Fischer, G., et al. (2000). "Chromosomal evolution in *Saccharomyces*." Nature **405**(6785): 451-4.
Fischer, G., et al. (2001). "Evolution of gene order in the genomes of two related yeast species." Genome Res **11**(12): 2009-19.
Fischer, G., et al. (2006). " Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages." PLoS Genet **2**(3):e32.
Fitzpatrick, D. A., et al. (2010). "Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser." BMC Genomics **11**: 290.
Fleischmann, R. D., et al. (1995). "Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*." Science **269**(5223): 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**(17): 5342-6.
Forche, A., et al. (2008). "The parasexual cycle in *Candida albicans* provides an alternative pathway to meiosis for the formation of recombinant strains." PLoS Biol **6**(5): e110.
Francino, M. P. (2005). "An adaptive radiation model for the origin of new gene functions." Nat Genet **37**(6): 573-7.
Fraser, C. M., et al. (1995). "The minimal gene complement of *Mycoplasma genitalium*." Science **270**(5235): 397-403.
Freer, S. N. and R. W. Detroy (1983). "Characterization of cellobiose fermentations to ethanol by yeasts." Biotechnol Bioeng **25**(2): 541-557.
Gadal, O., et al. (2002). "Hmo1, an HMG-box protein, belongs to the yeast ribosomal DNA transcription system." Embo J **21**(20): 5498-507.
Galagan, J. E., et al. (2003). "The genome sequence of the filamentous fungus *Neurospora crassa*." Nature **422**(6934): 859-68.
Ganley, A. R., et al. (2009). "The effect of replication initiation on gene amplification in the rDNA and its relationship to aging." Mol Cell **35**(5): 683-93.
Ganley, A. R. and T. Kobayashi (2007). "Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data." Genome Res **17**(2): 184-91.
Garcia, S., et al. (2009). "Linkage of 35S and 5S rRNA genes in *Artemisia* (family Asteraceae): first evidence from angiosperms." Chromosoma **118**(1): 85-97.
Gelade, R., et al. (2003). "Multi-level response of the yeast genome to glucose." Genome Biol **4**(11): 233.
Goffeau, A., et al. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-7.
Gojkovic, Z., et al. (2004). "Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts." Mol Genet Genomics **271**(4): 387-93.
Gordon, J. L., et al. (2011). "Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents." PNAS **(in press)**.

Gordon, J. L., et al. (2009). "Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome." PLoS Genet **5**(5): e1000485.

Gouy, M., et al. (2010). "SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building." Mol Biol Evol **27**(2): 221-4.

Greig, D., et al. (2002). "Hybrid speciation in experimental populations of yeast." Science **298**(5599): 1773-5.

Guerreiro, P., et al. (1993). "Clusters of 5S rRNAs in the intergenic region of ubiquitin genes in *Tetrahymena pyriformis*." Biochim Biophys Acta **1216**(1): 137-9.

Gunderson, J. H., et al. (1987). "Structurally distinct, stage-specific ribosomes occur in *Plasmodium*." Science **238**(4829): 933-7.

Gutierrez, J., et al. (2002). "*Candida dubliniensis*, a new fungal pathogen." J Basic Microbiol **42**(3): 207-27.

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

Haft, D. H., et al. (2003). "The TIGRFAMs database of protein families." Nucleic Acids Res **31**(1): 371-3.

Hall, C., et al. (2005). "Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*." Eukaryot Cell **4**(6): 1102-15.

Hall, D. B., et al. (2006). " An HMG protein, Hmo1, associates with promoters of many ribosomal protein genes and throughout the rRNA gene locus in *Saccharomyces cerevisiae*." Mol Cell Biol **26**(9):3672-9.

Hani, J. and H. Feldmann (1998). "tRNA genes and retroelements in the yeast genome." Nucleic Acids Res **26**(3): 689-96.

Hart, C., et al. (2010). "Single-molecule sequencing: sequence methods to enable accurate quantitation." Methods Enzymol **472**: 407-30.

Hicks, J. and J. N. Strathern (1977). "Interconversion of mating type in *S. cerevisiae* and the Cassette model for gene transfer." Brookhaven Symp Biol(29): 233-42.

Hillis, D. M. and M. T. Dixon (1991). "Ribosomal DNA: molecular evolution and phylogenetic inference." Q Rev Biol **66**(4): 411-53.

Hittinger, C. T., et al. (2010). "Remarkably ancient balanced polymorphisms in a multi-locus gene network." Nature **464**(7285): 54-8.

Hittinger, C. T., et al. (2004). "Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts." Proc Natl Acad Sci U S A **101**(39): 14144-9.

Holliday, R. (1964). "A mechanism for gene conversion in fungi." Genet. Res. **5**: 282–304.

Hook-Nikanne, J., et al. (1998). "DNA sequence conservation and diversity in transposable element IS605 of *Helicobacter pylori*." Helicobacter **3**(2): 79-85.

Houseley, J. and D. Tollervey (2011). "Repeat expansion in the budding yeast ribosomal DNA can occur independently of the canonical homologous recombination machinery." Nucleic Acids Res.

Huang, X. and A. Madan (1999). "CAP3: A DNA sequence assembly program." Genome Res **9**(9): 868-77.

Huynen, M. A., et al. (2001). "Inversions and the dynamics of eukaryotic gene order." Trends Genet **17**(6): 304-6.

Ide, S., et al. (2010). "Abundance of ribosomal RNA gene copies maintains genome integrity." Science **327**(5966): 693-6.

James, T. Y., et al. (2006). "Reconstructing the early evolution of Fungi using a six-gene phylogeny." Nature **443**(7113): 818-22.
Jeffreys, A. J. (1979). "DNA sequence variants in the G gamma-, A gamma-, delta- and beta-globin genes of man." Cell **18**(1): 1-10.
Jeffries, T. W., et al. (2007). "Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*." Nat Biotechnol **25**(3): 319-26.
Johnson, F. B., et al. (1999). "Molecular biology of aging." Cell **96**(2): 291-302.
Johnston, M., et al. (1997). "The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII." Nature **387**(6632 Suppl): 87-90.
Kaeberlein, M., et al. (1999). "The SIR2/3/4 complex and SIR2 alone promote longevity in *Saccharomyces cerevisiae* by two different mechanisms." Genes Dev **13**(19): 2570-80.
Kalendar, R., et al. (2008). "Cassandra retrotransposons carry independently transcribed 5S RNA." Proc Natl Acad Sci U S A **105**(15): 5833-8.
Kapitonov, V. V. and J. Jurka (2003). "A novel class of SINE elements derived from 5S rRNA." Mol Biol Evol **20**(5): 694-702.
Kasahara, K., et al. (2007). "Assembly of regulatory factors on rRNA and ribosomal protein genes in *Saccharomyces cerevisiae*." Mol Cell Biol **27**(19):6686-705.
Kaur, R., et al. (2007). "A family of glycosylphosphatidylinositol-linked aspartyl proteases is required for virulence of *Candida glabrata*." Proc Natl Acad Sci U S A **104**(18): 7628-33.
Kellis, M., et al. (2004a). "Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*." Nature **428**(6983): 617-24.
Kellis, M., et al. (2004b). "Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery." J Comput Biol **11**(2-3): 319-55.
Kellis, M., et al. (2003). "Sequencing and comparison of yeast species to identify genes and regulatory elements." Nature **423**(6937): 241-54.
Kim, J. M., et al. (1998). "Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence." Genome Res **8**(5): 464-78.
Kircher, M. and J. Kelso (2010). "High-throughput DNA sequencing--concepts and limitations." Bioessays **32**(6): 524-36.
Kjeldsen, T., et al. (1997). "Synthetic leaders with potential BiP binding mediate high-yield secretion of correctly folded insulin precursors from *Saccharomyces cerevisiae*." Protein Expr Purif **9**(3): 331-6.
Knowles, D. G. and A. McLysaght (2009). "Recent de novo origin of human protein-coding genes." Genome Res **19**(10): 1752-9.
Kobayashi, T. (2008). "A new role of the rDNA and nucleolus in the nucleus--rDNA instability maintains genome integrity." Bioessays **30**(3): 267-72.
Kobayashi, T. (2011). "How does genome instability affect lifespan?: roles of rDNA and telomeres." Genes Cells. **16**(6): 617-624.
Kobayashi, T., et al. (2004). "SIR2 regulates recombination between different rDNA repeats, but not recombination within individual rRNA genes in yeast." Cell **117**(4): 441-53.
Koonin, E. V. and A. R. Mushegian (1996). "Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics." Curr Opin Genet Dev **6**(6): 757-62.

Koonin, E. V., et al. (1997). "Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea." Mol Microbiol **25**(4): 619-37.

Koren, S., et al. (2010). "An algorithm for automated closure during assembly." BMC Bioinformatics **11**: 457.

Korlach, J., et al. (2010). "Real-time DNA sequencing from single polymerase molecules." Methods Enzymol **472**: 431-55.

Koszul, R., et al. (2004). "Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments." Embo J **23**(1): 234-43.

Koszul, R., et al. (2006). "Stability of large segmental duplications in the yeast genome." Genetics **172**(4): 2211-22.

Kress, H., et al. (2001). "Evolution of 5S rRNA gene families in *Drosophila*." Chromosome Res **9**(5): 403-15.

Küberl, A, et al. (2011). " High-quality genome sequence of Pichia pastoris CBS7435." J Biotechnol **154**(4):312-20.

Kurtzman, C. P. (2011). Discussion of Telomorphic and Anamorphic Ascomycetous Yeasts and Yeast-like Taxa. The Yeasts, a Taxonomic Study. C. P. Kurtzman, J. W. Fell and T. Boekhout. Amsterdam, Elsevier**:** 293-307.

Kurtzman, C. P. and J. W. Fell (1998). The yeasts, a taxonomic study. Amsterdam, Elsevier.

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**(4): 417-32.

Kwast, K. E., et al. (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**(1): 250-65.

Labib, K. and B. Hodgson (2007). "Replication fork barriers: pausing for a break or stalling for time?" EMBO Rep **8**(4): 346-53.

Lalo, D., et al. (1993). "Two yeast chromosomes are related by a fossil duplication of their centromeric regions." C R Acad Sci III **316**(4): 367-73.

Lau, A. Y., et al. (2000). "Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG." Proc Natl Acad Sci U S A **97**(25): 13573-8.

Levine, M. T., et al. (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**(26): 9935-9.

Li, D., et al. (2010). "A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand." Cell Res **20**(4): 408-20.

Li, L., et al. (2007). "*Candida glabrata*: an emerging oral opportunistic pathogen." J Dent Res **86**(3): 204-15.

Libkind, D., et al. (2011). "Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast." Proc Natl Acad Sci U S A **108**(35): 14539-44.

Lin, J., et al. (2004). "A universal telomerase RNA core structure includes structured motifs required for binding the telomerase reverse transcriptase protein." Proc Natl Acad Sci U S A **101**(41): 14713-8.

Linardopoulou, E. V., et al. (2005). "Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication." Nature **437**(7055): 94-100.

Liolios, K., et al. (2010). "The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata." Nucleic Acids Res **38**(Database issue): D346-54.

Liti, G., et al. (2009). "Population genomics of domestic and wild yeasts." Nature **458**(7236): 337-41.

Liti, G. and E. J. Louis (2005). "Yeast evolution and comparative genomics." Annu Rev Microbiol **59**: 135-53.

Llorente, B., et al. (2000a). "Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of gene redundancy compared to *Saccharomyces cerevisiae*." FEBS Lett **487**(1): 122-33.

Llorente, B., et al. (2000b). "Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*." FEBS Lett **487**(1): 101-12.

Long, M., et al. (2003). "The origin of new genes: glimpses from the young and old." Nat Rev Genet **4**(11): 865-75.

Louis, E. J. and J. E. Haber (1990). "The subtelomeric Y' repeat family in *Saccharomyces cerevisiae*: an experimental system for repeated sequence evolution." Genetics **124**(3): 533-45.

Louis, E. J., et al. (1994). "The chromosome end in yeast: its mosaic nature and influence on recombinational dynamics." Genetics **136**(3): 789-802.

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**(5): 955-64.

Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**(5494): 1151-5.

Mais, C., et al. (2005). "UBF-binding site arrays form pseudo-NORs and sequester the RNA polymerase I transcription machinery." Genes Dev **19**(1):50-64.

Maleszka, R. and G. D. Clark-Walker (1993). "Yeasts have a four-fold variation in ribosomal DNA copy number." Yeast **9**(1): 53-8.

Manchado, M., et al. (2006). "Molecular characterization and chromosomal mapping of the 5S rRNA gene in *Solea senegalensis*: a new linkage to the U1, U2, and U5 small nuclear RNA genes." Genome **49**(1): 79-86.

Marcet-Houben, M. and T. Gabaldon (2010). "Acquisition of prokaryotic genes by fungal genomes." Trends Genet **26**(1): 5-8.

Marchler-Bauer, A., et al. (2009). "CDD: specific functional annotation with the Conserved Domain Database." Nucleic Acids Res **37**(Database issue): D205-10.

Marchler-Bauer, A., et al. (2011). "CDD: a Conserved Domain Database for the functional annotation of proteins." Nucleic Acids Res **39**(Database issue): D225-9.

Margulies, M., et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-80.

Marques-Bonet, T., et al. (2009). "The origins and impact of primate segmental duplications." Trends Genet **25**(10): 443-54.

Masai, E., et al. (2003). "Roles of the enantioselective glutathione S-transferases in cleavage of beta-aryl ether." J Bacteriol **185**(6): 1768-75.

Melnick, L. and F. Sherman (1993). "The gene clusters ARC and COR on chromosomes 5 and 10, respectively, of *Saccharomyces cerevisiae* share a common ancestry." J Mol Biol **233**(3): 372-88.

Merz, K., et al. (2008). "Actively transcribed rRNA genes in *S. cerevisiae* are organized in a specialized chromatin associated with the high-mobility group protein

Hmo1 and are largely devoid of histone molecules." Genes Dev **22**(9): 1190-204.

Mieczkowski, P. A., et al. (2006). "Recombination between retrotransposons as a source of chromosome rearrangements in the yeast *Saccharomyces cerevisiae*." DNA Repair (Amst) **5**(9-10): 1010-20.

Miranda, I., et al. (2006). "Evolution of the genetic code in yeasts." Yeast **23**(3): 203-13.

Mitrovich, Q. M., et al. (2010). "Evolution of yeast noncoding RNAs reveals an alternative mechanism for widespread intron loss." Science **330**(6005): 838-41.

Miura, F., et al. (2006). "A large-scale full-length cDNA analysis to explore the budding yeast transcriptome." Proc Natl Acad Sci U S A **103**(47): 17846-51.

Muir, G., et al. (2001). "Three divergent rDNA clusters predate the species divergence in *Quercus petraea (Matt.) Liebl.* and *Quercus robur L.*" Mol Biol Evol **18**(2): 112-9.

Nakao, Y., et al. (2009). "Genome sequence of the lager brewing yeast, an interspecies hybrid." DNA Res **16**(2): 115-29.

Negrini, S., et al. (2007). "DNA breaks are masked by multiple Rap1 binding in yeast: implications for telomere capping and telomerase regulation." Genes Dev **21**(3): 292-302.

Nei, M., et al. (1997). "Evolution by the birth-and-death process in multigene families of the vertebrate immune system." Proc Natl Acad Sci U S A **94**(15): 7799-806.

Nei, M. and A. P. Rooney (2005). "Concerted and birth-and-death evolution of multigene families." Annu Rev Genet **39**: 121-52.

Neuveglise, C., et al. (2000). " Genomic exploration of the hemiascomycetous yeasts: 9. *Saccharomyces kluyveri*." FEBS Lett **487**(1):56-60.

Neuveglise, C., et al. (2002). "Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts." Genome Res **12**(6): 930-43.

Nosek, J., et al. (2009). "Biology and genetics of the pathogenic yeast *Candida parapsilosis*." Curr Genet **55**(5): 497-509.

Oesterhelt, D. and J. Tittor (1989). "Two pumps, one principle: light-driven ion transport in halobacteria." Trends Biochem Sci **14**(2): 57-61.

Ogata, H., et al. (1999). "KEGG: Kyoto Encyclopedia of Genes and Genomes." Nucleic Acids Res **27**(1): 29-34.

OhEigeartaigh, S. S., et al. (2011). "Systematic discovery of unannotated genes in 11 yeast species using a database of orthologous genomic segments." BMC Genomics **12**: 377.

Oliver, S. G., et al. (1992). "The complete DNA sequence of yeast chromosome III." Nature **357**(6373): 38-46.

Otsuka, Y., et al. (2003). "Detection and characterization of a novel extracellular fungal enzyme that catalyzes the specific and hydrolytic cleavage of lignin guaiacylglycerol beta-aryl ether linkages." Eur J Biochem **270**(11): 2353-62.

Panaretou, B. and P. W. Piper (1992). "The plasma membrane of yeast acquires a novel heat-shock protein (hsp30) and displays a decline in proton-pumping ATPase levels in response to both heat shock and the entry to stationary phase." Eur J Biochem **206**(3): 635-40.

Pandey, V., et al. (2008). Applied Biosystems SOLiD™ System: Ligation-Based Sequencing. Next Generation Genome Sequencing: Towards Personalized Medicine M. Janitz, Wiley**:** 29-41.

Pankratov, T. A., et al. (2007). "*Mucilaginibacter paludis gen. nov., sp. nov.* and *Mucilaginibacter gracilis sp. nov.*, pectin-, xylan- and laminarin-degrading

members of the family Sphingobacteriaceae from acidic Sphagnum peat bog." Int J Syst Evol Microbiol **57**(Pt 10): 2349-54.

Panozzo, C., et al. (2002). "Aerobic and anaerobic NAD+ metabolism in *Saccharomyces cerevisiae*." FEBS Lett **517**(1-3): 97-102.

Parra, G., et al. (2000). "GeneID in *Drosophila*." Genome Res **10**(4): 511-5.

Parra, G., et al. (2007). "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes." Bioinformatics **23**(9): 1061-7.

Paule, M. R. and R. J. White (2000). "Survey and summary: transcription by RNA polymerases I and III." Nucleic Acids Res **28**(6): 1283-98.

Payen, C., et al. (2008). "Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms." PLoS Genet **4**(9): e1000175.

Pelliccia, F., et al. (2001). "5S ribosomal and U1 small nuclear RNA genes: a new linkage type in the genome of a crustacean that has three different tandemly repeated units containing 5S ribosomal DNA sequences." Genome **44**(3): 331-5.

Petersen, R. F., et al. (1999). "Karyotypes of *Saccharomyces sensu lato* species." Int J Syst Bacteriol **49 Pt 4**: 1925-31.

Piekna-Przybylska, D., et al. (2007). "New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA." Rna **13**(3): 305-12.

Piper, P. W., et al. (1997). "Hsp30, the integral plasma membrane heat shock protein of *Saccharomyces cerevisiae*, is a stress-inducible regulator of plasma membrane H(+)-ATPase." Cell Stress Chaperones **2**(1): 12-24.

Piskur, J. (2001). "Origin of the duplicated regions in the yeast genomes." Trends Genet **17**(6): 302-3.

Piskur, J., et al. (2006). "How did *Saccharomyces* evolve to become a good brewer?" Trends Genet **22**(4): 183-6.

Poláková, S., et al. (2009). "Formation of new chromosomes as a virulence mechanism in yeast Candida glabrata." Proc Natl Acad Sci U S A **106**(8):2688-93.

Pratt-Hyatt, M. J., et al. (2006). "Increased recombination between active tRNA genes." DNA Cell Biol **25**(6): 359-64.

Prokopowich, C. D., et al. (2003). "The correlation between rDNA copy number and genome size in eukaryotes." Genome **46**(1): 48-50.

Pryde, F. E., et al. (1997). "Chromosome ends: all the same under their caps." Curr Opin Genet Dev **7**(6): 822-8.

Qari, S. H., et al. (1994). "Blood and sporozoite stage-specific small subunit ribosomal RNA-encoding genes of the human malaria parasite *Plasmodium vivax*." Gene **150**(1): 43-9.

Quinlan, A. R., et al. (2008). "Pyrobayes: an improved base caller for SNP discovery in pyrosequences." Nat Methods **5**(2): 179-81.

Rachidi, N., et al. (1999). "Multiple Ty-mediated chromosomal translocations lead to karyotype changes in a wine strain of *Saccharomyces cerevisiae*." Mol Gen Genet **261**(4-5): 841-50.

Reedy, J. L., et al. (2009). "Mechanistic plasticity of sexual reproduction and meiosis in the *Candida* pathogenic species complex." Curr Biol **19**(11): 891-9.

Regnacq, M. and H. Boucherie (1993). "Isolation and sequence of HSP30, a yeast heat-shock gene coding for a hydrophobic membrane protein." Curr Genet **23**(5-6): 435-42.

Renault, P., et al. (2009). "Genetic characterization and phenotypic variability in *Torulaspora delbrueckii* species: Potential applications in the wine industry." Int J Food Microbiol **134**(3): 201-10.

Rhind, N., et al. (2011). "Comparative functional genomics of the fission yeasts." Science **332**(6032): 930-6.
Riccombeni, A., et al. (2012). " Analysis of the Genome of the Pathogenic Yeast Candida orthopsilosis." PLoS One **7**(4):e35750.
Rice, P., et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-7.
Rogers, J. H. (1985). "The origin and evolution of retroposons." Int Rev Cytol **93**: 187-279.
Rolland, T., et al. (2009). "Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes." PLoS One **4**(8): e6515.
Ronaghi, M., et al. (1996). "Real-time DNA sequencing using detection of pyrophosphate release." Anal Biochem **242**(1): 84-9.
Rooney, A. P. (2004). "Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in Apicomplexans." Mol Biol Evol **21**(9): 1704-11.
Rooney, A. P. and T. J. Ward (2005). "Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm." Proc Natl Acad Sci U S A **102**(14): 5084-9.
Rubin, G. M. and J. E. Sulston (1973). "Physical linkage of the 5 S cistrons to the 18 S and 28 S ribosomal RNA cistrons in *Saccharomyces cerevisiae*." J Mol Biol **79**(3): 521-30.
Salse, J., et al. (2008). "Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution." Plant Cell **20**(1): 11-24.
Scannell, D. R., et al. (2007a). "Yeast genome evolution--the origin of the species." Yeast **24**(11): 929-42.
Scannell, D. R., et al. (2007b). "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**(20): 8397-402.
Scannell, D. R., et al. (2011). "The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the *Saccharomyces sensu stricto* Genus " G3 **1**(1): 11-25.
Schacherer, J., et al. (2004). "Recovery of a function involving gene duplication by retroposition in *Saccharomyces cerevisiae*." Genome Res **14**(7): 1291-7.
Schneider, J., et al. (2010). "RAPYD - Rapid Annotation Platform for Yeast Data." J Biotechnol **155**(1): 118-26.
Seoighe, C., et al. (2000). "Prevalence of small inversions in yeast gene order evolution." Proc Natl Acad Sci U S A **97**(26): 14433-7.
Seoighe, C. and K. H. Wolfe (1999). "Updated map of duplicated regions in the yeast genome." Gene **238**(1): 253-61.
Servant, F., et al. (2002). "ProDom: automated clustering of homologous domains." Brief Bioinform **3**(3): 246-51.
Shen, J. and L. A. Loeb (2001). "Unwinding the molecular basis of the Werner syndrome." Mech Ageing Dev **122**(9): 921-44.
Sinclair, D. A. and L. Guarente (1997). "Extrachromosomal rDNA circles--a cause of aging in yeast." Cell **91**(7): 1033-42.
Sinclair, D. A., et al. (1997). "Accelerated aging and nucleolar fragmentation in yeast sgs1 mutants." Science **277**(5330): 1313-6.
Soltis, E. D. and P. S. Soltis (2000). "Contributions of plant molecular systematics to studies of molecular evolution." Plant Mol Biol **42**(1): 45-75.

Souciet, J., et al. (2000). "Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies." FEBS Lett **487**(1): 3-12.
Souciet, J. L., et al. (2009). "Comparative genomics of protoploid Saccharomycetaceae." Genome Res **19**(10): 1696-709.
Spingola, M., et al. (1999). "Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*." Rna **5**(2): 221-34.
Stanke, M., et al. (2008). "Using native and syntenically mapped cDNA alignments to improve de novo gene finding." Bioinformatics **24**(5): 637-44.
Stanke, M. and S. Waack (2003). "Gene prediction with a hidden Markov model and a new intron submodel." Bioinformatics **19 Suppl 2**: ii215-25.
Steenkamp, E. T., et al. (2006). "The protistan origins of animals and fungi." Mol Biol Evol **23**(1): 93-106.
Stewart, C., et al. (2011). "A comprehensive map of mobile element insertion polymorphisms in humans." PLoS Genet **7**(8): e1002236.
Studer, R. A. and M. Robinson-Rechavi (2009). "How confident can we be that orthologs are similar, but paralogs differ?" Trends Genet **25**(5): 210-6.
Sudmant, P. H., et al. (2010). "Diversity of human copy number variation and multicopy genes." Science **330**(6004): 641-6.
Sullivan, M., et al. (2004). "Cdc14 phosphatase induces rDNA condensation and resolves cohesin-independent cohesion during budding yeast anaphase." Cell **117**(4): 471-82.
Sullivan, M. and F. Uhlmann (2003). "A non-proteolytic function of separase links the onset of anaphase to mitotic exit." Nat Cell Biol **5**(3): 249-54.
Szostak, J. W. and R. Wu (1980). "Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisiae*." Nature **284**(5755): 426-30.
Tanghe, A., et al. (2004). "Aquaporin-mediated improvement of freeze tolerance of *Saccharomyces cerevisiae* is restricted to rapid freezing conditions." Appl Environ Microbiol **70**(6): 3377-82.
Tatusov, R. L., et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.
Tautz, D. and T. Domazet-Loso (2011). "The evolutionary origin of orphan genes." Nat Rev Genet **12**(10): 692-702.
Tenney, A. E., et al. (2004). "Gene prediction and verification in a compact genome with numerous small introns." Genome Res **14**(11): 2330-5.
Thomson, J. M., et al. (2005). "Resurrecting ancestral alcohol dehydrogenases from yeast." Nat Genet **37**(6): 630-5.
Toh-e, A., et al. (1993). "Three yeast genes, PIR1, PIR2 and PIR3, containing internal tandem repeats, are related to each other, and PIR1 and PIR2 are required for tolerance to heat shock." Yeast **9**(5): 481-94.
Toll-Riera, M., et al. (2009). "Origin of primate orphan genes: a comparative genomics approach." Mol Biol Evol **26**(3): 603-12.
Torres-Machorro, A. L., et al. (2010). "Ribosomal RNA genes in eukaryotic microorganisms: witnesses of phylogeny?" FEMS Microbiol Rev **34**(1): 59-86.
Tsong, A. E., et al. (2003). "Evolution of a combinatorial transcriptional circuit: a case study in yeasts." Cell **115**(4): 389-99.
Tsong, A. E., et al. (2006). "Evolution of alternative transcriptional circuits with identical logic." Nature **443**(7110): 415-20.
Tzfati, Y., et al. (2003). "A novel pseudoknot element is essential for the action of a yeast telomerase." Genes Dev **17**(14): 1779-88.

Tzung, K. W., et al. (2001). "Genomic evidence for a complete sexual cycle in *Candida albicans*." Proc Natl Acad Sci U S A **98**(6): 3249-53.

Ureta-Vidal, A., et al. (2003). "Comparative genomics: genome-wide analysis in metazoan eukaryotes." Nat Rev Genet **4**(4): 251-62.

Vallenet, D., et al. (2006). "MaGe: a microbial genome annotation system supported by synteny results." Nucleic Acids Res **34**(1): 53-65.

van Hoof, A. (2005). "Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication." Genetics **171**(4): 1455-61.

Warringer, J., et al. (2011). "Trait variation in yeast is defined by population history." PLoS Genet **7**(6): e1002111.

Waters, A. P., et al. (1989). "Developmental regulation of stage-specific ribosome populations in *Plasmodium*." Nature **342**(6248): 438-40.

Watt, P. M., et al. (1996). "SGS1, a homologue of the Bloom's and Werner's syndrome genes, is required for maintenance of genome stability in *Saccharomyces cerevisiae*." Genetics **144**(3): 935-45.

Watt, P. M., et al. (1995). "Sgs1: a eukaryotic homolog of E. coli RecQ that interacts with topoisomerase II in vivo and is required for faithful chromosome segregation." Cell **81**(2): 253-60.

Weiner, A. M. (2002). "SINEs and LINEs: the art of biting the hand that feeds you." Curr Opin Cell Biol **14**(3): 343-50.

Wellinger, R. J. (2010). "When the caps fall off: responses to telomere uncapping in yeast." FEBS Lett **584**(17): 3734-3740.

Wendland, J., et al. (1999). "Compact organization of rRNA genes in the filamentous fungus *Ashbya gossypii*." Curr Genet **35**(6): 618-25.

Wicke, S., et al. (2011). "Restless 5S: The re-arrangement(s) and evolution of the nuclear ribosomal DNA in land plants." Mol Phylogenet Evol **61**(2): 321-32.

Wieland, J., et al. (1995). "The PMR2 gene cluster encodes functionally distinct isoforms of a putative Na+ pump in the yeast plasma membrane." Embo J **14**(16): 3870-82.

Will, J. L., et al. (2010). "Incipient balancing selection through adaptive loss of aquaporins in natural *Saccharomyces cerevisiae* populations." PLoS Genet **6**(4): e1000893.

Wolfe, K. (2004). "Evolutionary genomics: yeasts accelerate beyond BLAST." Curr Biol **14**(10): R392-4.

Wolfe, K. H. (2006). "Comparative genomics and genome evolution in yeasts." Philos Trans R Soc Lond B Biol Sci **361**(1467): 403-12.

Wolfe, K. H. and A. J. Lohan (1994). "Sequence around the centromere of *Saccharomyces cerevisiae* chromosome II: similarity of CEN2 to CEN4." Yeast **10 Suppl A**: S41-6.

Wolfe, K. H. and D. C. Shields (1997). "Molecular evidence for an ancient duplication of the entire yeast genome." Nature **387**(6634): 708-13.

Wood, V., et al. (2002). "The genome sequence of *Schizosaccharomyces pombe*." Nature **415**(6874): 871-80.

Wyman, S. K., et al. (2004). "Automatic annotation of organellar genomes with DOGMA." Bioinformatics **20**(17): 3252-5.

Yang, Z. (1997) " PAML: a program package for phylogenetic analysis by maximum likelihood." Comput Appl Biosci **13**(5):555-6.

Yang, Y., et al. (2010). "Annotation confidence score for genome annotation: a genome comparison approach." Bioinformatics **26**(1): 22-9.

Yu, C. E., et al. (1996). "Positional cloning of the Werner's syndrome gene." Science **272**(5259): 258-62.

Zappulla, D. C. and T. R. Cech (2004). "Yeast telomerase RNA: a flexible scaffold for protein subunits." Proc Natl Acad Sci U S A **101**(27): 10024-9.

Zhu, X. Y., et al. (2008). "Molecular and cytological characterization of 5S rDNA in *Oryza* species: genomic organization and phylogenetic implications." Genome **51**(5): 332-40.

Zimmer, E. A., et al. (1980). "Rapid duplication and loss of genes coding for the alpha chains of hemoglobin." Proc Natl Acad Sci U S A **77**(4): 2158-62.

Zou, H. and R. Rothstein (1997). "Holliday junctions accumulate in replication mutants via a RecA homolog-independent mechanism." Cell **90**(1): 87-96.

Zou, S., et al. (1996). "The *Saccharomyces* retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci." Genes Dev **10**(5): 634-45.

Appendix I

Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents

Jonathan L. Gordon, David Armisén, Estelle Proux-Wéra, Seán S. ÓhÉigeartaigh, Kevin P. Byrne, and Kenneth H. Wolfe[1]

Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland

[1] To whom correspondence should be addressed. E-mail khwolfe@tcd.ie

This manuscript is in press in Proceedings of the National Academy of Sciences of the USA.

Abstract

We investigate yeast sex chromosome evolution by comparing genome sequences from 16 species in the family Saccharomycetaceae, including new data from the genera *Tetrapisispora*, *Kazachstania*, *Naumovozyma* and *Torulaspora*. We show that although most yeast species contain a mating-type (*MAT*) locus and silent *HML* and *HMR* loci that are structurally analogous to those of *Saccharomyces cerevisiae*, their detailed organization is highly variable and indicates that the *MAT* locus is a deletion hotspot. Over evolutionary time, chromosomal genes located immediately beside *MAT* have continually been deleted, truncated, or transposed to other places in the genome in a process that is gradually shortening the distance between *MAT* and *HML*. Each time a gene beside *MAT* is removed by deletion or transposition, the next gene on the chromosome is brought into proximity with *MAT* and is in turn put at risk of removal. This process has also continually replaced the triplicated sequence regions, called Z and X, that allow *HML* and *HMR* to be used as templates for DNA repair at *MAT* during mating-type switching. We propose that the deletion and transposition events result from evolutionary accidents during mating-type switching, combined with natural selection to keep *MAT* and *HML* on the same chromosome. The rate of deletion

accelerated greatly after the whole-genome duplication, probably because genes were redundant and could be deleted without requiring transposition. Our results show that mating-type switching imposes a significant mutational cost on the genome, one that must be outweighed by the evolutionary advantages of being able to switch.

Keywords: genome evolution | gene transposition | gene truncation | DNA repair | *Saccharomyces*

The *MAT* locus is the only site in the *Saccharomyces cerevisiae* genome that is continually cleaved and repaired as part of the normal life cycle (1, 2). The *MAT* locus exists in two versions (idiomorphs) that contain either *MAT*a or *MAT*α genes, enabling it to specify three cell types: haploid **a**, haploid α, and diploid **a**/α. Mating-type switching is a programmed DNA rearrangement process that occurs in haploid cells and converts a *MAT*a idiomorph into a *MAT*α idiomorph, or *vice versa*. During switching, DNA at the *MAT* locus is removed and replaced with DNA copied from either the *HML* or the *HMR* locus. *HML* and *HMR* are 'silent cassettes' that store the α-specific and **a**-specific sequence information, respectively, but are transcriptionally inactive due to chromatin modification (2-4).

The ability to switch mating type does not exist in all fungi, but originated independently at least twice (Lee et al. 2010): once in the family Saccharomycetaceae that includes *S. cerevisiae* and *Kluyveromyces lactis*, and once in the Schizosaccharomycetaceae that includes *Schizosaccharomyces pombe*. In the Saccharomycetaceae, switching evolved in a two-step process (6, 7). The first step was the origin of the *HML* and *HMR* cassettes, which occurred at the base of this family after it had diverged from other families such as Debaryomycetaceae and the *Candida albicans* clade (8). All species having silent cassettes, for example *Lachancea waltii* (Di Rienzi et al. 2011), are probably able to switch mating types using the homologous recombination machinery. However, in some clades a second evolutionary step increased the rate and/or precision of switching by directing a dsDNA break to the *MAT* locus in cells that are about to switch. This second step occurred independently, by two different mechanisms, in two groups of yeasts. In the 'post-WGD' clade (species that

underwent whole-genome duplication (Wolfe and Shields 1997)), and their closest relatives such as *Zygosaccharomyces rouxii*, the dsDNA break is made by the HO endonuclease (6, 11). The *HO* gene does not exist outside this clade. In the genus *Kluyveromyces*, the dsDNA break is made by the excision of a mobile genetic element from the *MAT*α idiomorph, during the switch from *MAT*α to *MAT*a (12, 13). The mobile element contains a gene, α3, that is only present in *Kluyveromyces*. To switch in the opposite direction, from *MAT*a to *MAT*α, *Kluyveromyces* induces a dsDNA break at *MAT*a by a different but uncharacterised mechanism (Barsoum et al. 2010).

In all species that have silent cassettes, DNA repair at *MAT* is guided by two regions of sequence (the Z and X regions) that are almost identical between *MAT*, *HML* and *HMR*. In *S. cerevisiae*, the Z region contains the 3' end of the α1 gene, and the X region contains the 3' end of α2 and the 5' end of the neighboring chromosomal gene *BUD5* (Fig. 1). The idiomorph-specific region between them is called Y, the two versions of which (Yα and Ya) have no sequence similarity to one another. In *S. cerevisiae*, switching begins when the HO endonuclease cleaves the Y/Z junction in the *MAT* locus (1, 2, 14). The old *MAT*-Y region is degraded. The Z and X sequences direct the use of *HML* or *HMR* as a template for repair, during which both strands of DNA at *MAT*-Y are newly synthesized in an error-prone fashion (15, 16). Repair is initiated by invasion of a 3' end from *MAT*-Z into the *HM* donor (17, 18), so the first strand of the *MAT*-Y region is always synthesized in the direction from Z to X. Switching takes about one hour (Hicks et al. 2011).

Switching does not occur during every cell cycle, but is a strategy that enables a 'lonely' haploid yeast cell (that is, an isolated single cell that cannot find a partner of the opposite mating type) to produce diploid descendants (19, 20). The haploid cell buds mitotically, the mother cell switches mating type, and the mother and daughter cells then mate to produce a homozygous diploid that can continue to replicate mitotically (Mortimer 2000). In natural populations of *S. paradoxus*, switching has been estimated to occur approximately once per 20,000 cell generations (Liti et al. 2009). The average generation time of natural yeast populations is not known, but generation times of 100 minutes (Warner 1989), 100 hours or 100 days would correspond approximately to one mating-type switch per 4 years, 200 years, or 5000 years respectively. Even at the

lowest of these rates, two yeast species that diverged 10 million years ago would each have gone through 2000 switches since they shared a common ancestor, so switching needs to be efficient and accurate. In this paper, we report evidence that switching errors do accumulate along evolutionary lineages and have had a profound effect on the structure of the *MAT*-containing chromosome in post-WGD species.

Results

Conservation of MAT-HML linkage. We compared *MAT* locus organization in 16 species of the family Saccharomycetaceae (Kurtzman 2011). We augmented existing data with new genome sequences for seven species: two each from the post-WGD genera *Kazachstania*, *Tetrapisispora* and *Naumovozyma*, and one from the non-WGD genus *Torulaspora*. The new data support previous hypotheses that the 3-cassette structure (*MAT*, *HML*, *HMR*) originated at the base of the Saccharomycetaceae (6-8), that the HO endonuclease is younger than the 3-cassette structure (6, 12), and that the loss of the *MAT*a2 gene (6, 25, 26) occurred on the same branch of the phylogenetic tree as the WGD.

Among the 14 species in which mating-type switching appears to be possible, we find that *MAT* and *HML* are always on the same chromosome (86–310 kb apart) and the genotype of *HML* is always α. *HMR* is often on a different chromosome (Dujon et al. 2004), and some species have two *HMR* loci (Scannell et al. 2007b). *HML* and *HMR* are usually but not invariably subtelomeric. The conservation of *HML* and *MAT* in *cis*, and of the α genotype at *HML*, is probably due to conservation of the recombination enhancer (RE) site among species. The RE, which has so far only been found in *S. cerevisiae* (29, 30), is located in the interval between *HML* and *MAT*. It increases the frequency of productive switching by biasing the choice of donor (Szostak and Wu 1980), and operates by binding the α2 protein (31, 32). The two species that may be unable to switch mating type are *Lachancea kluyveri* which has no *HML* or *HMR* (6, 33), and *Kazachstania africana* which appears to have separate *MAT*α and *MAT*a loci due to a genomic rearrangement and has lost *HML*, *HMR* and the *HO* endonuclease gene.

Turnover of Z and X regions. Although the *MAT* loci of most of the species are organized in a manner analogous to that of *S. cerevisiae*, the detailed structure of the Z and X regions varies extensively in terms of which *MAT* genes and neighboring chromosomal genes extend into them (Fig. 2). The X regions of *S. cerevisiae* and *Kaz. naganishii*, for instance, have nothing in common. This variation is surprising because the Z and X regions are virtually identical among the three copies within each genome, and were previously found to be among the most slowly-evolving sequences in the genome (with >96% identity) among four species in the genus *Saccharomyces* that are separated by tens of millions of years (34). Therefore the Z and X region sequences have low rates of nucleotide substitution but can be completely replaced. There is an evolutionary requirement for triplicated sequences flanking *MAT*, *HML* and *HMR* to guide mating-type switching, but the requirement is for triplication *per se* and not for any particular sequence.

A general principle of *MAT* locus organization is that the idiomorph-specific region Yα must contain parts of both the α1 and α2 genes, and Ya must contain parts of the a1 and (where present) a2 gene, so that the gene fragments in the *MAT*-Z and *MAT*-X regions are incapable of expression in cells with the 'wrong' genotype. Beyond this principle, however, it does not seem to matter which *MAT* genes extend into Z and X (Fig. 2), although in species with the HO endonuclease the Y/Z junction has been stabilized to a site in *MAT*α1. *Tetrapisispora phaffii* is puzzling because it seems to violate the principle: it has no Ya region (there is no DNA between the Z and X regions in its *MAT*a idiomorph), so it is not clear how (or if) *MAT*a1 expression is prevented in *MAT*α cells of this species.

Collision and truncation of chromosomal genes flanking *MAT*. The Z and X regions often include parts of flanking chromosomal genes whose functions are not related to cell identity (colored blue in Fig. 2), again with much variation among species. These genes are partially duplicated at *HML* and *HMR*. Remarkably, there is often almost no intergenic DNA between the flanking genes and the *MAT* genes and in some cases they overlap (Fig. 2; Fig. S1). Some flanking genes are truncated, such as *S. cerevisiae BUD5* whose start codon overlaps the stop codon of *MAT*α2. The Bud5 protein is only

half the length of its orthologs in other species, losing an SH3 domain at its N-terminus (Wong et al. 2003). *SLA2*, *SWI6* and *LAA1* in other species are all similarly truncated at their ends closest to *MAT* (Fig. S2). These features are all suggestive of a process that tends to delete nonessential DNA beside the *MAT* locus.

Progressive DNA deletion beside *MAT*. To investigate how the *MAT* locus acquired different flanking genes in different post-WGD species, we compared the genomes to the 'Ancestral' gene order (Gordon et al. 2009) inferred to have existed just before the WGD occurred. In the Ancestral genome nomenclature (Gordon et al. 2009), *HML* and *MAT* are on chromosome 1 (Anc_1), with *HML*α1 and *HML*α2 being the first two genes on this chromosome (Anc_1.1 and Anc_1.2) and the *MAT* locus about 120 genes further along (positions Anc_1.120 to Anc_1.122) (Fig. 3). The genes ancestrally flanking *MAT* are *SLA2* and *DIC1*, an arrangement that appears to be quite old and stable because it is conserved in *Komagataella phaffii (Pichia pastoris)* (37) and *Ogataea* (*Hansenula*) *polymorpha* (Butler et al. 2004). Ancestral chromosome 1 was duplicated as part of the WGD, giving rise to two daughter chromosomes. We call one daughter the '*MAT* chromosome' because it retained the *MAT* and *HML* loci, and the other the 'non-*MAT* chromosome' because it lost its copies of these loci. Both chromosomes underwent further rearrangement after the WGD, but in each post-WGD species the chromosomal regions derived from the *MAT* and the non-*MAT* chromosomes can be identified by tracing the products of each rearrangement event (Gordon et al. 2009) and are shown in Fig. 3.

Strikingly, large deletions are seen on the *MAT* chromosome in each post-WGD species, beginning at the *MAT* locus and extending in the Z direction (leftwards as drawn in Fig. 3). These deletions brought genes that were originally further away in the interval between *HML* and *MAT* into direct proximity with *MAT*. In *Vanderwaltozyma polyspora*, for example, *SWI6* (Anc_1.60) is now the neighbor of *MAT*a1 (Anc_1.122) on the *MAT* chromosome, and almost all the ancestral genes between them were retained on the non-*MAT* chromosome instead (Fig. 3). This non-random distribution of genes between sister chromosomal regions contrasts with the usual pattern of gene losses after WGD (28, 38). The most obvious explanation is that 60 genes were deleted from the *MAT* chromosome in the *V. polyspora* lineage soon after WGD, while most of

its genome was still duplicated. The deletions have different endpoints in different post-WGD species, so that that among the nine post-WGD species in Fig. 3 the current neighbors of *MAT* on the Z side are *KCC4* (Anc_1.52), *SWI6* (Anc_1.60), *EMG1* (Anc_1.64), *TAF2* (Anc_1.76) and *CAN1* (Anc_1.83). A similar but less extensive deletion process has occurred on the other (X) side, where the genes flanking *MAT* are *RNH203* (Anc_1.130), *RCY1* (Anc_1.131) and *BUD5* (Anc_1.134) in different post-WGD species. In *T. blattae*, a translocation has joined the X side of *MAT* to a telomeric region. Rearrangements like this probably cannot occur on the Z side due to the evolutionary constraint to maintain *MAT* and *HML* on the same chromosome.

In contrast to the situation for post-WGD species, none of the non-WGD species show large deletions beside the *MAT* locus. They all retain an organization similar to *Torulaspora delbrueckii*, which is shown for illustration in Fig. 3. In different non-WGD species the genes neighboring *MAT* on the Z-side are *SLA2* (Anc_1.119) and *SUI1* (Anc_1.118), and on the X-side *DIC1* (Anc_1.123), *LAA1* (Anc_1.127), *RNH203* (Anc_1.130) and an unnamed zinc finger gene located between Anc_1.123 and 1.124 (Fig. 2). In *Z. rouxii*, similarly to *T. blattae*, a translocation has joined the X side of *MAT* to a telomeric region.

Gene transpositions provide a timeline. Instead of being deleted, some genes transposed away from the vicinity of the *MAT* locus. For instance *S. cerevisiae JJJ3* (Anc_1.113) is not found in the expected region of the *MAT* or the non-*MAT* chromosome (parts of chromosomes III and XIV, respectively), but instead is on chromosome X (*YJR097W*). *JJJ3* and its neighbor *YJR098C* (Anc_1.114) transposed from the *MAT* chromosome to a new genomic location descended from Ancestral chromosome 7, where they were inserted between genes Anc_7.468 (*YJR096W*) and Anc_7.470 (*YJR099W*). We found 39 separate events of transposition away from *MAT* and use letters A–Z and AA–MM to identify them (Fig. 3). Each transposition event moved 1–3 genes.

The transposition of *JJJ3* and *YJR098C* to the site on Ancestral chromosome 7 (event A) is shared by the genomes of six post-WGD species, so it must have occurred in their common ancestor. Further to the left (Z side) of the *MAT* locus, events B, C, D and E are transpositions shared by *S. cerevisiae* and *Candida glabrata* (they have the same

four insertion sites), but not other species. Further left again, events F, G, H and I are unique to *S. cerevisiae* and then we reach the gene (*TAF2*, Anc_1.76) that is the current neighbor of *MAT*. A similar pattern is seen in each other post-WGD species (Fig. 3; Table S1). It is evident that the genes transposed in a particular order, with those closest to Anc_1.120 moving before those further to the left, over a long time period during which the post-WGD lineages diverged from one another as shown by the phylogenetic tree in Fig. 3.

We therefore infer that the *MAT* locus tends to cause the deletion or transposition of the gene that is its immediate neighbor on the Z-side. When one neighbor is removed, the next comes under attack. During the 100–200 million years since WGD, this process has removed a series of 44–60 *MAT*-neighboring genes in different post-WGD species. A much smaller number of transpositions is seen on the X (right) side, but again an older transposition (event JJ) involved a gene that was ancestrally closer to the *MAT* locus than the younger transpositions (events KK–MM).

Discussion

We hypothesize that the evolutionary deletions, gene truncations, and transpositions beside the *MAT* locus were made during recovery from occasional accidents that occurred during mating-type switching. DNA synthesis during switching in *S. cerevisiae* is highly prone to errors including microhomology-mediated jumps to ectopic templates (Hicks et al. 2010). The evolutionary deletions resemble the long one-sided deletions found extending from the HO site, in the Z direction, in about 2% of *S. cerevisiae* cells in experiments in which the cleaved chromosome was repaired by microhomology-mediated end joining (MMEJ) because no donor sequence was available (Tatusov et al. 2003). During switching in *S. cerevisiae*, the HO double-strand break is processed to generate a long single-stranded tail that can include all the Z region and extend into the flanking gene (*TAF2*) beyond it (White and Haber 1990). If this tail broke and lost the Z region, no homologous donor would be available; to repair the chromosome in a way that satisfies the constraint (imposed by the RE) to keep *MAT* and *HML* in *cis* would require re-ligation by MMEJ, deleting part or all of *TAF2*. If instead the tail invaded some other place in the genome, it could cause transposition of

TAF2 before the *HML-MAT* linkage is restored. The greater extent of deletions and transpositions seen on the Z-side than on the X-side may be because DNA strand exchange initiates in the Z region (17, 18). Successful repair of the chromosome would also require the new sequence flanking *MAT* to be copied to *HML* and *HMR* to become a new Z region; the fact that different chromosomal genes extend into the Z and X regions in different species (Fig. 2) shows that such a feedback mechanism exists.

We infer that a tendency to delete DNA beside the *MAT* locus exists in non-WGD species as well as post-WGD species, because we see flanking gene truncations and some gene deletions in non-WGD species (Fig. 2; Fig. S2). However, the effects of the deletion process are much more visible in post-WGD species (Fig. 3). We hypothesize that the difference is because the WGD brought redundancy into the genome. Suddenly no genes beside the *MAT* locus were essential because they all had a second copy on the non-*MAT* chromosome, so large deletions were possible. As time progressed, duplicated genes were lost from throughout the post-WGD genome, and some genes in the interval between *HML* and *MAT* became single-copy. We propose that when the deletion process brought *MAT* adjacent to an essential single-copy gene, the process stalled until the gene transposed away from beside *MAT*. It is notable that some genes such as *TAM41* (Anc_1.86) transposed independently in multiple lineages to different genomic sites (events G, L, W and Z; Fig. 3 and Table S1). We suggest that its paralog on the non-*MAT* chromosome was lost soon after WGD, making *TAM41* essential and so requiring it to be relocated in each lineage when *MAT* encroached on it. Some patterns of transposition (events HH, AA, BB, and CC) also indicate that a gene can be 'trapped' in the Z region for a period of time while genes further to its left are deleted. *Eremothecium gossypii SUI1* (Anc_1.118) may be an example of a trapped gene because *CWC25* (Anc_1.117) has transposed from between it and *VPS75* (Anc_1.116) (Fig. 2).

Our analysis suggests that errors during mating-type switching, combined with natural selection to keep *MAT* and *HML* on the same chromosome, have subjected the genes flanking the *MAT* locus to a continual process of attempted deletion and occasional transposition during evolution. Deletions were rampant in the immediate aftermath of the WGD but the rate at which *MAT* is moving towards *HML* is slowing because more genes are single-copy and need to be rescued by transposition. The deletion process

removes genes and is therefore likely to impact on the biology of the species in which it occurs. One likely casualty was a cyclin gene similar to *Candida albicans CCN1* (Whiteway et al. 1992), which has no ortholog in *S. cerevisiae*. This gene is located between positions Anc_1.77 and Anc_1.78 in non-WGD species. It has been lost from all post-WGD genomes, except in the genus *Kazachstania* where it survives because the *MAT* locus has only deleted Z-wards as far as Anc_1.83 in that genus (Fig. 3). Another likely casualty is the *MAT*a2 gene itself, whose loss led to rewiring of the cell identity pathway (25, 26).

Our observations about the *MAT* chromosome are reminiscent of the movement of genes out of the mammalian X chromosome (41, 42), but unlike that process we do not suggest that the 'out-of-*MAT*' gene movements are driven by natural selection. Instead, we propose that mating-type switching is accident-prone, and that recovery from these accidents erodes the flanking chromosomal DNA. The fact that switching has been an evolutionarily successful strategy (Liti et al. 2009) implies that it must confer a benefit that outweighs the mutational costs of the deletions described here, and of the error-prone DNA synthesis that occurs during switching (Hicks et al. 2010). What is this benefit? Unlike recombination, switching does not create or maintain any genetic diversity. And since switching occurs both in species that grow primarily as diploids (such as *S. cerevisiae* and most post-WGD lineages) and in others that grow primarily as haploids and sporulate immediately after mating (such as *K. lactis* and most non-WGD lineages), the benefit cannot simply be one of diploidy over haploidy. We suggest that the benefit of switching may be that, in effect, it makes spore germination reversible. Consider a single isolated spore that finds itself in a poor environment. In a yeast species that cannot switch mating types, if the spore germinates it commits itself irreversibly to mitotic growth until it finds a mating partner (Herman and Rine 1997). If the environment is too harsh, this cell lineage will go extinct. In contrast, in a species that can switch, a spore that germinates in a harsh environment can re-form new spores genetically identical to itself after just two mitotic cell divisions (Hicks and Strathern 1977), followed by switching, mating, and sporulation. In this way, mating-type switching may have the benefit of allowing spores to test environments of uncertain quality.

Materials and Methods

Sequencing. The new genomes were sequenced using Roche FLX technology with the aim of achieving high contiguity and establishing the order of genes along chromosomes. We sequenced the type strains, purchased from the Centraalbureau voor Schimmelcultures, of these species in the family Saccharomycetaceae (Kurtzman 2011): *Tetrapisispora phaffii* (CBS 4417; 17 scaffolds), *Tetrapisispora blattae* (CBS 6284; 10 scaffolds), *Naumovozyma dairenensis* (CBS 421; 12 scaffolds), *Kazachstania africana* (CBS 2517; 12 scaffolds), *Kazachstania naganishii* (CBS 8797; 13 scaffolds) and *Torulaspora delbrueckii* (CBS 1146; 7 scaffolds). We also completed the sequence of *Naumovozyma castellii* (CBS 4309; previously called *Saccharomyces castellii* or *Naumovia castellii*; 10 scaffolds), which was draft-sequenced by Cliften *et al.* (45, 46). Sequencing was done under contract by Eurofins MWG Operon. Each genome was sequenced to >20x coverage (> 1 million reads) using a Roche GS FLX instrument with Titanium reagents, with a mixture of paired (3 kb, 8 kb and 20 kb genomic DNA inserts; 1/4 of data each) and unpaired (1/4 of data) sequence reads. Data was assembled into contigs and scaffolds using the Celera assembler (Koren et al. 2010). All inter-contig joins in the scaffold data were checked manually by reference to the paired-end reads and by comparison to other species. All scaffolds appear to correspond to complete chromosomes, except for one unplaced 15-kb scaffold in *T. phaffii*. Ribosomal DNA was assembled and integrated into the scaffolds manually. Mitochondrial genomes were not assembled.

Annotation. We developed a pipeline, to be described in detail elsewhere, that utilizes gene order and sequence data from the YGOB database (Byrne and Wolfe 2005) to annotate new yeast genomes. The pipeline uses an approach based on TBLASTN to overcome frameshift sequencing errors.

Data access. Genomes can be viewed in the YGOB database (temporary URL: http://wolferine.gen.tcd.ie/reviewer with username review and password 3r0s10n). Sequences have been deposited in the EMBL database (accession numbers XXXXXX-XXXXXX).

ACKNOWLEDGMENTS. This work was supported by Science Foundation Ireland and the European Research Council. We thank A. Rourke and F.S. Dietrich for help, and G. Butler for comments on the manuscript.

References

1. Strathern JN *et al.* (1982) Homothallic switching of yeast mating type cassettes is initiated by a double-stranded cut in the *MAT* locus. *Cell* 31: 183-192.
2. Haber JE (1998) Mating-type gene switching in *Saccharomyces cerevisiae*. *Annu Rev Genet* 32: 561-599.
3. Hicks JB, Strathern JN, Herskowitz I (1977) in *DNA Insertion Elements, Plasmids and Episomes*, eds Bukhari A, Shapiro J, Adhya S (Cold Spring Harbor Laboratory Press, New York), pp. 457-462.
4. Herskowitz I, Rine J, Strathern JN (1992) in *The Molecular and Cellular Biology of the Yeast Saccharomyces*, eds Jones EW, Pringle JR, Broach JR (Cold Spring Harbor Laboratory Press, New York), pp. 583-656.
5. Lee SC, Ni M, Li W, Shertz C, Heitman J (2010) The evolution of sex: a perspective from the fungal kingdom. *Microbiol Mol Biol Rev* 74: 298-340.
6. Butler G *et al.* (2004) Evolution of the *MAT* locus and its Ho endonuclease in yeast species. *Proc Natl Acad Sci USA* 101: 1632-1637.
7. Bennett RJ, Johnson AD (2005) Mating in *Candida albicans* and the search for a sexual cycle. *Annu Rev Microbiol* 59: 233-255.
8. Dujon B *et al.* (2004) Genome evolution in yeasts. *Nature* 430: 35-44.
9. Di Rienzi SC *et al.* (2011) Genetic, genomic, and molecular tools for studying the protoploid yeast, *L. waltii*. *Yeast* 28: 137-151.
10. Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387: 708-713.
11. Kostriken R, Heffron F (1984) The product of the *HO* gene is a nuclease: purification and characterization of the enzyme. *Cold Spring Harb Symp Quant Biol* 49: 89-96.
12. Barsoum E, Martinez P, Astrom SU (2010) Alpha3, a transposable element that promotes host sexual reproduction. *Genes Dev* 24: 33-44.
13. Rusche LN, Rine J (2010) Switching the mechanism of mating type switching: a domesticated transposase supplants a domesticated homing endonuclease. *Genes Dev* 24: 10-14.
14. Haber JE (2006) Transpositions and translocations induced by site-specific double-strand breaks in budding yeast. *DNA Repair* 5: 998-1009.
15. Ira G, Satory D, Haber JE (2006) Conservative inheritance of newly synthesized DNA in double-strand break-induced gene conversion. *Mol Cell Biol* 26: 9424-9429.
16. Hicks WM, Kim M, Haber JE (2010) Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. *Science* 329: 82-85.
17. White CI, Haber JE (1990) Intermediates of recombination during mating type switching in *Saccharomyces cerevisiae*. *EMBO J* 9: 663-673.
18. Hicks WM, Yamaguchi M, Haber JE (2011) Real-time analysis of double-strand DNA break repair by homologous recombination. *Proc Natl Acad Sci USA* 108: 3108-3115.
19. Replansky T, Koufopanou V, Greig D, Bell G (2008) *Saccharomyces sensu stricto* as a model system for evolution and ecology. *Trends Ecol Evol* 23: 494-501.
20. Greig D, Leu JY (2009) Natural history of budding yeast. *Curr Biol* 19: R886-890.

21. Mortimer RK (2000) Evolution and variation of the yeast (*Saccharomyces*) genome. *Genome Res* 10: 403-409.
22. Tsai IJ, Bensasson D, Burt A, Koufopanou V (2008) Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle. *Proc Natl Acad Sci USA* 105: 4957-4962.
23. Warner JR (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* 24: 437-440.
24. Kurtzman CP (2011) in *The Yeasts, a Taxonomic Study (5th edition)*, eds Kurtzman CP, Fell JW, Boekhout T (Elsevier, Amsterdam), Vol. 2, pp. 293-307.
25. Tsong AE, Miller MG, Raisner RM, Johnson AD (2003) Evolution of a combinatorial transcriptional circuit: a case study in yeasts. *Cell* 115: 389-399.
26. Tsong AE, Tuch BB, Li H, Johnson AD (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature* 443: 415-420.
27. Fabre E *et al.* (2005) Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing and subtelomeres. *Mol Biol Evol* 22: 856-873.
28. Scannell DR *et al.* (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci USA* 104: 8397-8402.
29. Wu X, Haber JE (1996) A 700 bp cis-acting region controls mating-type dependent recombination along the entire left arm of yeast chromosome III. *Cell* 87: 277-285.
30. Zhou Z, Sun K, Lipstein EA, Haber JE (2001) A *Saccharomyces servazzii* clone homologous to *Saccharomyces cerevisiae* chromosome III spanning *KAR4*, *ARS 304* and *SPB1* lacks the recombination enhancer but contains an unknown ORF. *Yeast* 18: 789-795.
31. Szeto L, Fafalios MK, Zhong H, Vershon AK, Broach JR (1997) Alpha2p controls donor preference during mating type interconversion in yeast by inactivating a recombinational enhancer of chromosome III. *Genes Dev* 11: 1899-1911.
32. Wu C *et al.* (1998) Mcm1 regulates donor preference controlled by the recombination enhancer in *Saccharomyces* mating-type switching. *Genes Dev* 12: 1726-1737.
33. Souciet JL *et al.* (2009) Comparative genomics of protoploid Saccharomycetaceae. *Genome Res* 19: 1696-1709.
34. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241-254.
35. Wong S, Fares MA, Zimmermann W, Butler G, Wolfe KH (2003) Evidence from comparative genomics for a complete sexual cycle in the "asexual" pathogenic yeast *Candida glabrata*. *Genome Biol* 4: R10.
36. Gordon JL, Byrne KP, Wolfe KH (2009) Additions, losses and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet* 5: e1000485.
37. De Schutter K *et al.* (2009) Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nat Biotechnol* 27: 561-566.
38. Dietrich FS *et al.* (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* 304: 304-307.
39. Yu X, Gabriel A (2003) Ku-dependent and Ku-independent end-joining pathways lead to chromosomal rearrangements during double-strand break repair in *Saccharomyces cerevisiae*. *Genetics* 163: 843-856.

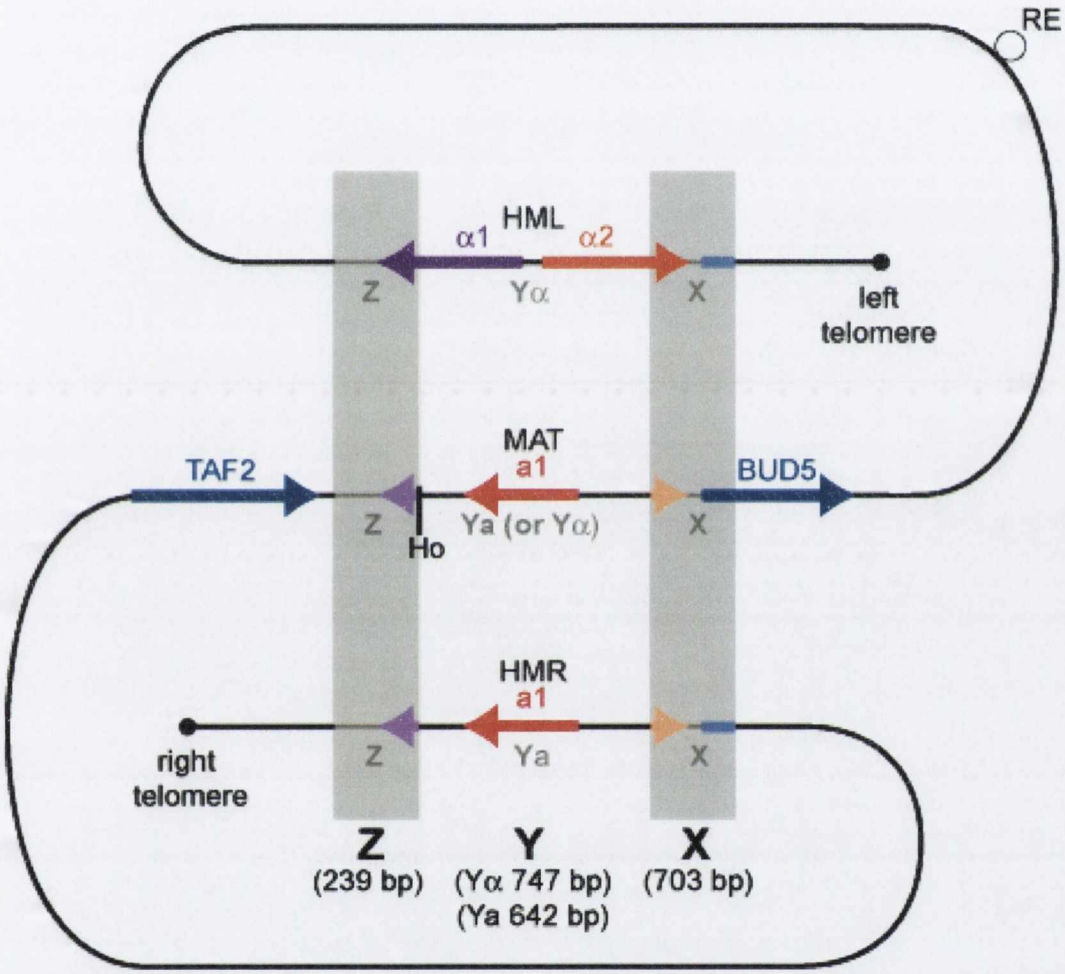
40. Whiteway M, Dignard D, Thomas DY (1992) Dominant negative selection of heterologous genes: isolation of *Candida albicans* genes that interfere with *Saccharomyces cerevisiae* mating factor-induced cell cycle arrest. *Proc Natl Acad Sci USA* 89: 9410-9414.
41. Emerson JJ, Kaessmann H, Betran E, Long M (2004) Extensive gene traffic on the mammalian X chromosome. *Science* 303: 537-540.
42. Potrzebowski L *et al.* (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol* 6: e80.
43. Herman PK, Rine J (1997) Yeast spore germination: a requirement for Ras protein activity during re-entry into the cell cycle. *EMBO J* 16: 6171-6181.
44. Strathern JN, Herskowitz I (1979) Asymmetry and directionality in production of new cell types during clonal growth: the switching pattern of homothallic yeast. *Cell* 17: 371-381.
45. Cliften P *et al.* (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301: 71-76.
46. Cliften PF, Fulton RS, Wilson RK, Johnston M (2006) After the duplication: gene loss and adaptation in *Saccharomyces* genomes. *Genetics* 172: 863-872.
47. Koren S, Miller JR, Walenz BP, Sutton G (2010) An algorithm for automated closure during assembly. *BMC Bioinformatics* 11: 457.
48. Byrne KP, Wolfe KH (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* 15: 1456-1461.
49. Astell CR *et al.* (1981) The sequence of the DNAs coding for the mating-type loci of *Saccharomyces cerevisiae*. *Cell* 27: 15-23.

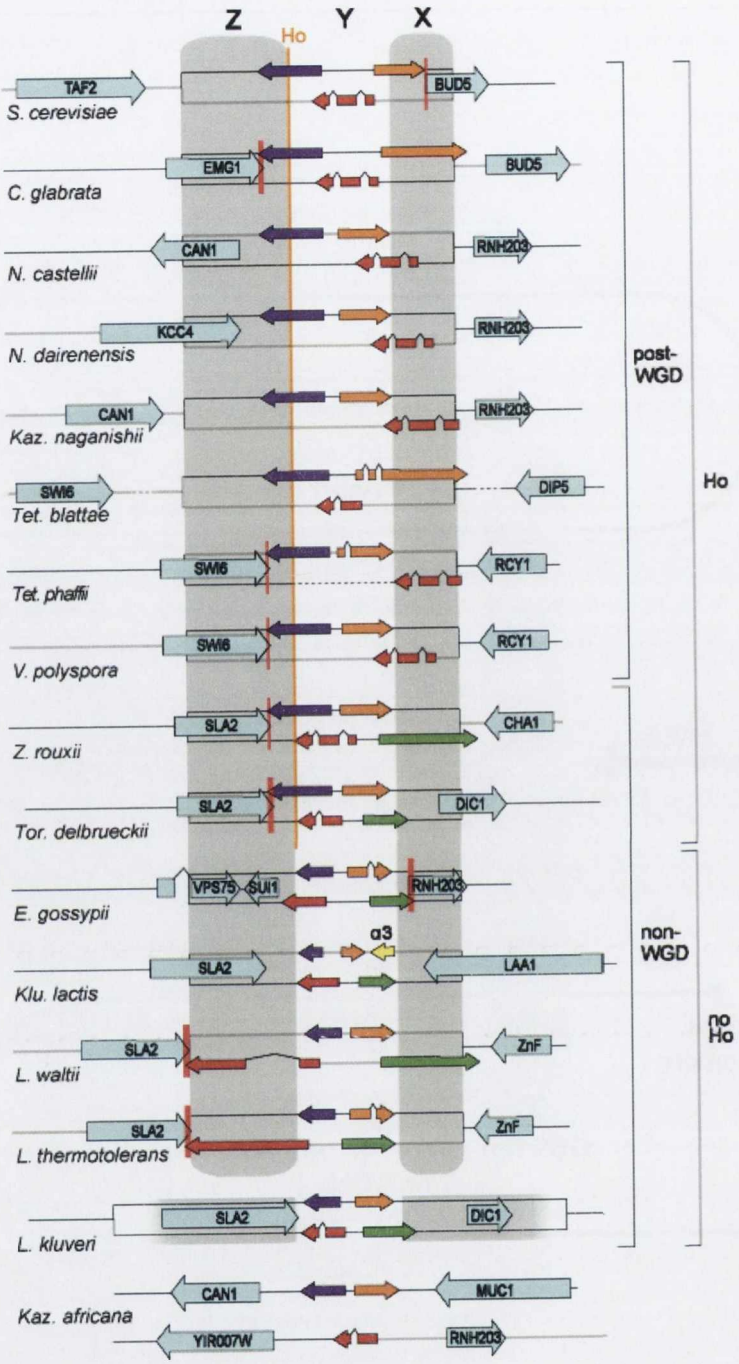
Figure legends

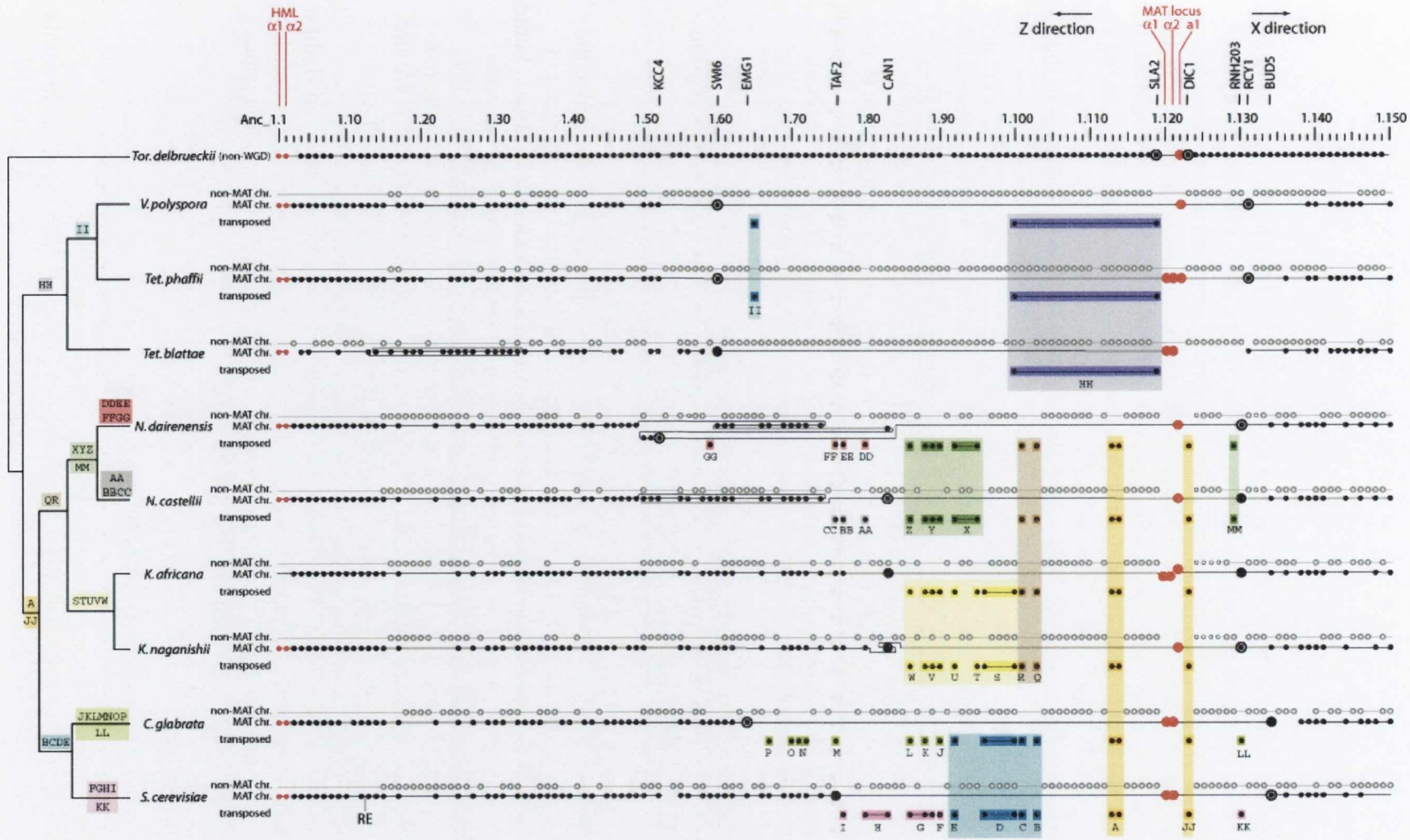
Fig. 1. Organization of the *MAT*, *HML* and *HMR* loci on chromosome III of a *MAT*a *S. cerevisiae* cell. The Z and X regions occur in three copies in parallel orientation and include parts of the α1, α2 and *BUD5* genes. The Y region between them occurs in two versions (idiomorphs), Yα and Ya, which are completely dissimilar. This diagram is reversed relative to the standard *S. cerevisiae* orientation (Haber 1998) to maintain compatibility with Figures 2 and 3 despite species-specific inversions in *S. cerevisiae* (Dujon et al. 2004). Note on nomenclature: We define X and Z as the regions that occur in three copies. In *S. cerevisiae* (2, 49) these are usually called X and Z1, and two duplicated regions that extend the similarity between *MAT* and *HML* (but not *HMR*) beyond them are called W and Z2. There are similar duplicated extensions at the outer edges of the triplicated regions in the other species studied here, but we did not see any consistent patterns of organization.

Fig. 2. Schematic organization (not to scale) of the *MAT* locus in 16 species. Both possible versions of the Y region are shown for each species. Yα contains the genes α1 (purple) and α2 (orange). Ya contains the genes a1 (red) and a2 (green, only in non-WGD species (6, 25)). Caret symbols indicate introns. Gray shading indicates the extent of the Z and X regions. HO endonuclease, where present, cleaves the *MAT* locus at the Y/Z boundary at a site in the α1 gene. Flanking chromosomal genes are shown in blue. Pink vertical bars indicate gene overlaps (broad bars) or intergenic distances ≤5 bp (narrow bars). In *L. kluyveri* there are no *HML* and *HMR* cassettes (33) but the sequenced strain is diploid so only the inner boundaries of Z and X are defined. In *K. africana* there are two *MAT*-like regions and no *HO* gene. The dashed line for *T. phaffii* Ya represents zero length of sequence. Fig. S3 shows the same regions drawn to scale.

Fig. 3. Progressive loss of genes flanking the *MAT* locus by deletion and transposition. The scale indicates gene positions along part of Ancestral chromosome 1, from Anc_1.1 to Anc_1.150. Each circle represents a gene, with *HML* and *MAT* genes in red (each genome sequence is arbitrarily either *MAT*α or *MAT*a). Horizontal lines connect genes that are currently neighbors; zigzags show inversions. For each post-WGD species, genes are assigned to three groups: those derived from the *MAT* chromosome (the chromosome that retained the *MAT* locus after WGD; black circles); those derived from the non-*MAT* chromosome (the paralogous chromosome that lost the *MAT* locus after WGD; open circles); and those that transposed from the *MAT* chromosome to other places in the genome (letters A-Z and AA-MM; colored backgrounds). Each transposition can be inferred to have occurred on a particular branch of the phylogenetic tree on the left, based on the clade of species that share the insertion site, as shown by the different colors. Genes named above the scale are the current neighbors of the *MAT* locus in the species shown here; these genes are identified by bullseyes (for flanking genes that extend into the Z or X regions) or large black circles. Due to a large inversion in *S. cerevisiae* that spans the *MAT* locus and the centromere (Dujon et al. 2004), the Z and X directions as indicated at the top correspond to rightwards and leftwards, respectively, on chromosome III. More details are given in Fig. S4.


RE
HML
α1
α2
Z
Yα
X
left
telomere
MAT
a1
TAF2
BUD5
Z
Ho
Ya (or Yα)
X
HMR
a1
right
telomere
Z
Ya
X
Z
(239 bp)
Y
(Yα 747 bp)
(Ya 642 bp)
X
(703 bp)


Z
Y
X
Ho
TAF2
BUD5
S. cerevisiae
EMG1
BUD5
C. glabrata
CAN1
RNH203
N. castellii
KCC4
RNH203
N. dairenensis
CAN1
RNH203
Kaz. naganishii
SWI6
DIP5
Tet. blattae
SWI6
RCY1
Tet. phaffii
SWI6
RCY1
V. polyspora
SLA2
CHA1
Z. rouxii
SLA2
DIC1
Tor. delbrueckii
VPS75
SUI1
RNH203
E. gossypii
α3
SLA2
LAA1
Klu. lactis
SLA2
ZnF
L. waltii
SLA2
ZnF
L. thermotolerans
SLA2
DIC1
L. kluveri
CAN1
MUC1
YIR007W
RNH203
Kaz. africana
post-WGD
non-WGD
Ho
no Ho


HML
α1 α2
Z direction
MAT locus
α1 α2 a1
X direction
KCC4
SWI6
EMG1
TAF2
CAN1
SLA2
DIC1
RNH203
RCY1
BUD5
Anc_1.1
1.10
1.20
1.30
1.40
1.50
1.60
1.70
1.80
1.90
1.100
1.110
1.120
1.130
1.140
1.150
Tor.delbrueckii (non-WGD)
V.polyspora
Tet.phaffii
Tet.blattae
N.dairenensis
N.castellii
K.africana
K.naganishii
C.glabrata
S.cerevisiae
non-MAT chr.
MAT chr.
transposed
II
HH
DDEE
FFGG
XYZ
MM
AA
BBCC
QR
A
JJ
STUVW
JKLMNOP
LL
BCDE
PQHI
KK
RE

SUPPLEMENTARY INFORMATION

Supplementary Figure 1. Details of the gene overlaps indicated in Fig. 2.

Supplementary Figure 2. Truncation of genes flanking the *MAT* locus. Shown are dot-matrix plots of concatenated protein sequences from all species for (**a**) Bud5, (**b**) Swi6, (**c**) Sla2, (**d**) Laa1, (**e**) Emg1 and (**f**) Rnh203. The number of amino acid residues in each protein is indicated. Pink highlighting denotes regions whose coding sequence lies inside the Z or X region. Each cell in the grids is a comparison between two species. If the proteins are alignable along their whole lengths, the diagonal signal runs into the corners of the cell. If one of the proteins is truncated, the signal does not go into a corner. For example, in **c**, the *Z. rouxii* Sla2 protein has no region homologous to the C-terminus of Sla2 in many other species including *S. cerevisiae*; and in **a**, the *S. cerevisiae* Bud5 protein has no region homologous to the N-terminus of Bud5 in many other species including *C. glabrata*. Bud5, Swi6, Sla2 and Laa1 are severely truncated in some species. Emg1 and Rnh203 are slightly shorter in the species in which they overlap the Z or X regions than in other species. In the comparison between the truncated Sla2 proteins of *T. delbrueckii* and *Z. rouxii*, the Z-overlapping region is seen to be conserved between the two species and the *T. delbrueckii* protein shows a deletion just upstream of the Z region. Plots were made using the program Dotter(Finn et al. 2010).

Supplementary Figure 3. Scale representation of *MAT* loci. Gray polygons show the extents of the Z and X regions.

Supplementary Figure 4. Detailed version of Fig. 3. Gene names are indicated at the ends of each segment of the *MAT* and non-*MAT* chromosomes that remains intact in each species. The names of transposed genes are given in Supplementary Table 1. Some small inversions, and genes not in the ancestral genome, are not shown. Small open circles (e.g. Anc_1.124 in *N. dairenensis*) represent genes whose assignment to the *MAT* or non-*MAT* chromosome was uncertain; these are arbitrarily shown on the non-*MAT* chromosome. Dashed lines denote large inversions in *S. cerevisiae* and *T. blattae* that span the *MAT* locus. In both of these species, the inversion of *MAT* has been compensated by a small inversion of the *HML* genes, maintaining the parallel orientation of *MAT* and *HML*. The topology of the phylogenetic tree is based on Hedtke *et al.*(Hedtke et al. 2006), except that we find that *Tetrapisispora phaffii* is more closely related to *V. polyspora* than to *T. blattae* (PhyML analysis of 30 protein-coding genes retained in duplicate in all species; data not shown).

References for Supplementary Information

1. Sonnhammer, E. L. & Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167, GC1-10 (1995).

2. Hedtke, S. M., Townsend, T. M. & Hillis, D. M. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55, 522-529 (2006).

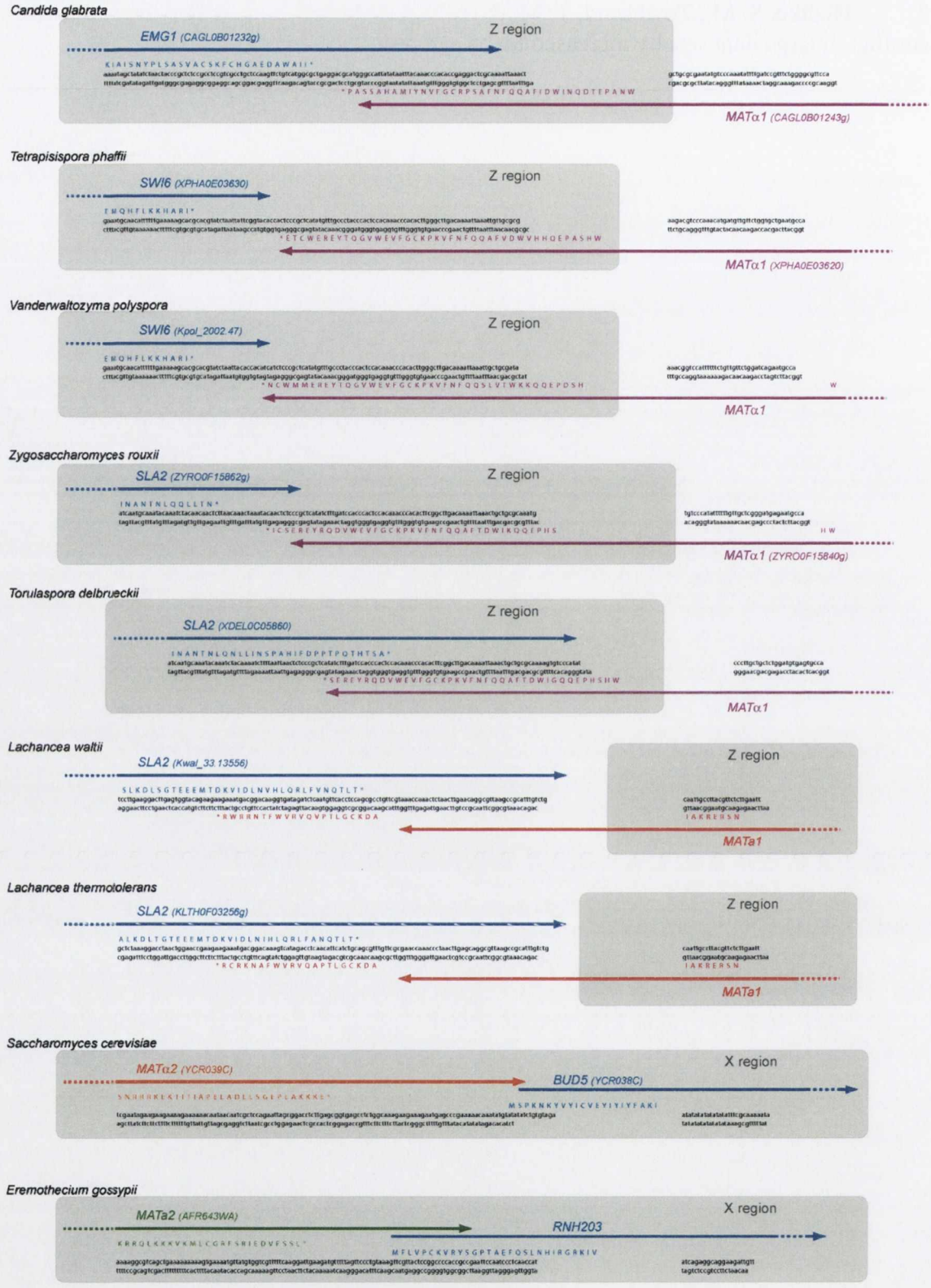

Candida glabrata
EMG1 (CAGL0B01232g)
Z region
MATα1 (CAGL0B01243g)
Tetrapisispora phaffii
SWI6 (XPHA0E03630)
Z region
MATα1 (XPHA0E03620)
Vanderwaltozyma polyspora
SWI6 (Kpol_2002.47)
Z region
MATα1
Zygosaccharomyces rouxii
SLA2 (ZYRO0F15862g)
Z region
MATα1 (ZYRO0F15840g)
Torulaspora delbrueckii
SLA2 (XDEL0C05860)
Z region
MATα1
Lachancea waltii
SLA2 (Kwal_33.13556)
Z region
MATa1
Lachancea thermotolerans
SLA2 (KLTH0F03256g)
Z region
MATa1
Saccharomyces cerevisiae
MATα2 (YCR039C)
BUD5 (YCR038C)
X region
Eremothecium gossypii
MATa2 (AFR643WA)
RNH203
X region

Figure S1

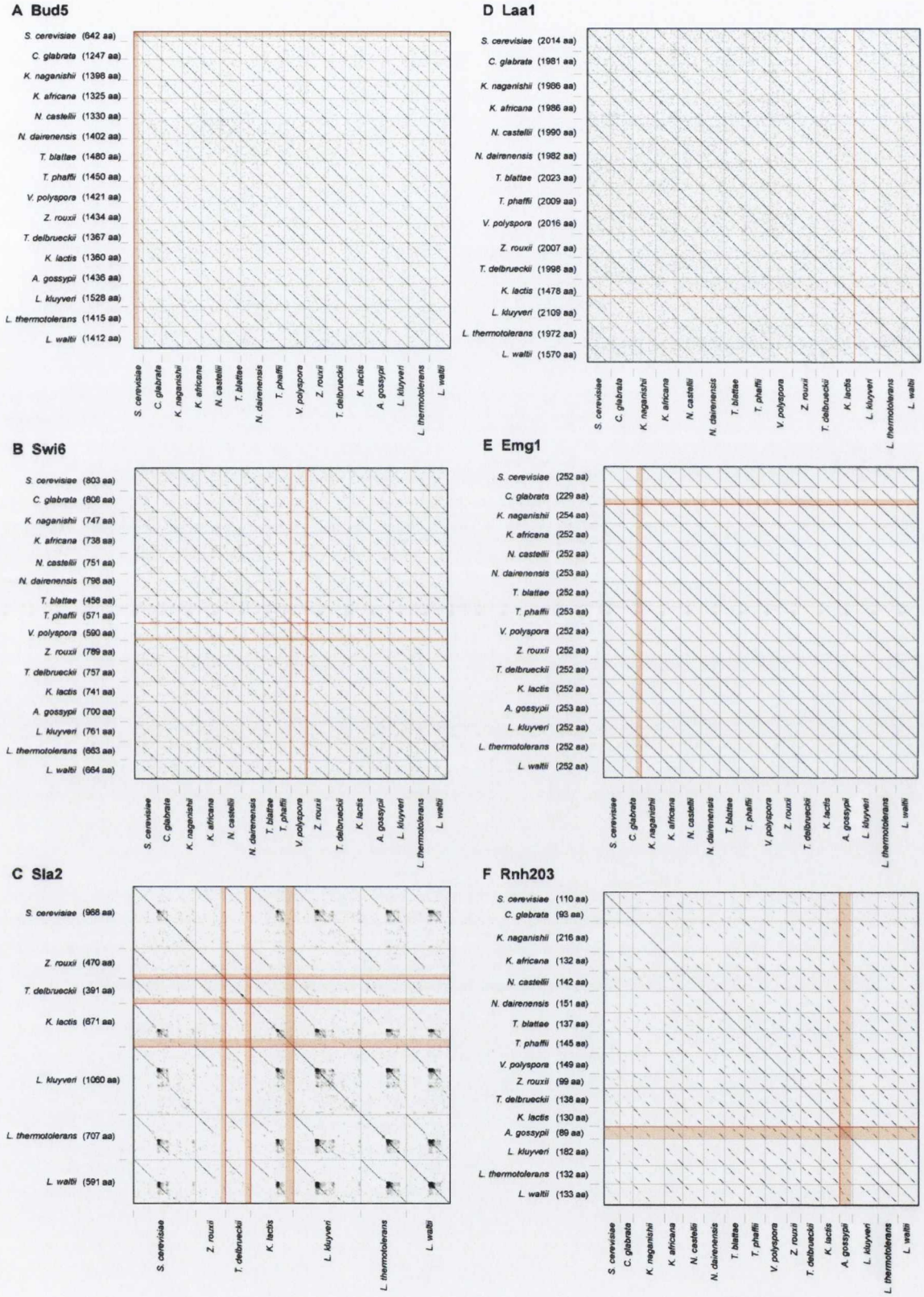

A Bud5
S. cerevisiae (642 aa)
C. glabrata (1247 aa)
K. naganishii (1398 aa)
K. africana (1325 aa)
N. castellii (1330 aa)
N. dairenensis (1402 aa)
T. blattae (1480 aa)
T. phaffii (1450 aa)
V. polyspora (1421 aa)
Z. rouxii (1434 aa)
T. delbrueckii (1367 aa)
K. lactis (1360 aa)
A. gossypii (1436 aa)
L. kluyveri (1528 aa)
L. thermotolerans (1415 aa)
L. waltii (1412 aa)
B Swi6
S. cerevisiae (803 aa)
C. glabrata (808 aa)
K. naganishii (747 aa)
K. africana (738 aa)
N. castellii (751 aa)
N. dairenensis (798 aa)
T. blattae (458 aa)
T. phaffii (571 aa)
V. polyspora (590 aa)
Z. rouxii (789 aa)
T. delbrueckii (757 aa)
K. lactis (741 aa)
A. gossypii (700 aa)
L. kluyveri (761 aa)
L. thermotolerans (663 aa)
L. waltii (664 aa)
C Sla2
S. cerevisiae (968 aa)
Z. rouxii (470 aa)
T. delbrueckii (391 aa)
K. lactis (671 aa)
L. kluyveri (1060 aa)
L. thermotolerans (707 aa)
L. waltii (591 aa)
D Laa1
S. cerevisiae (2014 aa)
C. glabrata (1981 aa)
K. naganishii (1986 aa)
K. africana (1986 aa)
N. castellii (1990 aa)
N. dairenensis (1982 aa)
T. blattae (2023 aa)
T. phaffii (2009 aa)
V. polyspora (2016 aa)
Z. rouxii (2007 aa)
T. delbrueckii (1998 aa)
K. lactis (1478 aa)
L. kluyveri (2109 aa)
L. thermotolerans (1972 aa)
L. waltii (1570 aa)
E Emg1
S. cerevisiae (252 aa)
C. glabrata (229 aa)
K. naganishii (254 aa)
K. africana (252 aa)
N. castellii (252 aa)
N. dairenensis (253 aa)
T. blattae (252 aa)
T. phaffii (253 aa)
V. polyspora (252 aa)
Z. rouxii (252 aa)
T. delbrueckii (252 aa)
K. lactis (252 aa)
A. gossypii (253 aa)
L. kluyveri (252 aa)
L. thermotolerans (252 aa)
L. waltii (252 aa)
F Rnh203
S. cerevisiae (110 aa)
C. glabrata (93 aa)
K. naganishii (216 aa)
K. africana (132 aa)
N. castellii (142 aa)
N. dairenensis (151 aa)
T. blattae (137 aa)
T. phaffii (145 aa)
V. polyspora (149 aa)
Z. rouxii (99 aa)
T. delbrueckii (138 aa)
K. lactis (130 aa)
A. gossypii (89 aa)
L. kluyveri (182 aa)
L. thermotolerans (132 aa)
L. waltii (133 aa)

Figure S2

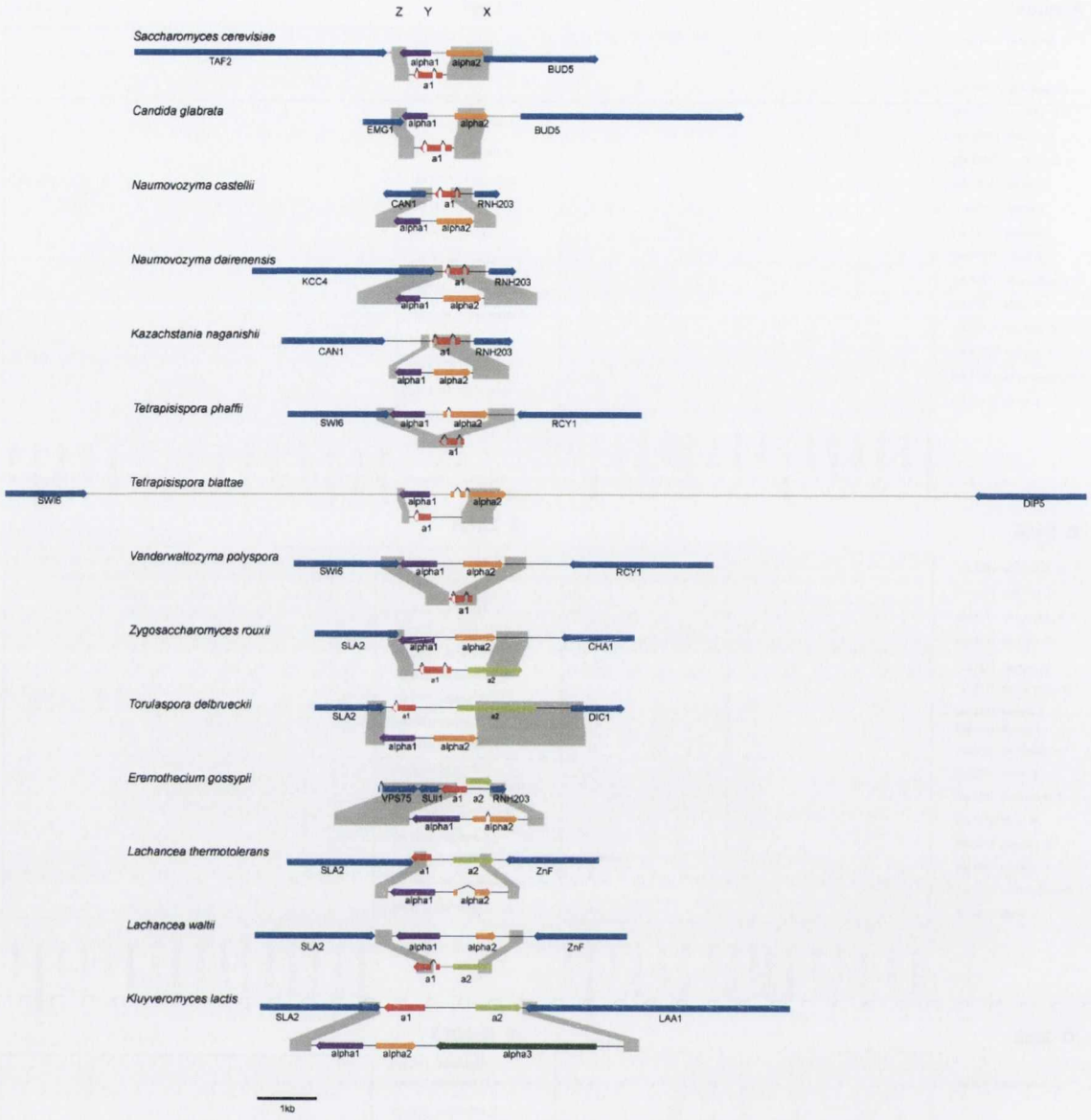

Z
Y
X
Saccharomyces cerevisiae
TAF2
alpha1
alpha2
a1
BUD5
Candida glabrata
EMG1
alpha1
alpha2
a1
BUD5
Naumovozyma castellii
CAN1
a1
RNH203
alpha1
alpha2
Naumovozyma dairenensis
KCC4
a1
RNH203
alpha1
alpha2
Kazachstania naganishii
CAN1
a1
RNH203
alpha1
alpha2
Tetrapisispora phaffii
SWI6
alpha1
alpha2
RCY1
a1
Tetrapisispora blattae
SWI6
alpha1
alpha2
DIP5
a1
Vanderwaltozyma polyspora
SWI6
alpha1
alpha2
RCY1
a1
Zygosaccharomyces rouxii
SLA2
alpha1
alpha2
CHA1
a1
a2
Torulaspora delbrueckii
SLA2
a1
a2
DIC1
alpha1
alpha2
Eremothecium gossypii
VPS75
SUI1
a1
a2
RNH203
alpha1
alpha2
Lachancea thermotolerans
SLA2
a1
a2
ZnF
alpha1
alpha2
Lachancea waltii
SLA2
alpha1
alpha2
ZnF
a1
a2
Kluyveromyces lactis
SLA2
a1
a2
LAA1
alpha1
alpha2
alpha3
1kb

Figure S3

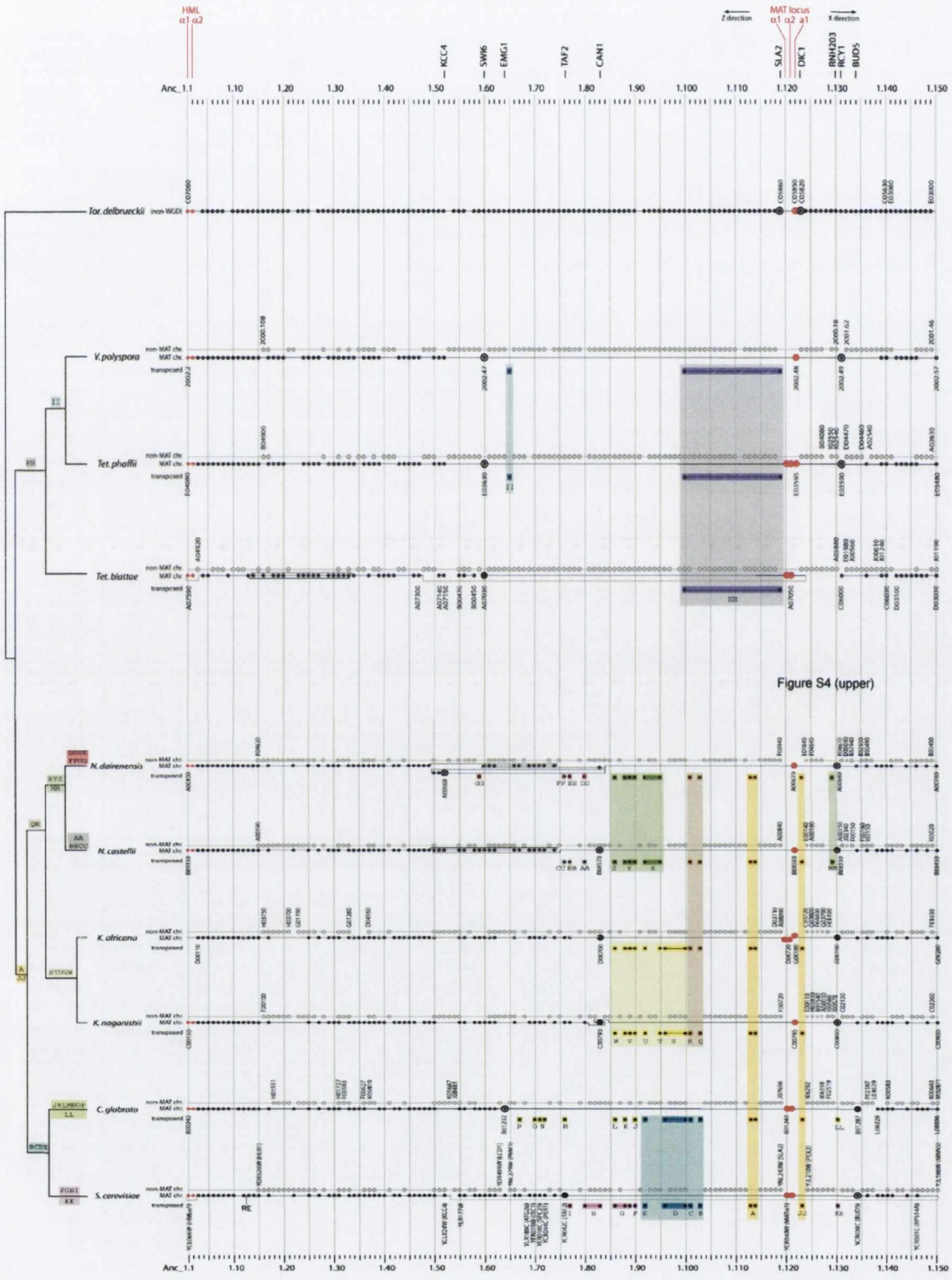


Figure S4 (upper)

Figure S4 (lower)

Table S1. Details of the gene transposition events shown in Figure 3. These events can be viewed in YGOB by entering a gene name or Anc name.

Transposition event	Ancestral (old) location of transposed genes (name of *S. cerevisiae* ortholog in parentheses)	Species with transposition	Transposed gene(s)	Current (new) location†			
				Left neighbor at insertion site		Right neighbor at insertion site	
				Anc name	Gene	Anc name	Gene
A	Anc_1.113 (JJJ3), Anc_1.114 (YJR098C)	Scer	YJR097W (JJJ3), YJR098C	Anc_7.468	YJR096W	Anc_7.470	YJR099W (YUH1)
		Cgla	CAGL0C04565g, CAGL0C04587g	Anc_7.468	CAGL0C04543g	Anc_7.470	CAGL0C04609g
		Knag	XNAG0D01360, XNAG0D01350	Anc_7.468	XNAG0D01370	Anc_7.470	XNAG0D01330
		Kafr	XAFR0G00910, XAFR0G00900	Anc_7.468	XAFR0G00920	Anc_7.470	XAFR0G00890
		Ncas	NCAS0H00550, NCAS0H00540	Anc_7.468	NCAS0H00560	Anc_7.470	NCAS0H00530
		Ndai	NDAI0D00500, NDAI0D00490	Anc_7.467	NDAI0D00510	Anc_7.470	NDAI0D00480
B	Anc_1.103 (COQ6)	Scer	YGR255C (COQ6)	Anc_5.62	YGR254W (ENO1)	Anc_5.55	YGR256W (GND2)
		Cgla	CAGL0F08239g	Anc_5.62	CAGL0F08261g	Anc_5.76	CAGL0F08217g
C	Anc_1.101 (TGL2)	Scer	YDR058C (TGL2)	Anc_3.306	YDR057W (YOS9)	Anc_3.308	YDR059C (UBC5)
		Cgla	CAGL0E04708g	Anc_3.306	CAGL0E04686g	Anc_3.308	CAGL0E04752g
D	Anc_1.96 (LIN1), Anc_1.100 (YSP1)	Scer	YHR156C (LIN1), YHR155W (YSP1)	Anc_5.89	YHR157W (REC104)	Anc_5.94	YHR154W (RTT107)
		Cgla	CAGL0I02002g, CAGL0I01980g	Anc_5.89	CAGL0I02068g	Anc_5.94	CAGL0I01958g
E	Anc_1.92 (CWC23)	Scer	YGL128C (CWC23)	Anc_6.116	YGL127C (SOH1)	Anc_6.227	YGL129C (RSM23)
		Cgla	CAGL0I09526g	Anc_1.336	CAGL0I09504g	Anc_6.227	CAGL0I09548g
F	Anc_1.90 (YGL140C)	Scer	YGL140C	Anc_2.335	YGL141W (HUL5)	Anc_6.241	YGL139W (FLC3)
G	Anc_1.86 (TAM41), Anc_1.88 (TFC4), Anc_1.89 (UFD1)	Scer	YGR046W (TAM41), YGR047C (TFC4), YGR048W (UFD1)	Anc_4.189	YGR044C (RME1)	Anc_4.191	YGR049W (SCM4)
H	Anc_1.80 (AVT1), Anc_1.83 (CAN1)	Scer	YEL064C (AVT2), YEL063C (CAN1)	Anc_6.17	YEL062W (NPR2)	none (telomeric)	YEL065W (SIT1)
I	Anc_1.77 (SLP1)	Scer	YOR154W (SLP1)	Anc_5.500	YOR153C (PDR5)	Anc_5.501	YOR155C (ISN1)
J	Anc_1.90 (YGL140C)	Cgla	CAGL0K08206g	Anc_1.194	CAGL0K08184g	Anc_1.220	CAGL0K08228g
K	Anc_1.88 (TFC4)	Cgla	CAGL0A03630g	Anc_2.398	CAGL0A03586g	Anc_3.122	CAGL0A03652g
L	Anc_1.86 (TAM41)	Cgla	CAGL0G03861g	Anc_4.32	CAGL0G03883g	Anc_4.35	CAGL0G03795g
M	Anc_1.76 (TAF2)	Cgla	CAGL0L00297g	Anc_1.176	CAGL0L00319g	none (telomeric)	CAGL0L00227g
N	Anc_1.71 (SPC25), Anc_1.72 (PER1)	Cgla	CAGL0M11858g, CAGL0M11880g	Anc_7.55	CAGL0M11902g	Anc_7.59	CAGL0M11836g
O	Anc_1.70 (ISC1)	Cgla	CAGL0E06556g	Anc_5.247	CAGL0E06534g	none (telomeric)	CAGL0E06600g
P	Anc_1.67 (MDL1)	Cgla	CAGL0E00385g	Anc_6.377	CAGL0E00407g	none (telomeric)	CAGL0E00363g
Q	Anc_1.103 (COQ6)	Knag	XNAG0B00490	Anc_3.530	XNAG0B00500	Anc_3.531	XNAG0B00480
		Kafr	XAFR0F04240	Anc_3.530	XAFR0F04230	Anc_6.359	XAFR0F04250
		Ncas	NCAS0E00580	Anc_3.530	NCAS0E00570	Anc_3.524	NCAS0E00590
		Ndai	NDAI0I02990	Anc_3.530	NDAI0I02980	Anc_3.531	NDAI0I03000
R	Anc_1.101 (TGL2)	Knag	XNAG0A07210	Anc_2.572	XNAG0A07200	Anc_2.577	XNAG0A07220
		Kafr	XAFR0I01710	Anc_2.572	XAFR0I01720	Anc_2.577	XAFR0I01690
		Ncas	NCAS0H02970	Anc_2.572	NCAS0H02980	Anc_2.585	NCAS0H02960
		Ndai	NDAI0C00700	Anc_2.570	NDAI0C00680	Anc_2.585	NDAI0C00710
S	Anc_1.96 (LIN1), Anc_1.100 (YSP1)	Knag	XNAG0K02480, XNAG0K02490	Anc_3.31	XNAG0K02500	Anc_3.35	XNAG0K02470
		Kafr	XAFR0A08480, XAFR0A08490	Anc_3.30	XAFR0A08500	Anc_3.35	XAFR0A08470
T	Anc_1.95 (ORC5*)	Knag	XNAG0D03580	Anc_4.148	XNAG0D03570	Anc_4.150	XNAG0D03590
		Kafr	XAFR0F03460	Anc_4.148	XAFR0F03450	Anc_4.150	XAFR0F03470
U	Anc_1.92 (CWC23)	Knag	XNAG0B01890	Anc_1.282	XNAG0B01880	Anc_1.284	XNAG0B01910
		Kafr	XAFR0A01530	Anc_1.282	XAFR0A01520	Anc_1.284	XAFR0A01540
V	Anc_1.88 (TFC4), Anc_1.89 (UFD1), Anc_1.90 (YGL140C)	Knag	XNAG0C00550, XNAG0C00540, XNAG0C00530	Anc_1.46	XNAG0C00520	Anc_1.48	XNAG0C00560
		Kafr	XAFR0D00470, XAFR0D00460, XAFR0D00450	Anc_1.47	XAFR0D00440	Anc_1.48	XAFR0D00480
W	Anc_1.86 (TAM41)	Knag	XNAG0H02120	Anc_2.270	XNAG0H02130	Anc_2.272	XNAG0H02110
		Kafr	XAFR0H03410	Anc_2.270	XAFR0H03420	Anc_2.272	XAFR0H03400
X	Anc_1.92 (CWC23), Anc_1.95 (ORC5*)	Ncas	NCAS0H03480, NCAS0H03490	Anc_2.661	NCAS0H03470	Anc_2.662	NCAS0H03520
		Ndai	NDAI0C00200, NDAI0C00190	Anc_2.661	NDAI0C00210	Anc_2.662	NDAI0C00180
Y	Anc_1.88 (TFC4), Anc_1.89 (UFD1), Anc_1.90 (YGL140C)	Ncas	NCAS0B08010, NCAS0B08020, NCAS0B08030	Anc_1.231	NCAS0B08000	Anc_1.235	NCAS0B08040
		Ndai	NDAI0B05320, NDAI0B05310, NDAI0B05300	Anc_1.231	NDAI0B05330	Anc_1.235	NDAI0B05290
Z	Anc_1.86 (TAM41)	Ncas	NCAS0B01420	Anc_8.649	NCAS0B01410	Anc_8.651	NCAS0B01430
		Ndai	NDAI0E01340	Anc_8.649	NDAI0E01350	Anc_8.651	NDAI0E01330
AA	Anc_1.80 (AVT2)	Ncas	NCAS0J01840	Anc_4.305	NCAS0J01830	Anc_4.308	NCAS0J01850
BB	Anc_1.77 (SLP1)	Ncas	NCAS0H02780	Anc_5.567	NCAS0H02770	Anc_5.569	NCAS0H02790
CC	Anc_1.76 (TAF2)	Ncas	NCAS0B07500	Anc_1.311	NCAS0B07510	Anc_1.312	NCAS0B07490
DD	Anc_1.80 (AVT2)	Ndai	NDAI0I00130	Anc_6.350	NDAI0I00140	none (telomeric)	NDAI0I00120
EE	Anc_1.77 (SLP1)	Ndai	NDAI0F02310	Anc_6.190	NDAI0F02300	Anc_6.192	NDAI0F02320
FF	Anc_1.76 (TAF2)	Ndai	NDAI0B04850	Anc_4.29	NDAI0D04860	Anc_4.30	NDAI0D04840
GG	Anc_1.59 (VTA1)	Ndai	NDAI0J02120	Anc_4.227	NDAI0J02100	Anc_4.230	NDAI0J02130
HH	Anc_1.100 (YSP1), Anc_1.119 (SLA2)	Vpol	Kpol_1028.50, Kpol_1028.49	Anc_6.328	Kpol_1028.51	Anc_6.330	Kpol_1028.48
		Tpha	XPHA0D03870, XPHA0D03880	Anc_6.328	XPHA0D03860	Anc_6.330	XPHA0D03890
		Tbla	XBLA0I02320, XBLA0I02330	Anc_6.328	XBLA0I02310	Anc_6.330	XBLA0I02340
II	Anc_1.65 (PRM1)	Vpol	Kpol_388.9	Anc_4.21	Kpol_388.8	Anc_4.25	Kpol_388.10
		Tpha	XPHA0F00270	Anc_4.24	XPHA0F00280	Anc_4.25	XPHA0F00260
JJ	Anc_1.123 (DIC1)	Scer	YLR348C (DIC1)	Anc_4.180	YLR347C (KAP95)	Anc_4.182	YLR350W (ORM2)
		Cgla	CAGL0G01166g	Anc_4.180	CAGL0G01144g	Anc_4.182	CAGL0G01188g
		Knag	XNAG0B05730	Anc_4.180	XNAG0B05720	Anc_4.183	XNAG0B05740
		Kafr	XAFR0A06340	Anc_4.180	XAFR0A06350	Anc_4.182	XAFR0A06330
		Ncas	NCAS0A03540	Anc_4.180	NCAS0A03530	Anc_4.185	NCAS0A03550
		Ndai	NDAI0A03380	Anc_4.180	NDAI0A03370	Anc_4.185	NDAI0A03390
KK	Anc_1.130 (RNH203)	Scer	YLR154C (RNH203)	Anc_8.367	YLR153C (ACS2)	rDNA, Anc_8.376	YLR163C (MAS1)
LL	Anc_1.130 (RNH203)	Cgla	CAGL0F00341g	Anc_1.526	CAGL0F00363g	Anc_1.529	CAGL0F00319
MM	Anc_1.129 (NCE101)	Ncas	NCAS0D04660	Anc_6.349	NCAS0D04650	Anc_6.377	NCAS0D04670
		Ndai	NDAI0I00300	Anc_6.349	NDAI0I00310	Anc_6.377	NDAI0I00290

* The transposed gene is a paralog of *S. cerevisiae ORC5*.

† In most cases the left and right neighbors at the insertion site are a few positions apart in the Ancestral genome (they are not perfectly consecutive due to gene loss after WGD). In a few other cases (gray highlighting), the transposition is to a telomeric region, or coincides with a site of rearrangement relative to the Ancestral genome.