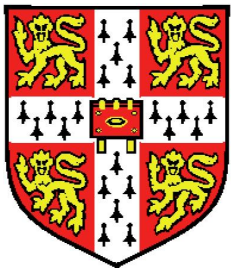
Evolution of Transcription Factor Repertoires in the *Saccharomycotina*



Jaqueline Hess

Clare Hall College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy in Biological Sciences

December 2010

Acknowledgements

I would like to thank my supervisor Nick Goldman for his ongoing support, encouragement and advice and giving me the opportunity to work in such an inspiring environment. I feel priviledged to have been able to work with the intelligent, interesting and most of all very nice present and past members of the Goldman group. Many thanks to Martin Taylor, Tim Massingham, Ari Loytynoja, Greg Jordan, Botond Sipos, Hazel Marsden, Fabio Pardi and Samuel Blanquart for help with many smaller and not so small problems, interesting discussions and a great atmosphere to work in.

I would also like to thank Nick Luscombe and the members of his group, especially Annabel Todd and Juanma Vaquerizas for providing me with valuable data without which this study would have not been feasible and discussions of my results that often opened new perspectives.

Finally, I'm heavily indebted to my friends who have provided me with ongoing support, inspiration and distraction. I'm especially grateful to Kate Downes and Anna Whitelock, who have been of exceptional support during the last few months my PhD. It is your coffee that is keeping me awake as I'm typing this. Many thanks, Jim Bending, Mike Croning, Chris Barnes, Emma Seach, Elodie Roy and Norman Mueller for filling my life with nice music and art, trips to the park and trips to the pub and making Cambridge such a nice place to be. Thanks also to the Herbert Street family, Rosie Peppin-Vaughan, Alastair Willoughby, Gareth Haslam, Nye Jones, Stephanie Grallert and Mike Dodds, who have provided me with a lovely home, lemsip, cups of tea and stews.

This thesis is dedicated to my parents without whose love, support and encouragement I would not be where I am right now.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Abstract

The nature of genetic changes underlying the evolution of phenotypic diversity has been a long-standing subject of evolutionary biology. Changes in gene regulation have been heavily implicated in the evolution of morphological complexity. Indeed, many studies have highlighted the great evolutionary plasticity of *cis*-regulatory regions. The role that transcription factors (TFs) play in the evolution and turnover of transcriptional regulatory networks has been the subject of intensive debate and as yet remains somewhat elusive on the systematic scale.

Here, I present a comprehensive study of TF repertoires in 15 species of yeasts belonging to the *Saccharomycotina*. These species are estimated to have shared their most recent common ancestor around 300 million years ago and include the well-studied model organism *Saccharomyces cerevisiae* as well as the human pathogen *Candida albicans*. Furthermore a whole-genome duplication (WGD) event has been inferred on the lineage leading to the modern *Saccharomyces* species, allowing for assessment of the impact of WGDs for TF repertoire evolution.

In a preliminary study presented in Chapter 2, I investigate phylogenomic approaches to infer a species phylogeny for subsequent evolutionary analyses. I examine the impact of evolutionary models accounting for between-gene heterogeneity through data partitioning and separate parameterisation. I find that partitioned models outperform concatenated models and show that the use of complex evolutionary models is important when analysing phylogenomic data.

The remaining chapters concentrate on various aspects of the evolutionary dynamics in TF repertoires. In Chapter 3, I describe a genome-wide screen for TFs and collection of the dataset which forms the basis of all following analyses. TF repertoires are analysed with respect to the types of DNA-binding domain (DBD) and domain architectures found. Chapter 4 describes an in-depth study of duplications and losses in different DBD families and results are discussed in relation to mechanistic differences in repertoire expansion as well as the regulatory network. In Chapters 5 and 6, I consider evolutionary rates of TFs in relation to their position in the regulatory network (Chapter 5) and examine differences between clades and how those relate to known biological differences (Chapter 6). Stress- and nutrient response signalling in particular appear to have undergone large-scale changes, supported by both expansion of signalling networks through gene duplication and evolutionary rate shifts between clades. This study has generated a number of interesting and experimentally testable examples of potential regulatory turnover, a few of which are discussed in greater detail in Chapter 7.

Contents

List of Figures

Chapter 1

Introduction

1.1 The Aim of this Thesis

One of the biggest challenges for evolutionary biologists in the post-genomic era has been the question of how phenotypic diversity arises. Mouse and human genomes for example share the largest part of their genes, and less than 1% of protein-coding genes in mouse do not have a detectable homolog in the human genome and *vice versa* (Mouse Genome Sequencing Consortium *et al.*, 2002). Even more similar are the human and chimpanzee genomes, where it is not only gene content that is almost identical but also nucleotide divergence is very low with a genome-wide nucleotide divergence rate of around 1.2% (Chimpanzee Sequencing and Analysis Consortium, 2005). A more recent study only found three candidate genes created *de novo* that have not arisen through gene duplication in the human lineage (Knowles & McLysaght, 2009). This amounts to about 0.01% of estimated protein-coding genes in human, thus rendering gene content a very poor explanation of phenotypic diversity.

The idea that not only the genes themselves but also when, how and in what combinations they are expressed provide an important contribution to differences between species was first proposed decades before whole-genome sequencing was even to be considered an achievable scientific endeavour (Britten & Davidson, 1971; Jacob, 1977; King & Wilson, 1975) and has since been the subject of intensive study. Recent technological developments have allowed us to study gene

regulation at various levels, from alternative splicing (e.g. Kim *et al.*, 2007), non-coding RNAs (e.g. Sempere *et al.*, 2006) and, maybe most prominently, changes in transcriptional regulation including *cis*-regulatory sequence turnover and the evolution of transcription factors, the regulatory proteins binding to those sequences (reviewed in Wray, 2007). Indeed, there has been a heated debate about the relative evolutionary importance of changes in *cis*-regulatory DNA and protein coding genes (e.g. Hoekstra & Coyne, 2007). *Cis*-regulatory DNA is often viewed as evolutionarily more flexible due to an alleged lower pleiotropic impact of mutations in non-coding sequences compared to mutations in protein-coding genes, transcription factors especially, which could essentially affect hundreds of downstream target genes (reviewed in Wray, 2007). While we are accumulating increasingly deep understanding about the evolutionary dynamics of *cis*-regulatory DNA, transcription factors have to date been studied either in an isolated manner or over very large evolutionary distances, making it difficult to assess their contribution towards adaptive phenotypes.

This thesis describes a systematic study of transcription factor repertoires, the full genomic complement of transcription factors found within the genome of a species, in 15 species of the *Saccharomycotina* which include the brewer's yeast *Saccharomyces cerevisiae* as well as the human pathogen *Candida albicans*. In the remainder of this Chapter, I will introduce the mechanisms involved in transcriptional initiation and the role that transcription factors play in this process. I will review the arguments of the *cis* versus *trans* debate and motivate why studying the evolution of transcription factors is fundamental to understanding the evolution of gene regulation, based on theoretical arguments and examples of regulatory evolution in the *Saccharomycotina*. Finally, I will discuss the most interesting aspects of the biology and ecology that distinguish the major clades included in my study to provide a reference of how my findings contribute to the understanding of adaptation to their respective niches.

1.2 Transcriptional Regulation in Eukaryotes

1.2.1 Transcriptional Initiation

The default transcriptional state of Eukaryotic genes is "off" and full transcriptional activation requires the execution of a series of molecular events and involvement of a large number of regulatory factors (Figure 1A). In Eukaryotes, DNA is wrapped around nucleosomes which are in turn organised into higher order chromatin structures. Each nucleosome covers 147 base pairs (bp) of DNA and consists of an octamer of histone proteins, incorporating two of each of the histones H2A, H2B, H3 and H4 (Luger *et al.*, 1997). The density of chromatin packing is an important determinant of transcriptional activity, providing a structural barrier for the binding of the transcriptional machinery and thereby preventing transcriptional initiation. Chromatin packing in turn is regulated by chemical modification of histone tails. These modifications are complex and can act alone or in combination to determine the packing state of DNA and association with cofactors that are able to alter condensation states. Known modifications include acetylation, methylation, phosphorylation and ubiqitination although at the moment their function and interactive behaviour is often not well understood (reviewed in Lee *et al.*, 2010a). Silent chromatin in *S. cerevisiae* for example has been found to be associated with hypoacetylation of histone H3 and H4 tails whereas their acetylation is associated with transcriptionally active genes (reviewed in Rusche *et al.*, 2003). This chromatin-mediated mode of gene inactivation is termed "silencing" and not to be confused with repression. As opposed to repression, which is gene-specific, silencing acts at a distance. Sequence-specific transcription factors (TFs) bind to silencer elements and recruit the silencing machinery. This can influence the chromatin structure of a number of surrounding genes such as is observed at telomeres. Silencing can also act in a gene-specific manner, however, as is the case at the silent mating type loci in *S. cerevisiae* (reviewed in Rusche *et al.*, 2003).

Most promoters in *S. cerevisiae* however have an "open" chromatin conformation, where nucleosomes are spaced between 160 and 200 bps apart (Lee *et al.*, 2007; Yuan *et al.*, 2005). Those often include a nucleosome free region (NFR)

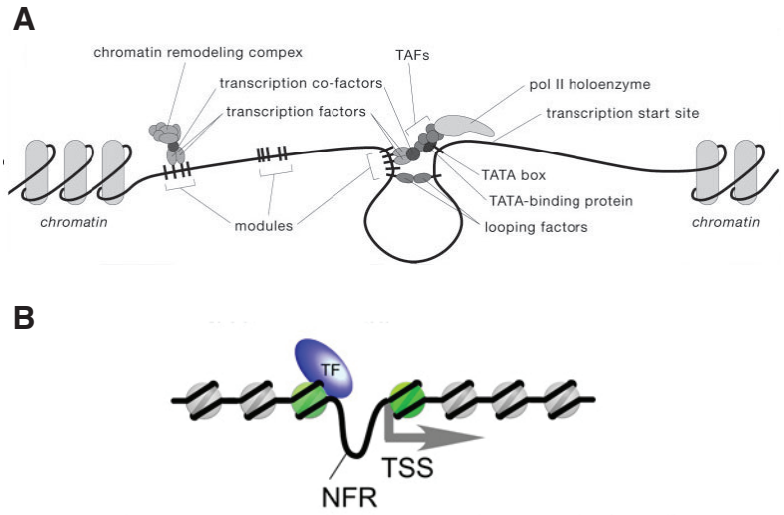


Figure 1.1: **A**: Promoter structure of a eukaryotic gene and the regulatory proteins involved in the modulation of transcriptional activity. Figure adapted from Wray *et al.* (2003). **B**: Nucleosome architecture around active promoters. Nucleosomes -1 and +1 (green) are located around 150 downstream and just upstream of the transcription start site respectively, exposing a nucleosome free region. Figure adapted from Venters & Pugh (2009).

near the transcription start site (TSS), spanning approximately 150 nucleotides located 200 bps upstream of annotated genes (Fig. 1B). This is in contrast to coding regions which were found to be highly occupied by nucleosomes in yeast (Lee *et al.*, 2007). Lee *et al.* furthermore found different categories of nucleosome density at promoters, and found that those correlated with expression level and functional classes of genes. Stress-responsive genes, for example, were most occupied in their data set, in contrast to genes involved in ribosome biogenesis and RNA and DNA metabolism which were found to be most depleted. This reflected the conditions under which the samples were taken, during the log growth phase, where stress-responsive genes are generally not needed and the cells are proliferating at high rates. Moreover, they found a strong statistical correspondence between nucleosome depletion and the presence of binding sites for TFs known to be localised in the nucleus in the sampled conditions, underlining the link between nucleosome organisation at promoters and activation of gene expression.

TFs are sequence-specific DNA-binding proteins that bind to *cis*-regulatory modules (Figure 1A). They perform a variety of roles by recruiting different co-factors, including chromatin remodelling complexes, general transcription factors, chromatin modifying complexes and RNA polymerase II itself via the mediator complex (reviewed in Venters & Pugh, 2009). Depending on the particular nuclear context, the same TF can often be both an activator and repressor by interacting with different condition-specific cofactors. The repressor-activator protein Rap1 is an archetypal example of such a multifunctional TF. Depending on the binding site context, it can act as an activator: e.g. by binding in combination with Gcr1, Gcr2 and Gal11 it induces the high-level expression of the glycolytic enzymes. Alternatively, it can function as a repressor: e.g. during silencing of the HMR and HML silent mating type loci in combination with Orc and Abf1 (reviewed in Shore, 1994). It has also been shown to be involved in maintenance of open chromatin and stimulation of meiotic recombination (reviewed in Morse, 2000).

Once the chromatin structure around a promoter is permissive, general TFs are recruited to the promoter to form the pre-initiation complex (PIC). The interactions between gene-specific TFs and the general transcriptional machinery are thought to be bridged by the Mediator complex, a 21 subunit protein complex which has been shown to interact with both gene-specific TFs and the unphosphorylated form of pol II which forms part of the PIC (reviewed in Björklund & Gustafsson, 2005). Other components of the PIC include the non-specific, broadly utilised general TFs TFIIA, -B, -D, -E, -F, and -H which assist the loading and release of RNA pol II at the TSS (reviewed in Thomas & Chiang, 2006). TFIID is the largest of those multiprotein complexes and includes the TATA-binding protein (TBP). TFIID nucleates the assembly of the PIC, either through direct recruitment or deposition of TBP only by the SAGA complex depending of promoter structure (reviewed in Venters & Pugh, 2009). The promoter occupation by TBP in turn is negatively regulated by the cofactors NC2 and Mot1 which, through combined action, mediate the removal of TBP from promoters. The effects of those negative regulators are counteracted by TFIIA and TFIIB which through interaction with TBP stabilise the TBP/DNA interactions and those together form the minimal PIC (reviewed in Venters & Pugh, 2009). The full PIC is subsequently assembled through binding of the remaining general TFs (TFIIF,

TFIIE and TFIIH) and pol II. TFIIH is the last component that is incorporated into the PIC and is crucial for the transition from transcriptional initiation to transcriptional elongation (reviewed in Venters & Pugh, 2009). TFIIH encodes a DNA helicase, required for the separation of the double strand to allow for the formation of an open promoter complex with pol II (Wang *et al.*, 1992). Furthermore, it contains a kinase subunit which phosphorylates the C-terminal domain of pol II, leading to its dissociation from the Mediator complex and the start of transcriptional elongation (Björklund & Gustafsson, 2005; Venters & Pugh, 2009). Transcriptional elongation itself is a highly regulated process requiring the action of numerous elongation factors and is accompanied by a series of interacting histone modifications, but in the interest of brevity it will not be discussed here.

1.2.2 Transcription Factors

Gene-specific TFs thus play an important role in determining whether or not a given gene is expressed at a given time. By definition TFs contain a sequence-specific DNA-binding domain (DBD) that recognises short motifs in the DNA called transcription factor binding sites (TFBSs), typically around 10 nucleotides in length. TFBSs in turn are organised into *cis*-regulatory modules (Fig. 1A) that often include 10 - 50 binding sites for 5 - 15 different TFs. These can be located within a few hundred base pairs of the promoter, as is commonly found in yeast, but especially in higher eukaryotes may be over tens of kilobases away from the genes they regulate. TFs can bind to certain motifs with varying affinity and whether or not a motif is occupied by a TF depends on a number of other factors, including whether or not the specific TF is present in the nucleus at this time, the condensation state of chromatin and/or binding of nearby TFs that might prevent binding through steric hindrance (reviewed in Wray *et al.*, 2003).

1.2.2.1 The DNA-binding Domain

Depending on the exact structural classification, the currently known number of DBDs ranges from around 150 to 300 (classification using Pfam or SUPERFAMILY respectively; Wilson *et al.*, 2008a, and references therein). Those fall into

eight broad structural classes that differ in their DNA-binding mode (Luscombe *et al.*, 2000). The two most abundant of those are the Helix-Turn-Helix (HTH) and zinc-coordinating DBDs (Wilson *et al.*, 2008a). A summary of the classes of DBDs discussed in Luscombe *et al.* and the families belonging to those is shown in Table 1.1. Note that this table is by no means complete, but is meant to illustrate the type of classification discussed here using well-studied and widely-known examples of DBD families. HTH DBDs, as the name suggests, are formed by two almost perpendicular α-helices that are connected by either a short β-turn or a linker loop. Prokaryotic HTH proteins tend to bind as homodimers to palindromic recognition sequences, whereas in eukaryotes they are frequently observed to bind as monomers or heterodimers to non-symmetrical target sites (Luscombe *et al.*, 2000). Winged HTH DBDs are characterised by the presence of an additional α-helix and an adjacent β-sheet which provide additional contacts with DNA. HTH proteins are the dominant classes of DBDs found in TF repertoires of both bacterial and archeal genomes (Charoensawan *et al.*, 2010a,b; Wilson *et al.*, 2008a).

The second important structural class of DBDs are the zinc-coordinating proteins. These include the C_2H_2 and binuclear cluster (Zn_clus) zinc fingers, which dominate eukaryotic TF repertoires. Zinc-coordinating proteins are characterised by the tetrahedral coordination of zinc ions, mediated by cysteine and histidine residues. In C_2H_2 zinc fingers this structure forms a finger-like protrusion that directly contacts DNA and each finger recognises a 3 bp motif. TFs often contain several adjacent such fingers, increasing the specificity of DNA-binding (reviewed in Luscombe *et al.*, 2000). Human C_2H_2-containing zinc finger TFs have been found to encode 8.5 zinc fingers on average, with some containing more than 30 repeats of the motif (Emerson & Thomas, 2009). Such poly-zinc fingers wrap around the DNA in a spiral-like manner with each protrusion contacting their 3 bp half-site, although especially in TFs with a very large number of fingers it is unlikely that all fingers make contact with DNA and selective binding of a subset of fingers has indeed been demonstrated before (Pavletich & Pabo, 1993).

Class	Family	Viral	Prokaryotes	Eukaryotes	Dimerisation
HTH	Cro and Repressor	x	—	—	Y
	Homeodomain	—	x	x	Y/N
	LacI	—	x	—	Y
	Prd paired domain	—	—	—	N
	Trp repressor	—	—	—	Y
	Diphteria tox	—	x	—	Y
Winged HTH	TFIIB	—	—	x	N
	Interferon regulatory factor	—	—	—	Y
	Catabolite gene activator (CAP)	—	—	—	Y
	Heat-shock and E2F/DP	—	—	x	N
	Ets domain	—	—	—	N
Zinc-coodinating proteins	C_2H_2	—	—	x	N
	Hormone receptor	—	—	x	Y
	Loop-sheet-helix	—	—	x	N
	Zn_clus	—	—	x	N
Zipper-type proteins	Leucine Zipper	—	—	x	Y
	Helix-loop-helix	—	—	x	Y
Other alpha-helix proteins	Pappilomavirus E2	x	—	—	Y
	Skn-1	—	—	x	N
	High-mobility group (HMG)	—	—	x	N
Beta-sheet proteins	MADS box	—	—	x	Y
	TBP	—	—	x	N
β-hairpin/ribbon proteins	MetJ repressor	—	x	—	Y
	T-domain	—	—	x	Y
Other	Arc repressor	x	—	—	Y
	STAT	—	—	x	Y

Table 1.1: Characterised structural classes of DNA-binding domains and the main families within those. This table is based on the data collected by Luscombe *et al.*, 2000, but omitting families and classes not relevant to transcriptional regulation.

1.2.2.2 Non-DNA-binding Regions of Transcription Factors

Besides the DBD, TFs often harbour additional "accessory domains" involved in regulating the activity of the TF itself by mediating protein-protein interactions, activation potential or metabolite binding (Fig. 1.2). As mentioned above, a TF's activity is often determined by context-dependent interaction with other regulatory components and indeed, many TFs are also capable of forming homo- or heterodimers (Table 1.1; Luscombe *et al.*, 2000). Especially in eukaryotes, heterodimerisation between TFs from the same family occurs often and provides mechanisms for combinatorial control. These include dimer- and hence context-dependent DNA-binding specificity as well as additional regulatory mechanisms by providing concentration-dependent switches through stochiometrical requirements for certain complexes to form. Heterotypic interactions with binding partners that lack DNA-binding ability or activation potential can lead to the complete inactivation of an interacting TF through dimerisation with the latter (reviewed in Amoutzias *et al.*, 2008) or concentration dependent effects where the inactive component acts as a molecular titer. Molecular titering in turn can result in ultrasensitive behaviour, where small changes in input concentrations, protein degradation rates or interaction strength of the partners involved can yield large changes in the concentration of the active TF (e.g. Buchler & Louis, 2008).

The sulfur metabolism regulators in *S. cerevisiae* provide a classic example of regulation through heterotypic interactions. Met4, the main regulator of sulfur metabolic genes, lacks intrinsic DNA-binding ability and relies on a number of different co-factors for tethering to its target promoters, whereas some of the co-factors, Met31 and Met32, lack transcriptional activation ability. In order to activate sulfur metabolic genes, Met4 complexes with either Cbf1 or Met31/Met32 and the set of target genes dependent on either of those complexes has been shown to be distinct, thus providing a fine-tuning of sulfur metabolism (Lee *et al.*, 2010c).

Three main types of activation domain have been characterised to date. These are acidic domains, which contain a large proportion of negatively charged residues, proline-rich domains and glutamine-rich domains. Acidic domains are among the best-understood of those and have been shown to activate transcription by interaction with the general transcription factors TFIID and TFIIB, the first

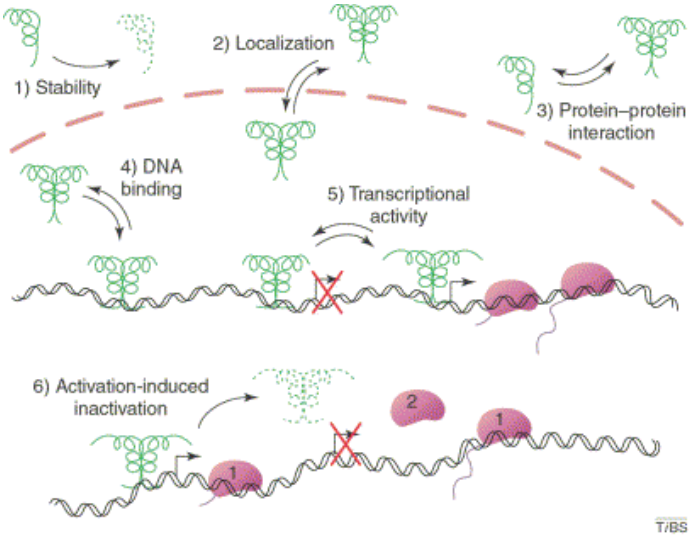


Figure 1.2: Regulating the regulator. Illustration of different mechanisms by which the activity of a TF can be controlled. Post-translational modification and induced conformational changes play an important role in all of these mechanisms and are discussed further in the text. Figure taken from Holmberg *et al.* (2002).

general TFs recruited to promoters to form the PIC (Latchman, 2007). Acidic domains have also been shown to interact directly with components of the mediator complex (Herbig *et al.*, 2010) as well as with the chromatin-remodelling and histone-modification machinery (reviewed in Erkine, 2004) and are thus able to act synergistically on transcriptional activation through a variety of mechanisms. More recent work is continuing to identify new modes of regulation, such as interaction with the universal, ubiquitin-like regulator SUMO through a sumo interaction motif that has been shown to act as an activation domain and resembles the acidic domains (Du *et al.*, 2010), underlining both the complexity of the interactions involved in transcriptional activation as well as how limited our understanding of these mechanisms is. Activation domains often lack an intrinsic structure, especially when unbound, and are thus somewhat idiosyncratic to detect without experimental evidence seeing that their main hallmark is high densities of either acidic amino acid, proline or glutamine residues.

Ligand-binding domains are common in bacterial TFs but are less frequently found in eukaryotic TFs. Eukaryotes have increasingly "outsourced" their metabo-

lite sensing components and instead integrate this information through protein-protein interactions which allow for greater regulatory flexibility (Charoensawan *et al.*, 2010b; Huang *et al.*, 1999). Recent studies indicate that a similar system exists in the archaea, whose DNA-binding proteins, although largely orthologous to bacterial TFs, often lack the additional ligand-binding domains required (Pérez-Rueda & Janga, 2010). Although less common than in prokaryotic TFs, there are however examples of ligand-binding in eukaryotic TFs, such as the *S. cerevisiae* TF Cup2 which contains a poly-copper binding element which confers Cup2 its regulatory activity based on the copper ion concentration in the cell (reviewed in Rutherford & Bird, 2004) or the nuclear hormone receptor TFs (reviewed in Mangelsdorf *et al.*, 1995).

Furthermore, TFs are heavily post-translationally regulated and other motifs found in non-DBD parts of these proteins are related to their own regulation. Those include nuclear localisation signals (NLS) and nuclear export signals (NES), short sequences that facilitate import and export from the nucleus. In non-activating conditions, NLSs are often masked through the binding of interaction partners which are released upon activation, leading to nuclear import of the TF. Similarly an exposed NES will mediate nuclear export once the TF is deactivated. In reality this process can be very complex, involving multiple additional layers of control and in some cases constant nucleocytoplasmic shuttling of TFs, where the predominant subcellular localisation of a TF is determined by the relative strength of import and export signals (reviewed in Ziegler & Ghosh, 2005). Also, phosphorylation, the addition or removal of one or several negatively charged phosphate groups by protein phosphatases or kinases, is an important way to regulate a TF's activity. Phosphorylation induces conformational changes or alters the affinity to interact with other proteins. TFs can be targeted by many different kinases, integrating information from different signalling pathways as well as harbouring multiple phosphorylation sites that have different effects on the TF's activity. For example, mammalian heat-shock factor Hsf1 has at least five phosphorylation sites some of which have activating and some repressive properties, and its overall transcriptional activity is thought to be determined by the balance of the modifications at different sites (reviewed in Holmberg *et al.*, 2002). Ubiquitination of some TFs and their subsequent degradation through the

ubiquitin-proteasome pathway has also been shown to be important for the disassembly of transcriptional complexes and removal of deactivated TFs (reviewed in Kodadek *et al.*, 2006). Similarly, SUMOylation, a related pathway, affects activating potential of TFs, although in this case conjugated proteins are not targeted for degradation through the proteasome but instead SUMOylation might prevent ubiquitination of the same residues as well as having been shown to exert repressive effects on transcriptional initiation, possibly through interaction with transcriptional co-repressors (reviewed in Anckar & Sistonen, 2007).

All together, these mechanisms interact to control a TF's activity by changing their interaction potential, subcellular localisation or protein stability, thus providing vast potential for combinatorial control and integration of various signals (Spoel *et al.*, 2010). So while the DBD influences where a TF is exerts its action, the non-DBD part determines when and how this action takes place, and whether the effect on transcriptional initiation is positive or negative, either through direct modifications in the non-DBD or the interaction with co-factors which are in turn under control of regulatory mechanisms of their own.

From a mechanistic perspective, molecular recognition of target sites and affinity for interaction with cofactors are dependent on the conformational state of the TF (e.g. Boehr *et al.*, 2009). Proteins do not exist in a single conformation at a certain point in time but rather form an ensemble of conformational states where each state is present at different concentrations. Ligand- or cofactor-binding, post-translational modification and DNA-binding can alter the relative concentrations of conformational states of TFs through the induction of allosteric changes (reviewed in Boehr *et al.*, 2009; Pan *et al.*, 2009 and Pan *et al.*, 2010). Whether or not a TF occupies a particular binding site and whether and how this affects transcriptional initiation are thus functions of the predominant conformational states of the TF, its nuclear localisation, accessibility of the binding site itself and the conformational states and localisation of its interaction partners.

TFs together "collect" information about the metabolic, developmental and environmental parameters that the cell is faced with and through combinatorial interaction translate this information into a dynamic output. Depending on its exact location, the impact of a mutation in a TF can conceivably range from small changes in relative concentrations of conformational states (e.g. through

increase or decrease in interaction potential with a cofactor) to complete loss of function (e.g. through loss of DNA-binding ability or activation potential). Understanding the evolutionary dynamics shaping the divergence of TFs is thus of great importance for the understanding of evolution of gene regulation as a whole.

1.2.2.3 Transcription Factor Repertoires — the Complement of Transcription Factors in a Genome

Due to the early availability of fully-sequenced prokaryotic genomes, many of the larger studies of TF repertoires initially focussed on bacterial and archaeal genomes (e.g. Aravind *et al.*, 2005; Minezaki *et al.*, 2005; Pérez-Rueda *et al.*, 2004). The first comparative study of TF repertoires in eukaryotes was published by Riechmann *et al.* (2000) and included the mustard weed *Arabidopsis thaliana*, *C. elegans*, *Drosophila melanogaster* and *S. cerevisiae*. A number of studies have followed since, focussing on plants (Shiu *et al.*, 2005), fungi (Shelest, 2008) and human (Vaquerizas *et al.*, 2009) as well as more recent cross-kingdom comparisons (Charoensawan *et al.*, 2010a,b). Two common themes about TF repertoires have emerged from these studies. Firstly, there is a correlation between the number of protein coding genes and the number of TFs encoded within a genome (Charoensawan *et al.*, 2010a; Pérez-Rueda *et al.*, 2004; Riechmann *et al.*, 2000; van Nimwegen, 2003). This relationship was found to be exponential: whereas very small genomes (such as found in obligate parasites or symbionts) include as little as 0.3% TFs among their protein-coding genes, larger genomes encode increasingly larger percentages of TFs (Babu *et al.*, 2006a; Charoensawan *et al.*, 2010a; Levine & Tjian, 2003; Pérez-Rueda *et al.*, 2004). This correlation was found to be stronger for prokaryotes, however, and the exponential increase in TF percentage with genome size was slower in eukaryotes, probably reflecting the more complex modes of regulation in eukaryotes (e.g. through combinatorial interactions; Ahnert *et al.*, 2008; Balaji *et al.*, 2006a and see above) thereby diminishing the need for increasingly larger number of TFs (Charoensawan *et al.*, 2010a).

Secondly, TF repertoires differ in their composition between the superkingdoms and between clades and are often highly asymmetrical with respect to the

contributing DBD families. Bacterial and archaeal TF repertoires are heavily dominated by HTH DBD proteins, with as much as 80% of TFs in those species belonging to this class (Charoensawan *et al.*, 2010a; Pérez-Rueda *et al.*, 2004). Approximately half of the bacterial DBD families are shared between the three major bacterial phyla (Charoensawan *et al.*, 2010b). Interestingly, although the transcriptional machinery in archaea is more similar to the eukaryotic one, the archaeal TF repertoires share most DBD families with the bacterial repertoires (Aravind & Koonin, 1999; Charoensawan *et al.*, 2010a; Minezaki *et al.*, 2005; Pérez-Rueda *et al.*, 2004). Eukaryotic TF repertoires in turn are mainly dominated by zinc-coordinating TFs and in contrast to bacterial repertoires show more distinct patterns of amplification between the metazoa, fungi and viridiplantae kingdoms (Riechmann *et al.*, 2000; Shelest, 2008; Shiu *et al.*, 2005). Indeed, very few families (three in total) are shared between the three superkingdoms (Charoensawan *et al.*, 2010b). Although most DBD families are shared between eukaryotes, different families have been heavily amplified in different lineages, e.g. the nuclear hormone receptor in *C. elegans*, KRAB-zinc fingers in human or the fungal-specific binuclear cluster zinc fingers ($Zn(II)_2Cys_6$) in *S. cerevisiae* and related fungi (Charoensawan *et al.*, 2010b; Riechmann *et al.*, 2000; Shelest, 2008; Vaquerizas *et al.*, 2009). In contrast to metazoan and fungal genomes however, where often a single DBD family dominates the TF repertoire, this asymmetry is less apparent in plants where several large families are found and in addition to a large number of C_2H_2 zinc fingers, Myb-DNA binding proteins, Helix-loop-helix and MADS box TFs also contribute large numbers of TFs to the repertoires (Riechmann *et al.*, 2000; Shiu *et al.*, 2005).

Overall these lineage-specific patterns of DBD occurrence underline the importance of comparatively recent and frequent gene duplication in TF repertoire growth. Currently we have relatively little idea about whether these lineage-specific expansions reflect functional differences between clades and this is not aided by our very limited functional knowledge about most TFs in many species. There is some indication that differences in lineage-specific occurrence might be of functional importance, e.g. differences in repertoires between vertebrates and invertebrates where a DBD involved in body plan and organogenesis is amplified in vertebrates but not invertebrates and another involved in neural development

is absent from invertebrates all together (Charoensawan *et al.*, 2010b). Nevertheless, especially in eukaryotes where metabolite-binding and signal transduction are largely separated from DNA-binding (see above), a TF is not restricted to regulating a certain set of genes but might instead be coopted into new regulatory roles and these large lineage-specific patterns might merely reflect the mutational mechanisms generating them and the concomitant evolutionary plasticity of gene regulatory networks.

1.2.3 The Regulatory Network

Transcription factors do not function in isolation but rather interact with each other to form complex networks that respond to environmental inputs and determine developmental programs. Many genes are expressed in a coordinated fashion where groups of functionally related genes are induced or repressed at the same time. In two seminal studies in the early 2000s Lee *et al.* (2002) and Harbison *et al.* (2004) have characterised the *S. cerevisiae* transcriptional regulatory network (TRN) through genome-wide location analysis of the binding of the majority TFs encoded in this genome. These studies have set the foundations for the systematic study of the architectural and dynamic properties of the networks resulting in coordinated expression of functionally related genes in eukaryotic genomes. The earlier of the two studies surveyed genome-wide binding patterns of 106 TFs in rich media and uncovered important topological properties of the TRN. The regulatory interactions can be seen as a directed graph where a TF is connected to a target gene (TG) via a directed edge. Analysis of the resulting graph structure revealed that regulatory interactions are often organised into so-called network motifs (subgraphs; Fig. 1.3A) that are overrepresented in the TRN and display distinct properties with regards to modulating regulatory dynamics (Fig. 1.3A; Lee *et al.*, 2002; Milo *et al.*, 2002). The three most commonly occurring motifs are the feed-forward loop (FFL), single-input motif (SIM) and multiple-input motif (MIM). The SIM, where one TF regulates multiple TGs, functions in coregulating a set of related genes under the same signal, whereas the MIM, where multiple TFs regulate one or more TGs, can potentially integrate

various signals for the regulation of downstream genes. The FFL, the most common of the three, relays information from one TF via a second TF and both then regulate the same TG. FFLs display several interesting dynamic properties based on whether the two TFs act as repressors and/or activators, e.g. if both TFs are activators the FFL can serve as a buffer against noisy expression of the first TF so that the TG is only activated after a period of stable induction (Mangan & Alon, 2003). More recently, it was proposed that the "bi-fan" motif, a type of MIM, is the most highly overrepresented motif in both the *Escherichia coli* and *S. cerevisiae* networks (Artzy-Randrup *et al.*, 2004). Like the FFL motif, the bi-fan was shown to be able to produce a large range of dynamical transcriptional outputs depending on the nature and biochemical properties of the regulators involved (Ingram *et al.*, 2006).

Besides these smaller-scale properties, TRNs also display large-scale topological properties that have been analysed using graph-theoretic approaches (Fig. 1.3B). The indegree distribution of a gene, the number of TFs it is regulated by, was found to be exponentially distributed, meaning that most genes are regulated by a relatively small number of TFs (e.g. 93% of genes in *S. cerevisiae* are regulated by one to four TFs; Guelzim *et al.*, 2002). The outdegree of TFs, the number of TGs they regulate in turn was found to follow a power-law distribution, where a small number of TFs regulates a very large number of TGs. These TFs were named regulatory hubs and correspond to the master regulators in the TRN (Guelzim *et al.*, 2002; reviewed in Babu *et al.*, 2004). TRNs thus have a "scale-free" architecture and are often also found to be highly-clustered (many interconnections between groups of nodes). These properties have important implications for the evolvability of such networks and robustness to perturbations. Random deletion of TFs in scale-free networks is more likely to affect regulators with few target genes, lending such networks robustness against random gene inactivation (e.g. Albert *et al.*, 2000). Evolvability in turn is a feature of the clustered nature of regulatory networks. Many genes are coregulated by multiple TFs (Balaji *et al.*, 2006a), allowing for evolutionary flexibility through the usage of alternative regulatory pathways.

Furthermore, TRNs can be organised into hierarchical structures (Fig. 1.3C), with three main hierarchical components. These are the top layer, which contains

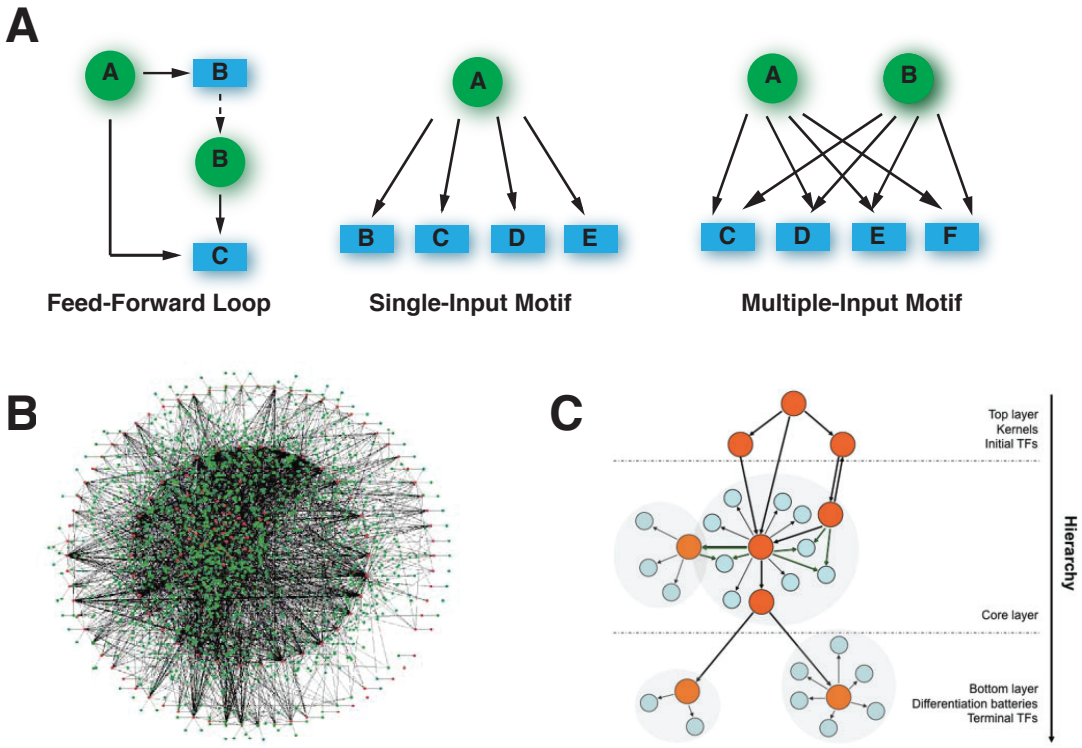


Figure 1.3: Properties of regulatory networks. **A**: Common submotifs found within regulatory networks. **B**: Global structure of the regulatory network. Adapted from (Babu *et al.*, 2004). **C**: Schematic of the hierarchical structure found in regulatory networks. Adapted from (Nowick & Stubbs, 2010).

TFs that are themselves not regulated by other TFs; a densely connected middle layer; and a bottom layer which contains TFs that do not regulate other TFs (e.g. Bhardwaj *et al.*, 2010b; Cosentino Lagomarsino *et al.*, 2007; Jothi *et al.*, 2009; Ma *et al.*, 2004; Yu & Gerstein, 2006). These layers differ in their topological, functional and evolutionary properties. Whereas the top and core layers are generally involved in the regulation of many biological processes and contain TFs with large outdegrees, the bottom layer represents a specialist layer containing TFs that perform largely stand-alone functions such as amino acid metabolism or carbohydrate catabolism (Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009; Yu & Gerstein, 2006). Moreover, the core layer is distinct from the top layer in that it forms a tightly connected component, is highly enriched for regulatory hubs and shows the greatest level of collaborative regulation between TFs (Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009). Together, the core layer TFs regulate 87% of characterised target genes (compared to 35% and 25% in the top and bottom layers respectively) and as such is the main processor of regulatory information (Jothi *et al.*, 2009). Indeed, there is almost no collaboration between TFs in the bottom layer (Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009). Analysis of motifs found within and across different layers suggest the presence of a feed-forward structure, in which top-layer TFs regulate core- and bottom-layer TFs and core-layer TFs regulate bottom layer TFs (Jothi *et al.*, 2009; Ma *et al.*, 2004). Figure 1.4 shows the percentages of the two most commonly occurring motifs involving regulators from each layer in the *S. cerevisiae*, *E .coli* and *Mycobacterium tuberculosis* TRNs. Whereas most MIMs are formed between core-layer TFs regulating non-TF target genes, FFLs often involve a top- and core-layer TF, relaying information from top to bottom (Bhardwaj *et al.*, 2010b).

Differences between the dynamic properties of TFs also reflect possible mechanisms for information processing across the network, e.g. top-level TFs are generally much higher in protein abundance, most likely due to slow degradation, whereas both core and bottom-level TFs are more rapidly degraded (Jothi *et al.*, 2009). Together with the observation of frequent occurence of FFLs between top- and core layer TFs, this suggests the presence of a constantly "armed" system that can be rapidly modulated through availability of "second in command" TFs. Interestingly, protein abundance of top-level TFs was also found to be relatively

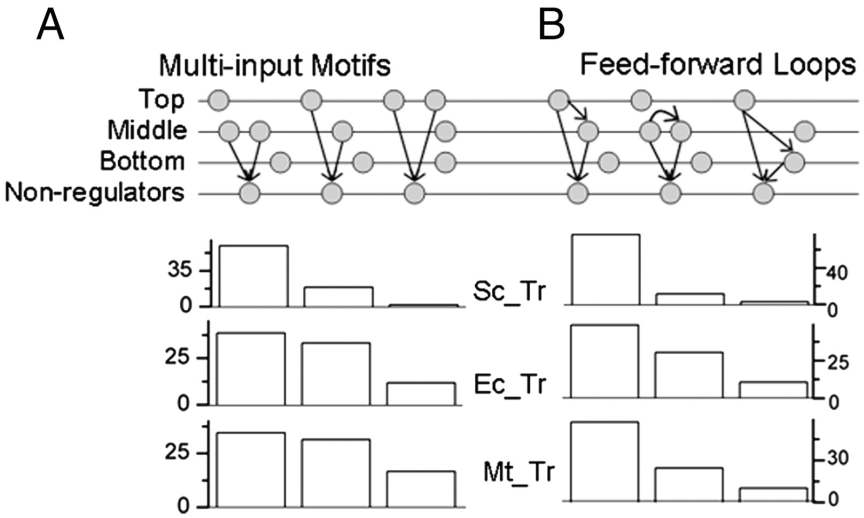


Figure 1.4: The distribution of network motifs across different hierarchical layers of the *S. cerevisiae* (Sc), *E. coli* (Ec) and *Mycobacterium tuberculosis* (Mt) based on the study by Bhardwaj *et al.* (2010b). Each bar corresponds to the percentage of network motifs that correspond to the structure shown directly above

noisy and it was proposed that this might aid differential responses to the same stimuli between cells and thus enhance adaptability within a population (Jothi *et al.*, 2009).

Top and core layer TFs were furthermore found to be more evolutionary conserved than bottom layer TFs (Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009), although paradoxically Yu & Gerstein (2006), while they also found top-level regulators to be more influential, found that bottom level TFs were more likely to be lethal when deleted. It is possible that this reflects both a low degree of redundancy in the bottom level due to the often stand-alone regulatory role of the TFs as well as potentially greater evolutionary plasticity through fast turnover and rewiring of these TFs seeing that they regulate few target genes and show low levels of coregulation and as such a possibly more amenable to "replacement" by other TFs.

TRNs thus are heavily substructured entities whose complex patterns of coregulation and dynamic properties can generate both highly-coordinated, rapid and specific regulatory output. Indeed, coregulation is wide-spread (the average num-

ber of TFs a TG is regulated by in yeast is 2.9 [Balaji *et al.*, 2006a]) and the tendency for TFs to coregulate TGs increases with organismal complexity (Bhardwaj *et al.*, 2010b). Moreover, the yeast coregulatory network shows a more distributed architecture than the scale-free transcriptional regulatory network and as such is more robust to the targeted removal of regulatory hubs, explaining why only few TF hubs are essential genes (Balaji *et al.*, 2006b). Extensive coregulation thus appears to facilitate evolvability through the presence of alternative pathways to regulate a given target gene.

1.3 How do Complex Phenotypes Evolve? — The *Cis* versus *Trans* Debate

The genomic changes underlying the evolution of phenotypic diversity and particularly form have been the subject of much recent debate. While the importance of changes in gene expression has been appreciated early on (e.g. King & Wilson, 1975; see above), the availability of fully sequenced genomes, transcriptomes and genome-wide TF binding data has raised new questions and challenges of exactly how genotypic variation gives rise to phenotypic variation. The observation that homologous developmental regulators that have been evolving independently for hundreds of millions of years of evolution and are able to substitute for each other functionally in distantly related species has given rise to a school of evolution of development ("evo-devo") whose strongest proponents have placed heavy if not exclusive emphasis on changes in *cis*-regulatory sequences for the generation of morphological variation (e.g. Carroll, 2005; Prud'homme *et al.*, 2007; Stern, 2000; Wray, 2007; Wray *et al.*, 2003). Pax6, for example, is a master regulator of eye development and can perform this function when substituted between humans, mice and *Drosophila* (Halder *et al.*, 1995; Onuma *et al.*, 2002). Further support for the "*cis* hypothesis" came from a handful of compelling examples of regulatory rewiring of a conserved "regulatory toolkit" (e.g. in producing *Drosophila* wing pigmentation patterns: Gompel *et al.*, 2005; Prud'homme *et al.*, 2006 or morphology of the pelvis in stickleback fish: Chan *et al.*, 2010; Shapiro *et al.*, 2004).

In his book "Endless Forms Most Beautiful", Sean Carroll states: "regulatory evolution must be the major contributor to the evolution of form".

The arguments that are being put forward in favour of such hypotheses generally encompass three lines of thought. First, changes in *cis*-regulatory sequences are likely to be less pleiotropic than changes in protein-coding sequences (Stern, 2000; Wray, 2007; Wray *et al.*, 2003). This dwells on the fact that whereas most changes to protein-coding genes are likely to be deleterious, especially in the case of TFs where mutations might affect tens if not hundreds of downstream targets in a number of different conditions, *cis*-regulatory sequences are modular and changes in a TFBS are likely to affect only the gene in question and only at a particular point in space and time, and have thus less pleiotropic constraint. Second, changes in *cis*-regulatory DNA are more likely to be codominant and are therefore more likely to be exposed to selection than mutations in protein-coding genes which are likely to be recessive (Wray, 2007; Wray *et al.*, 2003). Third, a distinction is made between the evolution of physiology and the evolution of morphology. While evo-devo proponents generally accept the importance of protein-coding evolution for adaptive changes in physiological traits, evolution of morphology is proposed to be exclusively due to *cis*-regulatory changes (e.g. Carroll, 2005). In a critical review of these arguments, Hoekstra & Coyne (2007) argue convincingly that in fact this distinction lacks a true biological justification and that it is hard to conceive why the deleterious effect of protein-coding mutations would be less strong for physiological and biochemical traits than for anatomical ones. Furthermore, they argue that while there is ample evidence for adaptive evolution in protein-coding genes, there is lack of evidence for adaptive changes in *cis*-regulatory regions. Although the adaptive potential of *cis*-regulatory mutations is clearly more difficult to determine, often associations between gene expression level and phenotype have not been linked to causal mutations (there is no *a priori* reason to believe that changes in gene expression are the result of *cis*-regulatory mutations). Equivalently, fast-evolving non-coding regions have rarely been linked to a phenotypic outcome and their evolutionary significance remained questionable (Hoekstra & Coyne, 2007).

Recent technological advances have made it possible to both link changes in TFBSs to changes in expression level (e.g. Chen *et al.*, 2010) and determine the

adaptive significance of changes in gene expression level (e.g. Emerson *et al.*, 2010; Fraser *et al.*, 2010) which should allow for more thorough empirical studies of "the locus of evolution". Indeed, the question of whether regulatory divergence resides in *cis* or *trans* has been the subject of a plethora of studies, primarily in yeast (e.g. Brem *et al.*, 2002; Bullard *et al.*, 2010; Chang *et al.*, 2008; Emerson *et al.*, 2010; Sung *et al.*, 2009; Tirosh *et al.*, 2009; Wang *et al.*, 2007; Yvert *et al.*, 2003) but also in *Drosophila* (e.g. Osada *et al.*, 2006; Wittkopp *et al.*, 2004) and mammals (Wilson *et al.*, 2008b), at both the intra- and interspecies level. Commonly used approaches to addressing such questions experimentally are either the mapping of quantitative trait loci (QTLs) that are associated with gene expression variation in crosses between different strains or species (e.g. Brem *et al.*, 2002; Yvert *et al.*, 2003) or, more recently, measurement of allele-specific expression (ASE) in a F_1 hybrid of the two parental strains or species and comparison of relative expression patterns to those observed in the parents (e.g. Bullard *et al.*, 2010; Chang *et al.*, 2008; Emerson *et al.*, 2010; Sung *et al.*, 2009; Tirosh *et al.*, 2009; Wang *et al.*, 2007. This comparison works based on the hypothesis that in the hybrid the *trans* environment is the same for both parental alleles. If both alleles are expressed at equal levels in the F_1 hybrid, the relevant *cis*-regulatory elements are considered to be conserved and differences in expression level between the parental strains are solely due to *trans* effects. Similarly, if the expression of alleles in the hybrid mirrors that of the parental strains, differences in expression are postulated to be solely due to *cis* effects. A mixture of both patterns points to more complex histories of inheritance that can neither be explained by *cis* or *trans* effects alone.

Generally, these studies revealed a large number of loci (e.g. around a quarter of protein-coding genes in yeast) that showed divergence in expression levels between strains (e.g. Brem *et al.*, 2002; Yvert *et al.*, 2003) and species (e.g. Bullard *et al.*, 2010; Tirosh *et al.*, 2009). Within-species crosses mainly implicated *trans*-acting factors as the prevalent source for expression divergence (Brem *et al.*, 2002; Chang *et al.*, 2008; Emerson *et al.*, 2010; Sung *et al.*, 2009; Wang *et al.*, 2007; Yvert *et al.*, 2003), whereas between-species crosses found greater importance of *cis*-regulatory variation (Osada *et al.*, 2006; Tirosh *et al.*, 2009; Wilson *et al.*,

2008b; Wittkopp *et al.*, 2004, 2008). In a comparison between inter- and intra-specific ASE data in yeast, Emerson *et al.* (2010) furthermore found that there was an excess of *cis* polymorphism between species when compared to the level of polymorphism expected based on within-species comparisons, especially when compared to the proportions of *trans*-acting variation, indicating that *cis*-acting variation might be the subject of directional selection. While these studies are useful for determining the direct influence of mutations in the *cis*-regulatory region on the expression divergence of that gene, they conceptually fail to address the questions posed by the initial argument, i.e. an associated *trans* change does not necessarily imply that the underlying mutation(s) is (are) in a protein-coding gene. Similarly, a *cis*-only effect could mask the combined action of a co-evolved change in protein-protein interaction (PPI) potential and creation of a new binding site, where, all else being equal, neither the PPI nor the binding site alone are sufficient for the recruitment of a novel TF. Indeed, especially the later studies (e.g. Emerson *et al.*, 2010; Sung *et al.*, 2009; Tirosh *et al.*, 2009) and studies on specific regulatory modules (e.g. Chang *et al.*, 2008) have generally revealed complex patterns of *cis*- and *trans*-dependence with only a minority of expression divergence explained solely due to *cis* or *trans* effects. Tirosh *et al.* (2009) investigated the relative contributions of *cis* and *trans* effects to expression divergence in four different conditions using an ASE approach. Interestingly, they discovered that, while *cis* divergence between species is mainly independent of environmental conditions (effects persisted in all four conditions tested), *trans* effects were largely condition-dependent, somewhat contradicting the first argument outlined above that argues for greater modularity and condition-specific impact of changes in *cis*-regulatory modules (although again it needs to be emphasised that the nature of *trans* changes here remained undetermined). Recent approaches have attempted more careful dissection of the relationships between genetic and expression divergence, e.g. through modelling of the underlying regulatory systems (Ye *et al.*, 2009) or integration of a variety of other sources of data (Choi & Kim, 2008) which should allow for the refinement of the *cis* versus *trans* hypotheses.

It is thus clear that neither source of variation alone is able to explain the large amounts of expression divergence observed in the species studied here and

that components other than *cis*-regulatory modules deserve our attention when studying the evolution of gene regulation. Indeed, in a thorough evaluation of the arguments advocated by Carroll, Wray and others (Carroll, 2005; Prud'homme *et al.*, 2007; Stern, 2000; Wray, 2007; Wray *et al.*, 2003), using the available evidence at the time Lynch & Wagner (2008) have made a strong case for the importance of TF divergence in the evolution of gene regulatory networks which I will outline below. First and foremost, there is overwhelming evidence for adaptive evolution in TFs across all major domains of life (Lynch & Wagner (2008) cite 17 studies in support of this). Secondly, studies very similar to those that initially established the notion that TFs are "evolutionarily rigid" equally highlighted examples where developmental regulators failed to complement for each other in divergent clades as is the case for the heart developmental regulator *tinman* (reviewed in Lynch & Wagner, 2008). Furthermore, genome-wide ChIP-chip studies such as the one by Borneman *et al.* (2007) and Tuch *et al.* (2008a), which will be discussed in more detail in the next section, have shown that the presence of a conserved TFBS does not predict TF binding and have implicated the importance of other factors such as PPIs, post-translational modification and local chromatin organisation in determining whether or not a binding site is occupied (see Thompson & Regev, 2009; Wohlbach *et al.*, 2009 and Pan *et al.*, 2009 and for review). Based on these observations and after a model outlined by Tuch *et al.* (2008b) (Figure 1.5A), Lynch & Wagner (2008) have proposed a model of TF recruitment via the evolution of novel PPIs which is shown in Figure 1.5. They hypothesise that a scenario where a novel interaction evolves prior to a change in *cis*-regulatory elements (Fig. 1.5C) will have much greater fitness effects, seeing that it will be immediately dominant and affecting all downstream targets of the regulator simultaneously. In the case of the evolution of a novel *cis*-regulatory element at a single gene (Fig. 1.5B), fitness effects would be very weak initially seeing that TF recruitment would be limited to just this one allele. As such the first scenario is arguably much more likely.

Moreover, the authors provide several lines of argument against the notion that mutations in TFs necessarily have high pleiotropic effects. First, the role of a TF is context-dependent and is determined by the combinatorial interactions with other regulators that are expressed at the same time and in the same tissue

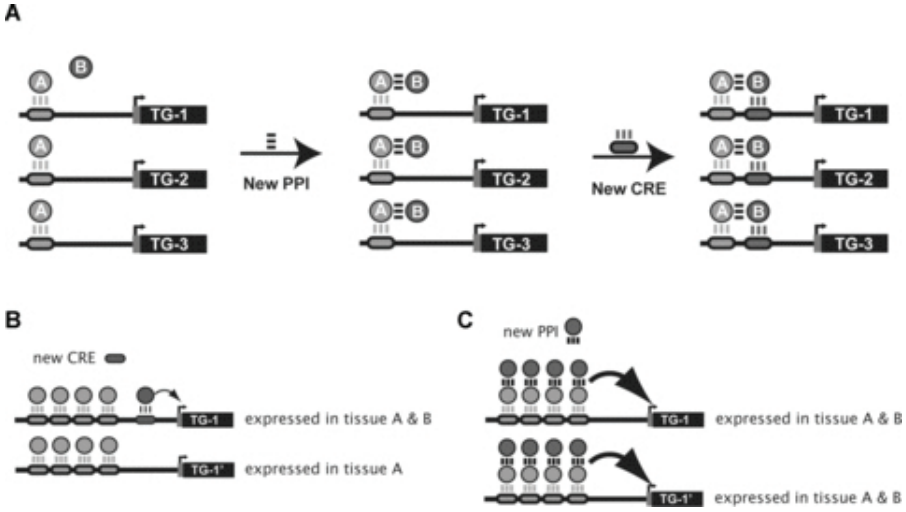


Figure 1.5: **A**: A proposed model for the regulatory rewiring of regulons (after Tuch *et al.*, 2008b). **B**: Rewiring scenario through the evolution of a novel TFBS. **C**: Rewiring scenario where a novel PPI arises before changes in *cis*-regulatory regions. Figure taken from Lynch & Wagner (2008).

(e.g. Balaji *et al.*, 2006a; Luscombe *et al.*, 2004; see above). As such, gain or loss of some functional properties of TFs (e.g. PPIs) are likely only to be important in a certain spatio-temporal context, providing a large number of combinatorial possibilities to be explored without wide-spread pleiotropic effects. Indeed, the study by Tirosh *et al.* (2009), who found *trans* effects to be predominantly condition-specific, provides indirect empirical evidence in support of such claims. Gene duplication is also an important mechanism for overcoming pleiotropic constraints (Hoekstra & Coyne, 2007; Lynch & Wagner, 2008). Duplicate copies of a TF will be redundant at first, removing strong selective constraints and allowing for sub- or neofunctionalisation of TFs through gain or loss of domains, interaction partners or changes in DNA-binding specificity. The fact that TF repertoires display such strong lineage-specific patterns of amplifications of DBD-types and domain architectures (Charoensawan *et al.*, 2010b; see above) suggests that gene duplication is indeed a major evolutionary force driving the divergence of TF repertoires. Lastly, pleiotropic effects can be overcome through alternative splicing and use of different TF isoforms in different tissues, e.g. the full transcript of the CF2 TF in *D. melanogaster* has at least three splice-isoforms that con-

tain three, six or seven repeats of the C_2H_2 zinc finger and although it is not clear whether the three-finger form binds DNA, the six- and seven-finger forms have different DNA-binding specificities (Hsu *et al.*, 1992). Furthermore, different isoforms are expressed in different developmental processes and tissues, e.g. either predominantly in female development (six-finger form), male development (seven-finger form) or testis-only (three-finger form).

Besides these regulatory and combinatorial properties, there are other structural features of TFs that underline their evolutionary potential, e.g the presence of short linear motifs (SLMs) or simple sequence repeats. SLMs are motifs that span between three and 10 amino acid residues and can mediate interactions with other proteins (Neduva & Russell, 2005). They can often arise through a single point mutation and as such are highly evolutionary plastic. In fact, over 85% of SLMs are found in intrinsically disordered regions of proteins reinforcing their evolutionary plasticity due to the lack of strong structural mutational constraints in these regions. Indeed intrinsically disordered regions are not only frequently found in TFs but also other components of the transcriptional machinery providing further evidence for an important role of SLMs (reviewed in Fuxreiter *et al.*, 2008). Overall this thus paints a rather dynamic picture of TF evolution, very far away from the static "toolkit" view proposed the Carroll (2005) and others.

1.4 The Evolution of Transcriptional Regulation

Our understanding of the evolution of transcriptional regulatory networks, very much like the regulatory processes themselves, spreads across several layers of complexity that we are slowly beginning to connect. These range from the divergence of gene expression patterns, the output of the regulatory networks, to the underlying genomic changes generating such divergence. Mutations affecting gene expression can arise in *cis*-regulatory regions, affecting transcriptional regulation locally, for example through generation or loss of transcription factor binding sites (TFBS), or in *trans*-acting factors, e.g. transcription factors, chromatin modifiers or signal transducing proteins thereby generating divergence at higher hierarchical levels that is able to influence transcriptional output globally and in a variety of regulatory processes. In the following sections, I will review our

current knowledge of the evolution of transcriptional regulation with a strong focus on the ascomycetous yeasts, those being the species that provided the largest contribution to our understanding of regulatory evolution as well as the focus of my study.

1.4.1 The *Cis*-regulatory Enigma

Coexpression between functionally related genes often shows strong conservation between species (e.g. Bergmann *et al.*, 2004; Chan *et al.*, 2009; Liao & Zhang, 2006; Stuart *et al.*, 2003; reviewed in Weirauch & Hughes, 2010). In contrast, both computational studies of *cis*-regulatory element conservation (e.g. Doniger & Fay, 2007; Gasch *et al.*, 2004; Raijman *et al.*, 2008; Tanay *et al.*, 2005) and experimental determination of TF binding (e.g. Borneman *et al.*, 2007; Lavoie *et al.*, 2010; Moses *et al.*, 2006; Odom *et al.*, 2007; Schmidt *et al.*, 2010; Zheng *et al.*, 2010) have revealed extensive variation in binding site turnover and occupancy between populations of *S. cerevisiae* (Zheng *et al.*, 2010) and between closely (Borneman *et al.*, 2007; Doniger & Fay, 2007; Moses *et al.*, 2006; Odom *et al.*, 2007; Raijman *et al.*, 2008) and more distantly (Gasch *et al.*, 2004; Lavoie *et al.*, 2010; Schmidt *et al.*, 2010; Tanay *et al.*, 2005) related species. Even when there is some degree of conservation of enriched motifs across functionally related sets of genes, it is rarely the primary promoter sequence that is conserved between species. In a study of ascomycetous fungi, Gasch *et al.* (2004) found that out of the 42 TF binding motifs that were overrepresented in coregulated genes in *S. cerevisiae*, the majority were also overrepresented in promoters of orthologous genes within the *Saccharomyces sensu stricto* species, while three quarters were conserved across the post-whole genome duplication (WGD) species and a third up to *Candida albicans* (see Fig. 1.9 for a phylogeny of these species). Despite conserved enrichment, spacing and orientation however, TFBSs were often not found in orthologous positions and are likely to have arisen independently. It is also worth noting that the enrichments observed here represent about a quarter of known TFs in *S. cerevisiae* and as such are likely to be a survey of only a few master regulatory pathways. In light of this, the findings that only a third of the regulation of such important pathways is conserved between *S. cerevisiae*

and *C. albicans* indicates the prevalence of large-scale rewiring even of important transcription programs.

This extent of evolutionary plasticity in promoter and enhancer sequences is not restricted to fungi but has been observed across a wide variety of organisms, including insects (Dermitzakis *et al.*, 2003), nematodes (Maduro & Pilgrim, 1996) and vertebrates (Fisher *et al.*, 2006). As such it appears to be a general feature of regulatory evolution. Studies of orthologous regions in promoters of closely related yeasts showed that TFBSs are frequently lost and gained and generally under very weak selective constraint (Doniger & Fay, 2007; Raijman *et al.*, 2008). Furthermore, promoters often tend to be enriched for multiple copies of a motif and the selective constraint experienced by a TFBS appears to be correlated with the number of redundant alternative sites in its vicinity (Raijman *et al.*, 2008). Indeed, 57% of the species-specific loss events inferred by Doniger & Fay (2007) could be explained by turnover through the gain of a compensatory binding site nearby.

Similarly, experimental TF binding data indicate large-scale turnover of promoter occupancy. The mapping of the two pseudohyphal growth regulators Tec1 and Ste12 in three *Saccharomyces sensu stricto* species revealed that while cooperative binding between the two regulators was strongly conserved, only 20% of binding events were conserved across all three species (Borneman *et al.*, 2007). Studies of liver-specific TFs involved in a regulatory program that is highly conserved in vertebrates yielded very similar results, with as much as 89% divergence in promoter occupancy for some TFs between human and mouse (Odom *et al.*, 2007). These differences were even more pronounced when the analysis was extended to five vertebrate species that shared their last common ancestor about 300 million years ago (mya; approximately the same time since the yeast species in my study shared their last common ancestor; see below), where binding events of one of the TFs, that were conserved across all five species amounted to 0.3% of the total binding events observed in human for that TF (Schmidt *et al.*, 2010). Moreover, although loss and gain of binding events could sometimes be related to a loss or gain of a TFBS in the corresponding region, there were numerous instances in all three studies where either binding was conserved but sequence was not or *vice versa*, implicating the importance of other factors such as local

chromatin structure or interaction (cooperative or competitive) with other TFs (Borneman *et al.*, 2007; Odom *et al.*, 2007; Schmidt *et al.*, 2010). High levels of divergent promoter occupancy were not only observed between species, but also more recently within species where genomic profiling of Ste12 in two *S. cerevisiae* parental strains and 43 segregants revealed frequent divergent binding, predominantly due to *cis* but also due to *trans* effects (Zheng *et al.*, 2010).

There often remains a "knowledge gap" between promoter occupancy by a TF and changes in gene expression (e.g. Borneman *et al.*, 2007; Gao *et al.*, 2004). Borneman *et al.* (2007) were able to relate about 20% of observed binding events to changes in gene expression and, more importantly, found no enrichment among those for events that were conserved between all three species surveyed. Extensive evolutionary plasticity of *cis*-regulatory regions is thus found at various levels, from frequently turning over promoter regions to sometimes sequence-unrelated gains or losses of binding events that are mediated or buffered by other factors, providing both a robustness against mutations as well as ample potential for evolutionary novelty through the compensatory and highly-context dependent impact of mutations. At the same time this makes inferences about the nature of adaptive changes a challenging task.

1.4.2 Evolution of Regulatory Networks — Lessons from Fungi

Due to being easily manipulable and a tractable model system, the ascomycetous fungi have contributed vastly to our understanding of regulatory evolution in eukaryotes. Studies investigating gene expression divergence, the evolution of *cis*-regulatory regions and comparative TF binding have begun to highlight the mechanisms underlying changes in gene regulation between species. Figure 1.6 shows the three major scenarios of regulatory evolution that have become apparent from those studies which I will review below, based on examples from ascomycetous fungi.

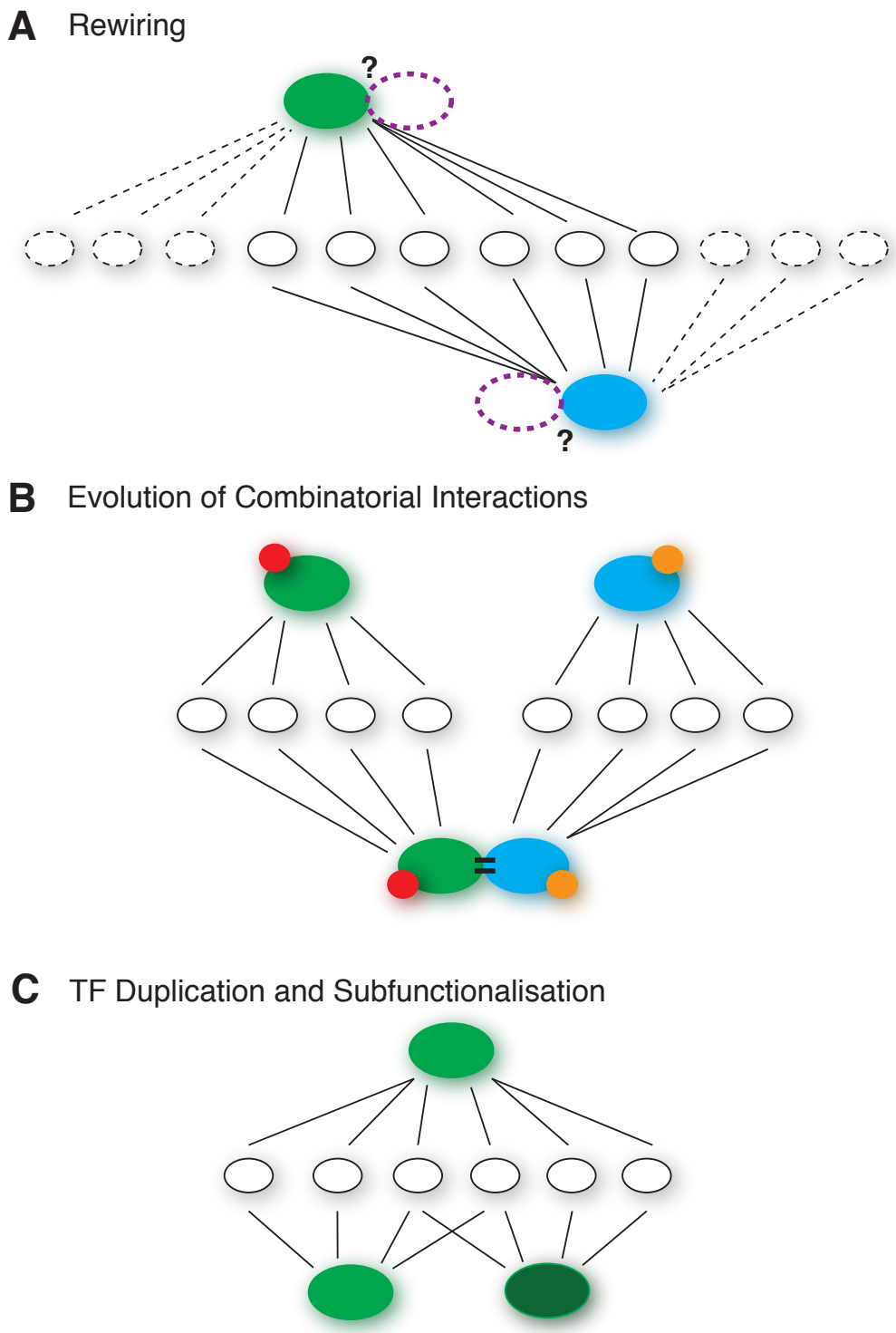


Figure 1.6: The three main mechanisms resulting in changes in gene regulation. **A**: Large-scale rewiring, **B**: Evolution of novel interactions, **C**: Gene duplication.

1.4.2.1 Large-scale Rewiring

Large-scale rewiring refers to a scenario where, despite conservation of TFs and their DNA-binding specificities, the control of entire regulons has been handed over from one or more TFs to one or more other TFs through turnover of *cis*-regulatory elements (Figure 1.6A). During rewiring, new target genes can be incorporated for coregulation or existing members can be lost from the regulon. Regulatory interactions can be gained and lost through point mutations in TFBSs themselves or through insertions or deletions nearby that alter the relative spacing of motifs with respect to motifs of other competitively or cooperatively binding TFs or the TSS. Transposable elements (TEs) also provide a potent mechanism for the creation of new regulatory modules (reviewed in Babu, 2010). Indeed, many TEs carry TFBSs and especially those carrying multiple binding sites for different TFs are frequently involved in the creation of new functional binding events when inserted in the vicinity of gene promoters, providing a mechanism for rapid and large-scale regulatory rewiring through transposition (Xie *et al.*, 2010). Other mechanisms by which new regulatory interactions can arise are gene duplication of regulators (see below) or target genes, doubling the number incoming (in the case of TF duplication) or outgoing (in the case of target gene duplication) interactions of target genes and TFs, respectively.

Several recent studies have detected such large-scale rewiring of important metabolic pathways in fungi (e.g. Askew *et al.*, 2009; Gasch *et al.*, 2004; Hogues *et al.*, 2008; Ihmels *et al.*, 2005; Lavoie *et al.*, 2010; Martchenko *et al.*, 2007; Tanay *et al.*, 2005). The most prominent and well-studied example of this is the rewiring of the ribosomal proteins (RPs) on the lineage including *S. cerevisiae* and its pre-WGD relatives (see Fig. 1.7). Over 90% of transcriptional activity in the cell is dedicated to ribosome biogenesis and the tightly coordinated expression of stochiometric quantities of over 70 RPs making up this macromolecular complex is crucial for its function (Deutschbauer *et al.*, 2005). In *S. cerevisiae* the main regulator of RPs is Rap1, a TF also involved in glycolytic gene regulation and silencing at the mating type loci. Rap1, together with its essential cofactors Hmo1, Ifh1 and Fhl1, regulates RP expression in a stress- and nutrient-dependent manner (Kasahara *et al.*, 2007; Martin *et al.*, 2004; Rudra *et al.*, 2005). Early

motif-enrichment analyses in the *Ascomycota* showed that the Rap1 motif was only enriched upstream of RP genes in the *Saccharomyces* species and close relatives, whereas other motifs were present in the remaining ascomycetous fungi (Gasch *et al.*, 2004; Tanay *et al.*, 2005). Recent studies have corroborated these results both informatically and biochemically and have identified the TFs Tbf1 and Cbf1 to be the main regulators of RP genes in *C. albicans* and likely most other ascomycetes too (Hogues *et al.*, 2008; Lavoie *et al.*, 2010).

Interestingly, while Rap1, Hmo1, Ifh1, Fhl1, Tbf1 and Cbf1 are conserved across the *Ascomycota*, the main regulators Rap1, Hmo1 and Tbf1 in particular have diverged significantly in function. Besides large changes in the number of bound promoter regions (e.g. 10-fold enrichment of Rap1 binding in *S. cerevisiae* compared to *C. albicans*), neither of these three TFs maintained a significant proportion of their target genes (Lavoie *et al.*, 2010). Tbf1 for example was found to be exclusively regulating RPs in *C. albicans* but has drifted from this specialist role to a generalist role in *S. cerevisiae* where it was detected at some RP genes but mainly telomeric and subtelomeric regions in addition to approximately 300 protein coding loci. Similarly, Rap1 is mainly bound near the telomeres in *C. albicans* whereas in *S. cerevisiae* it is known to regulate glycolytic genes and mating type silencing in addition to RP genes (see above). As such this rewiring event represents a complete handover of the, presumably heavily selectively constrained, RP regulon from a Tbf1-Cb1-based system to a Rap1-Hmo1-based system on the lineage leading to the *Saccharomyces* species, despite the conservation of their main regulators. Also of interest is that the other three regulators which provide condition-specific control have retained a significant part of their functionality, suggesting conservation of the condition-specific response. Intriguingly, Cbf1, which has a conserved role in the regulation of sulfur metabolism and respiratory chain genes but not RPs between the two species also regulates glycolytic enzymes in *C. albicans*. Because *C. albicans* has a preferentially aerobic lifestyle, this coordination between RPs, respiratory chain genes and glycolytic enzymes seems favourable and suggests that the regulator turnover in RP genes might have aided the decoupling of the regulation of respiratory chain genes and glycolytic enzymes and in turn their regulation from RPs in the *Saccharomyces* species which thrive under anaerobic, high-glucose conditions and repress respiratory

metabolism in exponential growth phases (see below). Indeed the glycolytic enzymes were "reunited" with the control of RP genes through combined regulation by Rap1 in the *Saccharomyces* species (Lavoie *et al.*, 2010). Similarly, the Rapid Growth Element (RGE), which is found across RPs in both species, has been lost from mitochondrial RPs (MRPs) in *S. cerevisiae* and relatives, which led to a decoupling of the expression of MRPs and RPs (Ihmels *et al.*, 2005). MRPs in *S. cerevisiae* are coupled with other aerobic and stress-responsive genes instead (Gasch *et al.*, 2000; Ihmels *et al.*, 2005).

Besides the RP regulon, there are numerous other known examples of rewiring in the *Ascomycota* (see Fig. 1.7), including carbohydrate metabolism (Askew *et al.*, 2009; Brown *et al.*, 2009; Hittinger *et al.*, 2004; Martchenko *et al.*, 2007), fatty acid catabolism and phospholipid biosynthesis (Hynes *et al.*, 2006) and amino acid biosynthesis (reviewed in Lavoie *et al.*, 2009).

Together these examples show that the regulation of some essential metabolic activities has been completely rewired within the 300 million years that separate *S. cerevisiae* and *C. albicans* and demonstrate the great evolutionary plasticity inherent in regulatory networks. Whether or not these rewirings represent instances of adaptive evolution or are mainly due to drift and which mechanisms allowed for such large-scale coordinated change is still being debated. It is clear however that we will be likely to be able to extend this list of examples in the future and that TF evolution has played an important role in the establishment of these new regulatory mechanisms (see below).

1.4.2.2 Evolution of Combinatorial Interactions

Apart from the gain and loss of *cis*-regulatory elements, evolution of new combinatorial interaction through novel PPIs between TFs can generate novel regulatory interactions and integration of regulons to be coexpressed under the same signals (Fig. 1.6B). It has been proposed that the evolution of a novel PPI is a likely pathway for subsequent rewiring through turnover in TFBSs where a new TF is recruited to all promoters in the regulon immediately and can be stabilised through increase in either the strength of protein-protein or protein-DNA

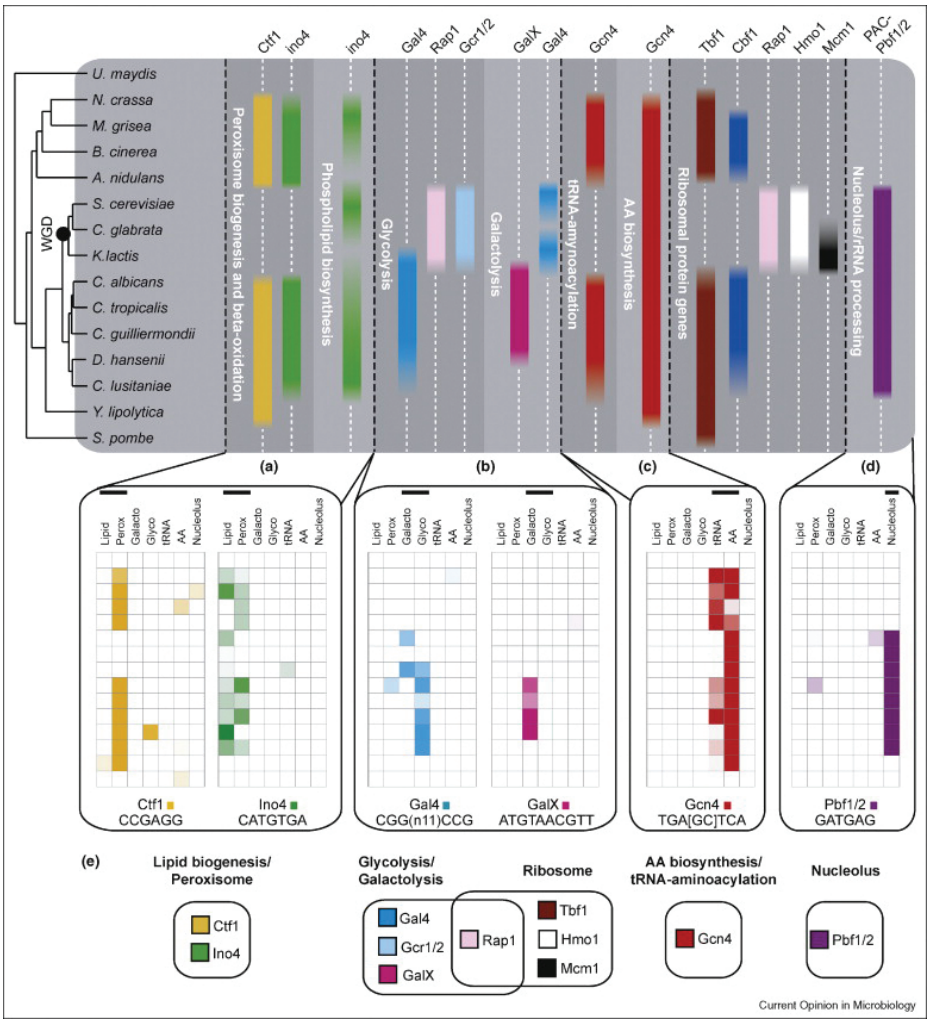


Figure 1.7: Known cases of regulatory rewiring in ascomycetous fungi. Figure taken from Lavoie *et al.* (2009).

interactions (Lynch & Wagner, 2008; Tuch *et al.*, 2008b; see above). The general regulator Mcm1 nicely illustrates this. Mcm1 is a ubiquitous TF that is itself not generally regulated but exerts its regulatory role through interaction with specific cofactors (reviewed in Tuch *et al.*, 2008a). Depending on whether it cooperates with Mat**a**2 or Matα1, Fkh2 and Ndd1, Yox1 or Arg80/Arg81, it regulates the expression of **a**-specific genes, α-specific genes, G2/M or M/G1 cell cycle transitions or arginine metabolic genes respectively. Comparative binding data of Mcm1 in *S. cerevisiae*, *Kluyveromyces lactis* and *C. albicans* suggested extensive gain and loss of target genes between those three species (Tuch *et al.*, 2008a). While the authors found most characterised Mcm1-cofactor interactions to be present to some extent across all three species, much of the turnover of binding events between species could be explained by the evolution of entirely new Mcm1-cofactor regulons.

The most interesting of those is probably the discovered interaction of Mcm1 and Rap1 at RP promoters in *K. lactis*. Mcm1 was detected to be bound to 70 out of the 101 cytosolic ribosome gene promoters which were also found to be enriched for a well-spaced Rap1 motif, suggesting coregulation of RPs by those TFs in *K. lactis* and a possible intermediate that might have facilitated the takeover of the RP regulon from Tbf1 to Rap1 (see above; Tuch *et al.*, 2008a). Subsequent motif analyses furthermore revealed that within its close relatives, the Mcm1-Rap1 promoter architecture is restricted to *K. lactis* but also appeared in other species (e.g. the post-WGD species *Candida glabrata* and the distantly-related *Yarrowia lipolytica*) which do not cluster phylogenetically. This is most-parsimoniously explained by a striking amount of convergent evolution (Tuch *et al.*, 2008a).

Another well-characterised example is the cofactor interaction with Mat**a**2 and Matα2 which is thought to have facilitated the switch from positive control to negative control of **a**-specific genes (**a**asgs) in the *Saccharomycotina* (Figure 1.8; Tsong *et al.*, 2003, 2006). In the haploid stage, *S. cerevisiae* and its relatives occur in one of two mating types, **a** and α that are able to form **a**/α diploids through reciprocal mating (reviewed in Soll *et al.*, 2009). Each cell-type expresses distinct sets of genes (**a**sgs and αsgs) that are required for mating with the other type and are turned off in **a**/α cells. In *C. albicans* and other ascomycetes, **a**sgs

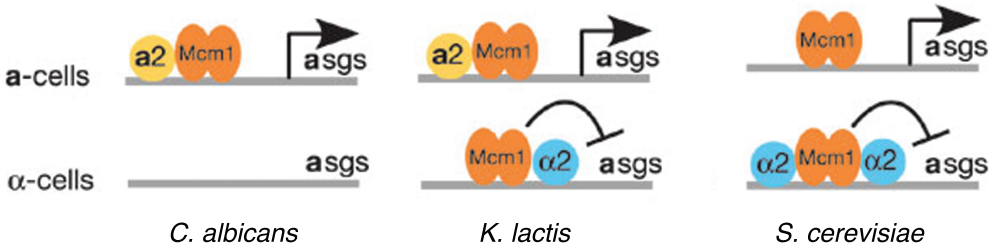


Figure 1.8: Regulatory rewiring from positive to negative regulation of **a**-specific genes between *S. cerevisiae* and *C. albicans*. Figure adapted from Tsong *et al.* (2006).

are turned off by default and in **a** cells induced by the TF **a**2. In *S. cerevisiae* in turn, **a**sgs are on by default but instead repressed by α2 in α cells, effectively leading to the same result (**a**-specific expression of **a**sgs), but using different regulatory logic (Tsong *et al.*, 2003). Investigation of the promoter regions of *S. cerevisiae* and *C. albicans* **a**sgs revealed well-spaced binding motifs for Mcm1, flanked by two α2 sites in *S. cerevisiae* and a single **a**2 in *C. albicans*, suggesting that Mcm1 might have facilitated this transition. Indeed, analysis of *K. lactis* **a**sg promoters revealed a tripartite structure as in *S. cerevisiae* but with one flanking **a**2 site and one that is similar to both the **a**2 and α2 motif (Tsong *et al.*, 2006). Further evidence for the potentially facilitating role of the evolution of an α2-Mcm1 interaction came from the study of the α2 interaction region which was found to be highly conserved in *Saccharomyces* and its close relatives and moderately conserved in *K. lactis* and relatives, but showed no conservation in the *Candida* species or beyond (Tsong *et al.*, 2006).

Furthermore, a novel Mcm1-cofactor interaction was discovered in *C. albicans* and its close relative *Candida dubliensis*. Here Mcm1 was found to bind to non-canonical sites in cooperation with Wor1, the main regulator of white-opaque switching in *C. albicans* (reviewed in Lohse & Johnson, 2009). The white and opaque states are two heritable cell types that can be assumed by *C. albicans* and display distinct expression programs that have a key role in the mating cycle and are also thought to be of importance for pathogenicity (Tuch *et al.*, 2010).

Moreover, the ability to switch from the white to the opaque state has been shown to be closely linked to cell cycle progression (reviewed in Lohse & Johnson, 2009), suggesting that the integration with Mcm1-regulation, whose core function is in cell cycle regulation, might have implications for the switching mechanism (Tuch *et al.*, 2008a). Again, this association was only found in the two *Candida* species suggesting recent origin and rapid rewiring.

This prevalence of Mcm1 in various rewiring events, already provides interesting hypotheses about the mechanisms and prerequisites for TFs involved in such events. Firstly, in at least two instances (Rap1 and Wor1), the evolution of the Mcm1-cofactor interaction and the associated rewiring seems to have happened recently suggesting that these events can generally happen fast and especially, if transient, will not necessarily be detectable in inter-species comparisons when phylogenetic distances are large. This might explain some of the puzzling rewirings, especially in the light of fast-evolving *cis*-regulatory sequences and it would be very interesting to delineate the exact evolutionary time-scale of such events. Second, the number of distinct cofactor interactions, but also the striking reoccurrence of Mcm1 at RP promoters, suggests that Mcm1 might act as a kind of "molecular glue" that is especially suited for establishment of co-regulatory interactions and handover of regulons. This is possibly mediated by a promiscuous and evolvable PPI interface (which it undoubtedly has) but also by the fact that Mcm1 is ubiquitous and acts in a cofactor-specific way and is therefore unlikely to link entirely unrelated regulons that might introduce deleterious cross-talk between each other. (We would expect the majority of TFs that Mcm1 might associate with to either not be present in the nucleus or inactive at the same time if they regulate functionally antagonistic genes, although there will clearly be exceptions).

1.4.2.3 Gene Duplication

The third evolutionary mechanism to create regulatory novelty is through gene duplication (Fig. 1.8C). This has become apparent in large scale analyses of regulatory networks where it has been shown that 77% and 69% of homologous TFs in *E. coli* and *S. cerevisiae* share at least one interaction with their duplicates,

accounting for 10% and 22% of interactions, respectively (Teichmann & Babu, 2004). Duplication of TFs is thus likely to be a major evolutionary force driving divergence in transcriptional regulatory networks. Once a TF is duplicated, one or both paralogs may be under relaxed selective constraint and accumulate mutations in the DBD and non-DBD regions and therefore be able to acquire new target genes or interactions with other regulatory proteins. This can allow for subfunctionalisation, i.e. the partitioning of regulatory roles, as is likely to be the case in the example outlined below, or neofunctionalisation through the acquisition of new target genes under the same signal or the opposite scenario, altered control of the TF while maintaining its target genes thereby being able to integrate several signals. The genomic region that is duplicated during such events can vary in size from small-scale duplication of individual genes up to the duplication of entire genomes. Especially the latter has been shown to be of great impact for regulatory evolution. After a WGD event on the yeast lineage, TFs were among the preferentially retained classes of genes and the post-WGD regulatory network has been shown to have diverged rapidly in function trough asymmetric loss of regulatory interaction in paralogous pairs of TFs (Byrne & Wolfe, 2007; Conant & Wolfe, 2006). WGD events can thus facilitate large, but most importantly, coordinated changes of entire regulatory programs through the creation of initial regulatory redundancy.

An example of post-duplication subfunctionalisation again comes from an Mcm1-cofactor interaction. In combination with the arginine-sensing TF Arg81, Mcm1 and Arg80 repress the expression of arginine biosynthetic genes and induce the expression of arginine catabolic genes (reviewed in Messenguy & Dubois, 2003). Arg80 is a duplicate of Mcm1 that has arisen around the time of the WGD. Although first thought to act as a heterodimer with Mcm1, later ChIP analyses indicated that the complex observed at arginine-responsive genes more likely incorporates a homodimer of Arg80 (see Tuch *et al.*, 2008a, and references therein). Unlike Mcm1, which regulates a broad array of functions (see above), Arg80 is specific to arginine-responsive genes and has a much reduced half-life (reviewed in Messenguy & Dubois, 2003) and appears to have taken over the role of Mcm1 in the regulation of arginine metabolism in the post-WGD species, thus representing a regulatory subfunctionalisation.

More recently, a TF duplication has also been implicated in the evolution of RP expression in the post-WGD species (see above). Ifh1 and Crf1 are condition-specific TFs that associate with Fhl1 at RP promoters to either activate (Ifh1) or repress (Crf1) RP expression by competetively binding to the Rap1-Hmo1-Fhl1 complex. Ihf1 is active during fast growth but is repressed during stress response, when Crf1 is expressed at high levels, thereby rapidly switching off the production of RPs (Wapinski *et al.*, 2010). Ifh1 and Crf1 are WGD duplicates of one another and although similar in some regions, Crf1 has lost the N-terminal acidic activation domain and has overall higher amino acid substitution rates (Wapinski *et al.*, 2010). Those authors showed that the observed lack of stress-dependent repression of RPs in *C. glabrata* is most likely due to the loss of the Crf1 ortholog in this species. This thus represents another example of subfunctionalisation; although here both TFs still regulate the same target genes but under different signals and with opposite effects.

A final interesting example of how the initial redundancy created through TF duplication can facilitate divergence and rewiring comes from a study of the YAP-family in yeasts (Kuo *et al.*, 2010). The YAP regulators are a small family of bZIP TFs that each preferentially bind to one of two target sites, YRE-O and YRE-A (Tan *et al.*, 2008), and binding specificity for either motif has been shown to be strongly influenced by the identity of residue 12 in the DBD (Kuo *et al.*, 2010). Despite their largely shared set of target genes, *C. glabrata* YAP1 differs from its homologs in the other yeasts studied at this critical position and, in contrast to YAP1 homologs in the other yeasts, preferentially binds to the YRE-A target site. Examination of promoter regions of shared target genes in *C. glabrata* revealed the coordinated change from a YRE-O to a YRE-A site, highlighting a case of co-evolution between DNA-binding specificity of the TF and the *cis*-regulatory regions in its target genes (Kuo *et al.*, 2010). Most likely, this was facilitated by the presence of a WGD duplicate of YAP1 which has been retained in most post-WGD species. Examination of multiple sequence alignments showed that the WGD paralog of CgYAP1 retained the otherwise conserved residue at the critical position in the DBD and as such likely also binding specificity for the YRE-O site (this study). Here the WGD paralog probably initially buffered against any deleterious effects of the substitution and change of specificity in the CgYAP1

DBD, allowing for gradual change from a YRE-O to a YRE-A motif in CgYAP1 targets.

In addition to gene duplication, new TFs can be gained through other mechanisms such as horizontal gene transfer (HGT) or *de novo* recruitment (reviewed in (Babu, 2010)). HGT is especially important in bacteria where a large number of regulators in the TF repertoire of *Escherichia coli* were found to have arisen by HGT (Price *et al.*, 2008). An example of *de novo* recruitment of a DBD is the eukaryotic WRKY-GCM1 zinc finger family, whose origin has been traced to the DBD part of a transposase (Babu *et al.*, 2006c). TEs thus not only provide "transport vehicles" for *cis*-regulatory DNA but also have the potential to contribute the DBD parts they encode as novel TFs, underlining their potential for the creation of regulatory novelty.

1.4.2.4 Promoter structure and the Evolution of Gene Regulation

Besides changes in the repertoire of transcriptional regulators and turnover in the targets they bind to, promoter structure also plays a role in the evolution of transcriptional regulation (Tsankov *et al.*, 2010). Amongst other factors, whether or not a TF is bound to a target site depends on the accessibility of the region. The binding site might be occluded by nucleosomes (see above). Similarly, binding of a TF might be necessary to alter the local chromatin environment into one that is permissive for transcriptional initiation. Local chromatin structure at promoters is furthermore influenced by the initiation rate of RNA pol II which can displace -1 nucleosomes (e.g. Weiner *et al.*, 2010) and "anti-nucleosomal" sequences mediating constitutively open promoter regions (e.g. Sekinger *et al.*, 2005), both resulting in elevated expression levels. Studying the evolution of gene expression of growth- and stress-related genes in the *Saccharomycotina* and the decoupling of respiratory metabolism from other growth-related genes after the WGD, Tsankov *et al.* found evidence for compensatory turnover of binding sites for different chromatin-associated TFs as well as transitions from one promoter structure to another, including transitions from constitutive to *trans*-regulated promoters trough the loss of antinucleosomal sequences and repositioning of regulatory motifs relative to the NFR. Divergence of promoter structure thus also

plays an important part in regulatory evolution and will be important to consider when interpreting results.

1.4.3 The Role of Transcription Factors

From the above examples, it is clear that evolutionary divergence of TFs is likely to play an important part in the evolution of regulatory networks. In most cases, regulatory rewiring was accompanied by changes in the TFs involved. Rap1 for example has gained a transactivation domain in the common ancestor of the pre- and post-WGD species which might have been fundamental for its ability to take over the regulation of RPs in those clades (Graham *et al.*, 1999; Tanay *et al.*, 2005). Also, the loss of the activation domain of Crf1, is likely to have been an important step for its establishment as a stress-induced repressor of RPs (Wapinski *et al.*, 2010).

Despite the accumulating information about evolutionary dynamics of *cis*-regulatory regions (see above), so far there have been limited efforts in characterising the nature, evolutionary plasticity and conservation of TF repertoires in a systematic way. A good attempt was made by Haerty *et al.* (2008) who collected and characterised TF repertoires in three species of nematodes. However, the functional data available with respect to the regulatory network for those species is relatively limited and complex thus making it more difficult to relate their results to known evolutionary changes and properties of TRNs. Most other studies have either been concerned with specific DBD families (e.g. Rodrigues-Pousada *et al.*, 2010; Thomas & Emerson, 2009), focussed on very wide groups of taxa (e.g. Charoensawan *et al.*, 2010b; Pérez-Rueda *et al.*, 2004; Riechmann *et al.*, 2000) or assumed a "focus species" perspective (e.g. Bussereau *et al.*, 2006; Vaquerizas *et al.*, 2009) and in most cases analysis was often restricted to DBD composition, domain architectures and lineage-specific amplifications. Since the work presented in this thesis was begun, two studies have been published assessing TF repertoires in ascomycetous yeasts, although again evolutionary analysis was relatively superficial with a main focus on DBD families (Park *et al.*, 2008; Shelest, 2008). The results of those studies will be discussed in greater detail in the introduction to Chapter 3.

1.5 The *Saccharomycotina*

The conservation of basic molecular mechanisms with respect to higher eukaryotic organisms and ease of cultivation in the laboratory coupled with rapid growth, simple genetic manipulation techniques and its industrial importance has advanced the yeast *Saccharomyces cerevisiae* to become one of the best-studied eukaryotic model organisms with respect to functional genomics and systems biology approaches (reviewed in Castrillo & Oliver, 2006) as well as comparative genomic studies (reviewed in Dujon, 2010). *S. cerevisiae* belongs to the *Saccharomycotina*, a clade containing over 1000 known members (Stajich *et al.*, 2009) that, owing to extensive sequencing efforts during the last 10 years, has become very well characterised with respect to patterns of genome evolution (e.g. Butler *et al.*, 2009; Dietrich *et al.*, 2004; Dujon *et al.*, 2004; Génolevures Consortium *et al.*, 2009; Kellis *et al.*, 2003). This together with the increasing amount of knowledge concerning the evolution of gene regulation in those species (reviewed above) made the *Saccharomycotina* ideally suited for a genome-wide comparison of TF repertoires. Furthermore, the *Saccharomycotina* display several interesting features with respect to genome evolution, e.g. a whole-genome duplication (WGD) event on the lineage leading to the *Saccharomyces* species, and contain the important human pathogen *Candida albicans*. It will be interesting to examine TF evolution in the light of these different mechanisms of genome evolution and determine their impact on the evolution of pathogenicity in the *Candida* clade.

1.5.1 Taxonomy and Genome Evolution

The *Saccharomycotina* are one of the three major clades within the *Ascomycota* (Figure 1.9) which furthermore encompass the *Pezizomycotina* (containing the filamentous *Aspergillus* species) and the *Thaphrinomycotina* (containing the fission yeast *Shizosaccharomyces pombe*; Hibbett *et al.*, 2007; Suh *et al.*, 2006). Their defining characteristics include a cell wall composition that mainly consists of β-glucans (compared to chitin in the *Basidiomycota*), GC content below 50% and their often more fermentative lifestyle. Furthermore, ascomycete yeasts are ecologically diverse and tend to be nutritional specialists, frequently associated with

plants and animals and found in liquid environments that are rich in organic carbon (reviewed in Suh *et al.*, 2006). The *Saccharomycotina* largely contain yeast forms occurring as haploids or diploids that predominantly go through long cycles of clonal growth by mitotic division ("budding") rather than sexual reproduction, although the majority of those species appears to be capable of going through a sexual cycle (reviewed in Knop, 2006; Fig. 1.9). Based on molecular estimates, the split leading to the divergence of the *Saccharomycotina* from the remaining *Ascomycetes* is thought to have occured between 300 and 500 mya (Taylor & Berbee, 2006).

Within the *Saccharomycotina* there are three major clades: the "CTG species" that are named after a change in the genetic code, leading to the CUG codon being translated into a serine instead of a leucine (Miranda *et al.*, 2006; Santos & Tuite, 1995) and encompass the *Candida* species as well as *Debaryomyces hansenii* and *Lodderomyces elongisporus*, and the pre-WGD and post-WGD species together referred to as the *Saccharomycetaceae*. The pre-WGD and post-WGD species are separated by a WGD event that is though to have occurred about 100 mya (Kellis *et al.*, 2004; Wolfe & Shields, 1997) and led to a doubling of chromosome numbers that is still visible in the 2:1 syntenic relationship of chromosomal regions between the post-WGD and pre-WGD species (reviewed in Dujon, 2010). The majority of WGD paralogs ("ohnologs") have subsequently been lost in a lineage- and species-specific manner, however, with *S.cerevisiae* having retained approximately 550 ohnologs, corresponding to about 10% of protein-coding genes (Byrne & Wolfe, 2005).

Most of the *Saccharomycotina* genomes encode between 5000 and 6000 protein-coding genes and post-WGD and CTG species generally encode more genes than the pre-WGD species (reviewed in Dujon, 2010). They tend to be intron-poor, especially in the *Saccharomycetaceae* where the number of intron-containing genes is around 3-6% and only a handful of genes have more than one intron (Génolevures Consortium *et al.*, 2009). The CTG species have a slightly larger percentage of introns in their protein-coding genes, although still very low at approximately 6-7% (reviewed in Dujon, 2010). Syntenic relationships within the three major clades are highly conserved, e.g. between *S. cerevisiae* and *Candida glabrata* approximately 58% of genes are found in syntenic regions (Montcalm & Wolfe,

2006). This is in contrast to the large amounts of sequence divergence observed. Conservation of protein-coding sequences between those two species has been estimated to be equivalent to levels of conservation between human and fish (Dujon *et al.*, 2004) and equivalent patterns are seen in the pre-WGD (Génolevures Consortium *et al.*, 2009) and CTG species (Butler *et al.*, 2009). This rapid sequence divergence has been attributed to the largely clonal lifestyle of these yeasts and presumably frequent population bottlenecks when new colonies are established (Dujon, 2010). Indeed outcrossing in natural populations appears to be rare (reviewed in Knop, 2006 and Dujon, 2010) and is also likely to be the cause of the scarcity of transposable elements and pseudogenes found in the *Saccharomycotina*.

1.5.2 Life style and Ecology

Ecological niches for the *Saccharomycotina* are relatively poorly defined. Many species, including *S. cerevisiae*, continue to be isolated from different environments, indicating that we have yet to define their full ecological ranges (e.g. Pennisi, 2005). Most species within the *Saccharomycetaceae* are associated with plants, e.g. *S. cerevisiae* has been isolated from grapes, the bark of oak trees, nectar and tree sap together with other *sensu stricto* species (e.g. Sampaio & Gonçalves, 2008). The *Saccharomycetaceae* however also contain *C. glabrata*, a human commensal and opportunistic pathogen (e.g. Pfaller & Diekema, 2007) and *Ashbya gossypii*, first isolated as a cotton pathogen, which has been subsequently found to be closely associated with a beetle (reviewed in Wendland & Walther, 2005). Most species in the CTG clade are generally associated with mammals, although again have been isolated from various locations. *C. tropicalis*, for example, has been found in soil, clinical samples, fermentation vats and rotten pineapples (Suh *et al.*, 2006). *Debaryomyces hansenii*, a close relative of the *Candida* species, is a marine yeast and has extremely high salt tolerance (Almagro *et al.*, 2000) but is rarely isolated from humans (Butler *et al.*, 2009).

1.5.2.1 Carbohydrate Metabolism

One of the defining features of *S. cerevisiae* and its close relatives is the ability to preferentially ferment glucose, even in the presence of oxygen. This is referred to

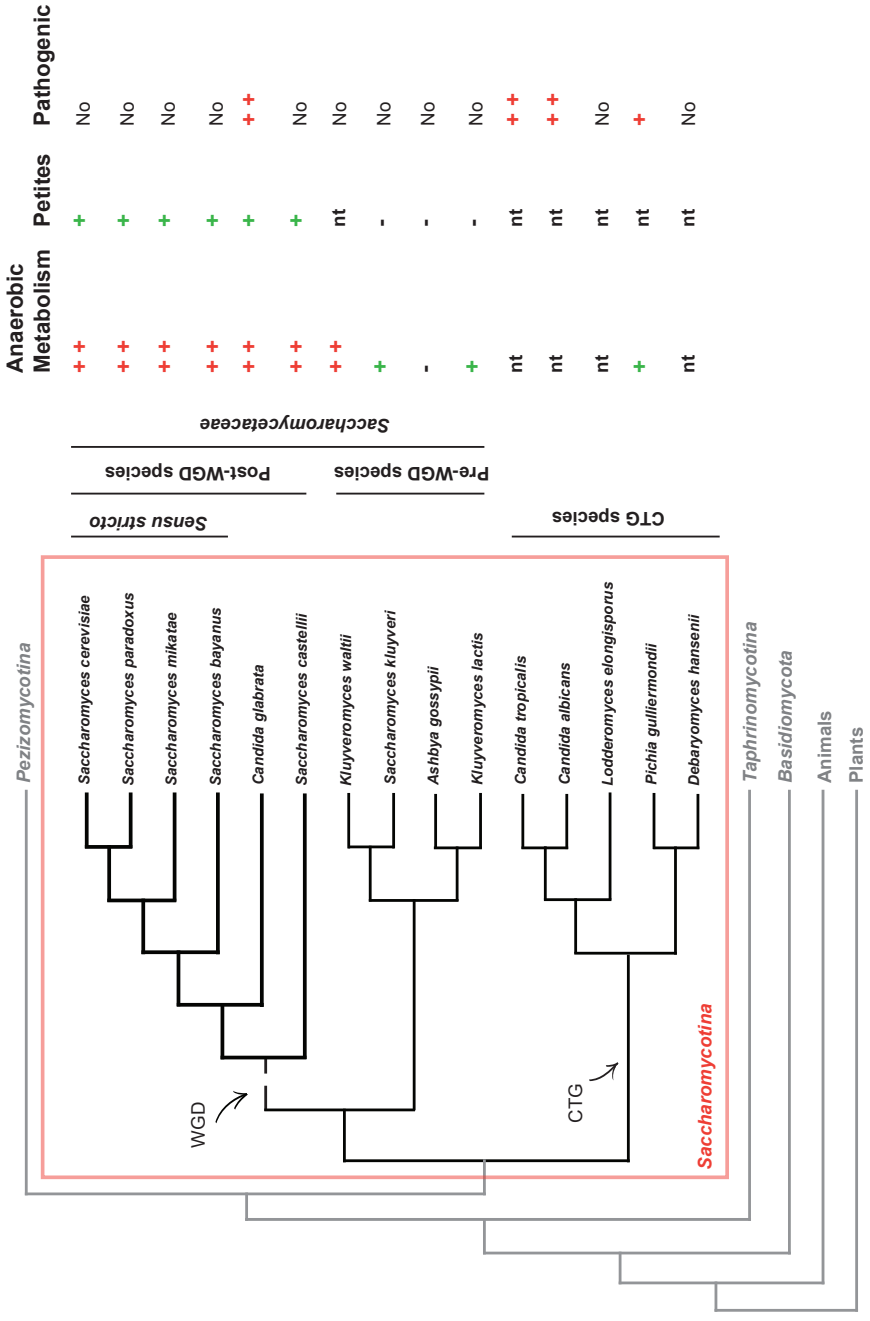


Figure 1.9: Phylogeny of the *Saccharomycotina* inferred in the phylogenomics study presented in Chapter 2. "Anaerobic metabolism" and "Petites" are based on the results from Merico *et al.* (2007) and refer to the ability to grow efficiently under anaerobic conditions in minimal media (++), grow anaerobically but in supplemented media and low biomass yield (+) or no growth in anaerobic media tested (-) as well as the ability to form respiratory-deficient petite mutants. "Pathogenicity" is based on Butler *et al.* (2009) and corresponds to no pathogenicity (No), weak pathogenicity (+) or strong pathogenicity (++).

as the Crabtree effect (De Deken, 1966). In glucose-rich conditions, *S. cerevisiae* ferments glucose to ethanol which is secreted from the cell. Once all the glucose is depleted, the cells switch to aerobic metabolism and consume the ethanol produced during fermentation. This is called the diauxic shift (Rolland *et al.*, 2002). Although the rate of ATP production is higher during fermentation, the net yield of ATP per molecule of glucose is lower and as such is not energetically favourable as a long-term nutritional strategy. Furthermore, anaerobic metabolism comes at an additional cost through the need to adjust the redox balance in absence of the respiratory chain to avoid accumulation of genotoxic reactive oxygen species and the need for molecular oxygen in a number of biosynthetic pathways, such as sterol and fatty acid biosynthesis, heme, NAD or uracil (reviewed in Merico *et al.*, 2007). *S. cerevisiae* is thought to have overcome these hurdles through a number of enzymatic adjustments (see below; reviewed in Piskur *et al.*, 2006) as well as the regulatory decoupling of oxidative metabolism and fermentative metabolism, e.g. through the rewiring of mitochondrial ribosomal proteins (see above; Ihmels *et al.*, 2005), repression of oxidative metabolism in high-glucose conditions (Carlson, 1999) and integration into intra- and extracellular signalling pathways to allow for a rapid response to changing environmental conditions (Brauer *et al.*, 2008). Populations of *S. cerevisiae* thus have a "make, accumulate, consume" strategy where growth is exponential during fermentation and then goes into a stationary phase in low glucose or high stress conditions, and cycles between the two depending on glucose availability (reviewed in Piskur *et al.*, 2006; Tu *et al.*, 2007). Furthermore, the ethanol produced and secreted during exponential growth acts as an antimicrobial agent against other organisms and as such might have conferred a competitive advantage.

Yeasts can be classified into obligate aerobes, that cannot survive without oxygen, facultative anaerobes, capable of both aerobic and anaerobic metabolism, and obligate anaerobes, that only use fermentative pathways, according to their oxygen metabolic requirements. Most species within the *Saccharomycotina* are facultative anaerobes, meaning they can survive in anaerobic conditions. The efficiency and nutritional requirements with which they do so however vary considerably between species (Merico *et al.*, 2007; Fig. 1.9; "Anaerobic Metabolism"

and "Petites"). The post-WGD species are generally Crabtree positive (i.e. perform aerobic fermentation), can grow under strict anaerobic conditions (on minimal media and with impaired respiratory chain) and are able to generate petite mutants, characterised by large deletion or losses of mitochondrial DNA. The lineages that did not undergo a WGD display a mosaic of these traits, with none of them displaying all traits at the same quantitative level as the post-WGD species (Merico *et al.*, 2007). It has been proposed that while the ability to survive in anaerobic environments is ancient and shared by many yeasts, the ability to efficiently ferment glucose was probably selected for during the end of the Cretaceous period when large amounts of fruit became available (Merico *et al.*, 2007; Piskur *et al.*, 2006).

Indeed, the importance of the WGD in the establishment of crabtree-positive lifestyles has been implicated a number of times (e.g. Conant & Wolfe, 2007; van Hoek & Hogeweg, 2009). For example, an ohnolog of the alcohol dehydrogenase gene *ADH* was retained after WGD and gained different enzymatic affinities allowing for efficient conversion of acetaldehyde to ethanol, an important step in alcoholic fermentation (Dujon, 2010). Also, a number of glycolytic enzymes and hexose transporters were retained after WGD and it has been proposed that this led to an effective increase in glycolytic flux (Conant & Wolfe, 2007) which has been further validated theoretically (van Hoek & Hogeweg, 2009) and experimentally (e.g. Lin & Li, 2010).

1.5.2.2 Pathogenicity

Three out of the five species belonging to the CTG clade included in my study and the post-WGD species *C. glabrata* are human commensals and opportunistic pathogens (see Fig. 1.9; "Pathogenic"). While most human contact with *Candida* species is benign and over 50% of humans harbour commensal populations, they can result in serious systemic infections especially in immunocompromised patients (e.g. following chemotherapy or organ transplants). Over 10 to 15% of bloodstream infections are caused by *Candida* species and mortality rates are relatively high. This is in part also due to the fact, that being eukaryotic and

thus more similar to humans in terms of their biochemistry, it has been very difficult to develop effective antifungal drugs and strains frequently develop resistance (reviewed in Brown, 2006; Hynes *et al.*, 2006).

Pathogenicity in those species has largely been attributed to the ability to undergo morphogenic transitions, secrete hydrolytic enzymes, respond to changes in the host environment such as changes in pH or osmolarity, and adhere to host surfaces (reviewed in Brown, 2006). Comparative genomics of pathogenic species has revealed expansions in gene families encoding extracellular lipases, transmembrane transporter families, surface adhesion proteins and proteins involved in iron acquisition that were absent or present only in low copy number in non-pathogenic species (Butler *et al.*, 2009). Interestingly, none of those families was equally amplified in the more distantly-related pathogen *C. glabrata* suggesting convergent evolution of pathogenicity by different molecular mechanisms despite the overall similarity of biological processes involved (Roetzer *et al.*, 2010). This has also become apparent in more detailed studies of factors known to be important for virulence in *C. albicans*, e.g. the TF Ace2 which regulates cell separation in *C. albicans* and the deletion of which results in attenuated virulence. Deletion of the *C. glabarata* ortholog surprisingly was found to increase virulence by 200 fold (reviewed in Hynes *et al.*, 2006), indicating related pathways but with drastically different outcomes. The ability to infect different host tissues and escape the human immune system is furthermore tightly tied to condition-specific regulation of metabolic activity, seeing that e.g. oxygen availability varies greatly across different tissues that thus represent different niches. This is reflected in the coordination of carbohydrate metabolism with phenotypic switching from white to opaque (see above) in *C. albicans* (reviewed in Brown, 2006) or the extreme resistance against starvation in *C. glabrata* in aid of its survival upon being engulfed by macrophages (reviewed in Roetzer *et al.*, 2010).

1.6 Transcription Factor Repertoires in the *Saccharomycotina*

The theoretical arguments and experimental evidence introduced above, thus established that TF divergence is an important part of regulatory evolution. To address evolutionary dynamics of TF repertoires on a systematic scale, I have collected TF repertoires from 15 species of yeasts belonging to the *Saccharomycotina*. Instead of relying on published datasets, I have developed a pipeline to assemble TF repertoires in yeasts to obtain an unbiased and high-coverage dataset of TFs. I then analysed the collected repertoires with respect to the types of DNA-binding domain (DBD) and domain architectures found (Chapter 3). This revealed interesting patterns of global DBD family evolution and numerous detailed examples of increase and decrease in copy numbers that are likely to be of functional importance to the evolution of the species studied. To gain a deeper understanding of regulator turnover in TF repertoires through gene duplication and loss, I have inferred evolutionary histories for each group of homologous TFs and investigated lineage-specific patterns of amplifications and losses and how those relate to the architecture of regulatory networks (Chapter 4). The patterns of gene duplication and loss across the clades studied were very different: while regulatory network growth in the post-WGD species was mainly due to retention of WGD duplicates and non-family specific, TF gain in the *Candida* species was dominated by ongoing lineage-specific amplification of the $Zn(II)_2Cys_6$ and fungal-specific zinc finger DBDs, the two largest TF families. Furthermore I found WGD duplicate TFs to be enriched for highly-connected TFs and stress- and nutrient-responsive regulators. I have investigated evolutionary rate constraints of TFs in relation to their position in the regulatory network (Chapter 5) and examined how these vary between different clades (Chapter 6). Analysis of evolutionary rates between clades reflected known examples of regulatory rewiring in the *Saccharomycotina* as well as again implicating extensive evolutionary changes in signal transduction pathways mediating the response to nutrient availability and different stresses. Examples of this are discussed in further detail in Chapter 7.

Furthermore, I have investigated the use of evolutionary models accounting for between-gene heterogeneity in phylogenomic approaches where hundreds of

genes are used to infer species phylogenies. Often in such studies, genes are concatenated into a single "supergene" that is then analysed using a single parameterisation of an evolutionary model. It is known however that evolutionary processes between genes can differ, which can lead to model violation and potentially the reconstruction of well-supported non-optimal trees. With the aim of obtaining a trustworthy species tree to be used in the evolutionary analyses described above, I collected a phylogenomics dataset for 18 species within the *Saccharomycotina* and tested the use of concatenated and partitioned models, where the data were partitioned by genes and every partition was parameterised separately thereby accounting for between-gene heterogeneity. Partitioned analysis consistently outperformed concatenated analysis, although here this did not affect the resulting species phylogeny. The choice of evolutionary model, especially accounting for differences in evolutionary processes at different codon positions, however did affect results, suggesting that within-gene heterogeneity resulted in stronger conflicting signal than between-gene heterogeneity and underlining the need to use good models of evolution.

Chapter 2

Finding the Yeast Phylogeny - Phylogenomic Approaches

2.1 Introduction

One of the most important pieces of information for any kind of comparative evolutionary analysis is the species phylogeny. It forms the basis of all downstream evolutionary analyses and having an accurate species tree is crucial for the inference of both qualitative and quantitative evolutionary properties. Phylogenomic methods that make use of genome-wide data have become a standard approach to resolving species phylogenies. Classic molecular systematic methods rely on one or a few genes that are considered to be phylogenetically informative such as ribosomal RNA or mitochondrial genes. In contrast, genome-wide analysis tries to utilise the maximum amount of information encoded in multiple genomes to reconstruct inter-species relationships (Philippe *et al.*, 2005). By combining data from different parts of the genomes we try to minimise the effect of sampling error which is encountered when a small number of characters (e.g. single genes) is analysed and which can affect phylogenetic reconstruction. It is well-known, however, that heterogeneities in the evolutionary process within single genes, such as different substitution rates across sites, can markedly affect phylogenetic reconstruction (e.g. Delsuc *et al.*, 2003, Brinkmann *et al.*, 2005, Nishihara *et al.*, 2007). So in order to gain maximum profit from the increased amount of data, which can only be expected to increase heterogeneity as data from different regions of

the genomes are included, it is necessary to explore the use of appropriate models that deal with variations in the evolutionary processes across different loci and between different species.

In a classic phylogenomics study encompassing 106 genes from seven species of yeast, Rokas *et al.* (2003) concluded that a supermatrix analysis, where all individual gene alignments were concatenated into a "superalignment" and analysed using a standard (homogeneous) evolutionary model, could give a confident species tree where analysis of the individual genes failed to find a congruent solution. As increasing amounts of data from yeasts and other organisms become available, it is an appropriate time to consider whether such methods are still practical and reliable when applied to larger datasets like those that we are able to assemble today and whether those conclusions hold for harder phylogenetic problems, for example spanning a larger period of evolution. I have investigated this, and generated a reliable phylogeny of an increased number of yeast species for which full genomes were available at the time, by an extended study of yeast phylogenomics [1].

2.1.1 Species Tree Reconstruction

The two most widely-used methods for inferring phylogenies from multiple genomic regions, supermatrix analysis and supertree reconstruction, have been reviewed by Delsuc *et al.* (2005). Both are based on the analysis of multiple sequence alignments of each of many orthologous genes found in the genomes of interest. In supertree analysis, the separate alignments are analysed on a gene-by-gene basis and individually reconstructed gene trees are combined by some criterion to yield a global hypothesis about the phylogenetic relationship of the species studied (Bininda-Emonds, 2005).

In supermatrix analysis, the alignments of the orthologous genes are analysed simultaneously in order to derive the species phylogeny, the aim being to amplify the phylogenetic support in the data, or in other words, increase the signal to noise ratio (Rokas *et al.*, 2003). The most commonly used approach is a simple concatenation where the alignments of all genes are concatenated into a single

[1] This work has been prepared for publication and is currently under review for submission.

superalignment and analysed as such. Further, more complex, model-based variations of this approach use, for example, partitioning of the superalignment into genes or codon positions which are often treated differently, in recognition of potentially diverse evolutionary dynamics — but these essentially still describe a supermatrix analysis. Whereas the term 'supermatrix analysis' often refers to the simple concatenation approach, in this chapter I will use it to refer to all such approaches and distinguish these levels of complexity by denoting them as either "concatenated" (all partitions treated equally) or "partitioned" (subset of the sites treated differently).

Furthermore, several probabilistic methods that explicitly model individual gene evolutionary histories of the loci analysed within the context of the coalescent to find the most likely species phylogeny have been proposed recently (e.g. Heled & Drummond, 2010; Kubatko *et al.*, 2009; for review see Degnan & Rosenberg, 2009). While those methods look very promising for the resolution of short internal branches where we expect large amounts of discordance between gene trees and species trees due to coalescent effects (e.g. Degnan & Rosenberg, 2006) they remain computationally intractable for phylogenomic datasets and therefore were not considered in the context studied here.

Simulation studies are indicating that supermatrix methods, given an appropriate model of evolution, generally are more accurate in recovering the underlying species phylogeny than supertree methods (e.g. Ren *et al.*, 2008). The only case in which supertree methods were found to outperform supermatrix analysis was in the presence of strong coalescent effects such as incomplete lineage sorting (Kupzcok et al., submitted). When incongruence through independent lineage sorting is frequent and the genes showing such incongruent gene trees contribute many columns to the supermatrix, the resulting erroneous signal becomes stronger than the signal introduced through gene tree reconstruction artefacts that otherwise dominate supertree methods. Here, I focussed on supermatrix analysis, to take advantage of this increased accuracy and to obtain results that are comparable to the study undertaken by Rokas *et al.*, and also because it relies on primary sequence data to generate hypotheses and allows the parameterisation and measurement of variation in the data used and test the importance of such parameterisation.

2.1.2 Considerations with the Supermatrix Approach

The rationale for the supermatrix approach is the assumption that most genes in a genome share their evolutionary history and are thus expected to bear the same phylogenetic signal. Concatenation of the genes aims to amplify this shared phylogenetic signal by increasing the sample size, to reduce the effect of stochastic processes that might dominate single gene phylogenies.

There are a number of well-known issues that can influence the accuracy of phylogenetic reconstruction in general. The most well-studied among those is probably across-site rate variation (Uzzell & Corbin, 1971) which is typically accommodated by adding gamma-distributed rates to the evolutionary model (Yang, 1994). Furthermore, not all substitutions occur at equal rates, e.g. A to G and T to C transitions occur at higher frequency than A,G to T,C transversions (Hasegawa *et al.*, 1985), and the placement of a residue within the secondary or tertiary structure of a protein can influence the patterns of substitutions observed: buried residues tend to evolve slower than exposed residues and not all substitutions are equally frequent at certain positions in different elements of secondary structure due to their physiochemical properties (Goldman *et al.*, 1998; Koshi & Goldstein, 1995; Le & Gascuel, 2010; Thorne *et al.*, 1996). In addition to variation in space, variation of the evolutionary process in time such as site-specific rate variation across lineages, referred to as heterotachy (Lopez *et al.*, 2002), can adversely affect the outcome of phylogenetic reconstruction. Whilst some solutions towards accommodating those processes are beginning to appear (e.g. Lartillot & Philippe, 2004, Pagel & Meade, 2008, Whelan, 2008), they remain less well-understood. Other issues that are difficult to account for *per se* include compositional bias (Lockhart *et al.*, 1994) and mutational saturation that can lead to long branch attraction artefacts (Ho & Jermiin, 2004).

Beside the difficulties associated with heterogeneity in substitution patterns, population genetic processes can result in discordance between gene trees and the species tree, essentially violating the assumptions of the supermatrix approach. Those include incomplete lineage sorting, i.e. ancestral polymorphisms that predate species divergence (reviewed in Degnan & Rosenberg, 2009), recombination

and gene conversion (Posada & Crandall, 2002; Schierup & Hein, 2000) or horizontal gene transfer (reviewed in Philippe & Douady, 2003). This is particularly problematic for short internal branches (Degnan & Rosenberg, 2006, 2009).

Other than problems arising from the nature of the sequences and their evolution themselves there are additional factors that can impact the accuracy of tree reconstruction. Those include gene duplication and loss that can lead to analysis of paralogous sequences that do not share the same history of substitutions and thus introduce conflicting signal. Gene sampling can also impact phylogenomic analyses as not all genes are equally informative to the resolution of a particular clade (Kuramae *et al.*, 2007). Furthermore, sequence alignment is another well-known contributor to variability in tree reconstruction, arising from misalignment of non-homologous positions (e.g. Löytynoja & Goldman, 2008; Marcet-Houben & Gabaldón, 2009). Whilst these issues provide serious challenges to phylogenomic studies, their effects can be alleviated by careful data collection and choice of methods to minimise their impact.

Seeing that the effects of heterogeneous evolutionary processes are already apparent even in single-gene studies, they can only be expected to gain in impact when data from multiple genomic regions are being analysed. Even when we account for processes such as different rates across sites, concatenation and subsequent supermatrix analysis using an evolutionary model with a single set of parameters for the entire dataset is assuming homogeneity, or rather a "constant heterogeneity", of the evolutionary process.

This is highly unlikely to hold and it has been shown that systematic errors resulting from the model violations mentioned above can be exacerbated by concatenation to the extent where highly-supported but incorrect topologies are recovered (Brinkmann *et al.*, 2005; Delsuc *et al.*, 2003; Nishihara *et al.*, 2007). A number of treatments to mitigate such effects have been proposed, e.g. increased taxon sampling to break up long branches and thereby reduce the effect of multiple substitutions (Philippe *et al.*, 2005); the removal of fast-evolving species, genes or sites as those are suspected to be most prone to accumulate nonphylogenetic signal (Jeffroy *et al.*, 2006; Nishihara *et al.*, 2007; Rodríguez-Ezpeleta *et al.*, 2007); and recoding of data as purines and pyrimidines only (RY-coding) for nucleotides (Phillips *et al.*, 2004) or according to functional categories for amino

acids, to reduce compositional bias (Rodríguez-Ezpeleta *et al.*, 2007). Although these measures seem to work in some cases (e.g. Rodríguez-Ezpeleta *et al.*, 2007), they are treating the symptoms of the problems, not the causes; typically they discard potentially informative parts of the data and it is unclear in what way this affects the signal-to-noise ratio.

Here, I prefer to address heterogeneity issues by using more sophisticated models to fit the data, aiming to retain all useful information, rather than discarding parts of the data to fit the models being used. Partitioned analysis in which parameters of the evolutionary model are estimated separately for each partition (in my case, each gene) in the dataset is a solution whose efficacy has recently been demonstrated in studies on simulated data (Ren *et al.*, 2008) as well as empirically by Nishihara *et al.* (2007) in a study of the branching order at the base of the mammals. Similarly, mixture models provide means for addressing heterogeneities in the data by using different substitution matrices (e.g. Lartillot & Philippe, 2004; Le *et al.*, 2008) or different sets of branch lengths (Pagel & Meade, 2008) for different pre-defined or learned partitions in the dataset. They are however likely to be computationally very expensive for large datasets and partitioned models, which are essentially a generalisation of mixture models, are, despite the large number of parameters estimated, computationally more feasible due to the ability to easily parellelise computation. I thus decided to focus on those.

2.1.3 Yeast Phylogenomics

The ascomycetous yeasts have been the focus of many smaller (Diezmann *et al.*, 2004; Kurtzman & Robnett, 2003; Schoch *et al.*, 2009; Tsui *et al.*, 2008) and larger (Bofkin, 2005; Cornell *et al.*, 2007; Fitzpatrick *et al.*, 2006; Jeffroy *et al.*, 2006; Kuramae *et al.*, 2006, 2007; Marcet-Houben & Gabaldón, 2009; Phillips *et al.*, 2004; Rokas *et al.*, 2003; Wang *et al.*, 2009) phylogenomic studies. While the smaller studies cited encompass a large range of species, they incorporate only a handful of purposely sequenced genes in their supermatrix analyses and thus add relatively little extra data. The first study attempting a larger scale analysis was conducted by Rokas *et al.* (2003) with 106 genes, focussing on the

relationships within the *Saccharomyces sensu stricto* species only. Fitzpatrick *et al.* (2006) extended this by using a slightly larger dataset (153 genes) and a wider phylogenetic range across the Ascomycota. More recent studies (Kuramae *et al.*, 2006; Marcet-Houben & Gabaldón, 2009) further increased the number of genes and species studied, analysing concatenated datasets of 531 and 1137 genes in 21 species respectively. The analysis conducted by (Wang *et al.*, 2009) does not use supermatrix analysis however, but instead employs a composition vector method that is based on genomic composition of the species analyzed. This approach lacks a reliable statistical framework and as such cannot really be assessed or compared to supermatrix approaches.

There is widespread agreement about the monophyly of the clade containing the species that underwent a whole-genome duplication (WGD) and their close relatives who form a sister clade to the WGD clade (Kellis *et al.*, 2004), as well as the monophyly of the clade including *Candida* and *Pichia* species and *Debaryomyces hansenii* (Figure 2.1A). The relationships within the *Saccharomyces sensu stricto* complex also seem well established. The maximum parsimony (MP) analysis by Kurtzman & Robnett (2003) shows some contradictory branching, placing *S. paradoxus* and *S. mikatae* as sister species (alternative branching not shown); this difference is however minor and it is otherwise well-accepted that *S. cerevisiae* and *S. paradoxus* are more closely related to each other.

In contrast, the relationships at the base of *Saccharomycs sensu stricto* species complex are unclear. Whereas most phylogenomic analyses recover *C. glabrata* as the species at the base of the WGD clade (Figure 2.1B), synteny data from an analysis of chromosomal inversions give strong evidence for *S. castellii* branching off earlier than *C. glabrata* (Scannell *et al.*, 2006) (Figure 2.1C) . In their supermatrix analysis of 153 genes, Fitzpatrick *et al.* (2006) recover *S. castellii* at the base of the WGD clade with strong support (bootstrap values $> 90\%$). Schoch *et al.* (2009) also recover this branching in their analysis of six genes, whereas Tsui *et al.* (2008) find *C. glabrata* and *S. castellii* as sister species at the root of the WGD clade (Figure 2.1D) . Supertree analysis in the study by Fitzpatrick *et al.* (2006) however does not recover this branching but instead sees *C. glabrata* branching off first. The Bayesian analysis by Jeffroy *et al.* (2006) using a similar-

sized dataset of amino acid sequences also supports the topology with *C. glabrata* at the base of the WGD species with significant support.

Whilst in recent phylogenomics studies the relationships between the pre-WGD species including *S. kluyveri*, *Ashbya gossypii*, *Kluyveromyces lactis* and *K. waltii* have usually been recovered as a clade where *K. lactis* and *A. gossypii* are sister species and *S. kluyveri* and *K. waltii* are sister species (Figure 2.1F), there has been some previous disagreement about their branching order (Cornell *et al.*, 2007; Kurtzman & Robnett, 2003; Wang *et al.*, 2009; Figure 2.1E) and it will be useful to confirm the relationships in this clade using sophisticated methods.

A further few uncertainties can be found within the clade encompassing the *Candida* and *Pichia* species. While *Candida albicans*, *Candida tropicalis* and *Lodderomyces elongisporus* are stably recovered as a monophyletic clade where *C. albicans* and *C. tropicalis* are sister species, the exact relationships between the remaining members of the clade are unclear. Tsui *et al.* (2008) find *D. hansenii*, *Pichia stipitis* and *Pichia gulliermondii* as a monophyletic clade where the two *Pichia* species are most closely related to each other to the exclusion of *D. hansenii* (Figure 2.1H). This clade is also recovered by Wang *et al.* (2009) but they found *D. hansenii* and *Pichia stipitis* to be sister species instead (Figure 2.1I). The very large study by Marcet-Houben & Gabaldón again recovers a different branching but here these species are paraphyletic with *P. stipitis* being most closely related to the clade containing the *Candida* species, followed by *D. hansenii* and *P. gulliermondii* (Figure 2.1J). The remaining studies largely recover the topology shown in Figure 2.1G (Fitzpatrick *et al.*, 2006; Jeffroy *et al.*, 2006; Kurtzman & Robnett, 2003; Schoch *et al.*, 2009).

The branching order at the base of the WGD species, as well as some of the relationships within the pre-WGD species and the *Candida* clade hence present problems that remain unsolved and we hope to obtain a more definite answer by performing more sophisticated analyses. It will be useful to examine the placement of those species on the ascomycete phylogeny when more appropriate methods for the analysis of multigene datasets are being used.

With the exception of the studies presented in Diezmann *et al.* (2004), Bofkin (2005), Tsui *et al.* (2008) and Schoch *et al.* (2009), all supermatrix analyses

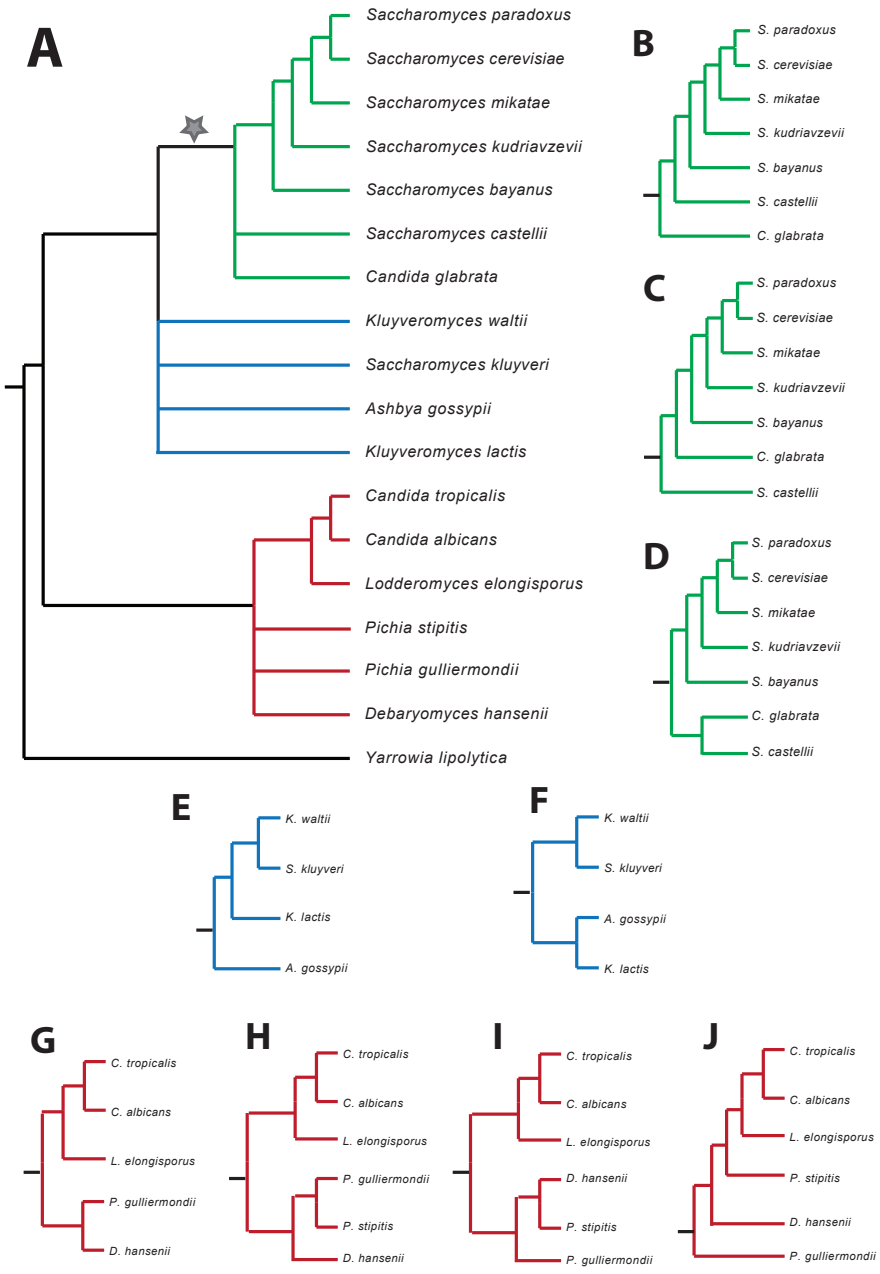


Figure 2.1: Topologies recovered by the four most inclusive previous supermatrix studies of fungi, trimmed to include only species that are considered in this study. The red branches indicate regions that show differences compared to a general consensus over these studies. The star indicates the position of the WGD event. **A:** Combined maximum likelihood analysis of four nuclear and two rDNA genes (Diezmann *et al.*, 2004). **B:** Tree recovered from Bayesian analysis of 106 amino acid sequences (Jeffroy *et al.*, 2006). **C:** Maximum likelihood tree from analysis of 153 amino acid sequences (Fitzpatrick *et al.*, 2006). **D:** Most parsimonious tree from a maximum parsimony analysis of 18S and 5.8S internal transcribed spacer, three 26S rDNAs, EF-1, mitochondrial SSU rDNA and COX II nucleotide sequences (Kurtzman & Robnett, 2003).

mentioned above have been carried out in a concatenated and hence rather simplistic manner. Some (e.g. Rokas *et al.*, 2003) do not address systematic error at all; others try to account for non-phylogenetic signal simply by removal of fast-evolving sites or by RY-recoding (Fitzpatrick *et al.*, 2006; Jeffroy *et al.*, 2006; Phillips *et al.*, 2004). As mentioned earlier, however, those treatments are not well-suited for a comparative analysis and do not directly address heterogeneity issues between the concatenated genes which can lead to phylogenetic inconsistency and reconstruction of wrong trees with high support. In order to obtain a high-quality species tree for 18 ascomycetous yeasts and to investigate how more data contribute towards solving more difficult phylogenetic problems, I have extended the well-known phylogenomics study of Rokas *et al.* by considering 10 additional species, increasing the diversity to a range of species that shared their last common ancestor about 250 million years ago and approximately trebling the number of genes to 343. This represents a phylogenomic dataset of a scale that is more typical of the problems that are studied today.

Furthermore I wanted to examine the effects of more sophisticated models accounting for both intra- and inter-gene heterogeneity of evolutionary dynamics, especially with regards to the conclusions drawn by Rokas *et al.* (2003), who claim to have obtained high-confidence results from rather simplistic analysis. It is known that over-simplification can lead to over-confidence (Sullivan & Swofford, 1997; Yang *et al.*, 1994) and it is interesting to see if those conclusions hold for more complex datasets and analyses such as ours.

The nucleotide sequences of the 343 genes were analysed both individually and as a supermatrix. I explored the signals present in the data when analysed as single genes and determined the impact of model choice and tree reconstruction method used. I examined the validity of using a supermatrix analysis on a dataset of this scale and investigated more sophisticated maximum likelihood (ML) analyses, accounting for heterogeneities between the genes while considering the entire dataset. Furthermore I was able to achieve robust estimates of controversial regions of the phylogeny using thorough modelling of supermatrix data. Analysis of the 343 genes' amino acid sequences confirmed these results.

2.2 Data Collection and Analysis

The core of the species that were selected for analysis were the eight species that were included in the Rokas *et al.* (2003) study. I considered 10 additional, more divergent, ascomycetous yeasts including the well-studied pathogenic fungus *Candida albicans* and the distantly related *Yarrowia lipolytica* as an outgroup (Dujon *et al.*, 2004). See Figure 2.1 for the complete list of species studied. Primary analyses were performed on nucleotide coding sequences. In order to reinforce the results obtained with the nucleotide dataset in concatenated and partitioned supermatrix analyses, I repeated those using amino acid data.

The Fungal Orthogroups Repository (FOR) at the Broad Institute (Wapinski *et al.*, 2007a) contains orthology assignments of protein-coding genes for 14 out of the 18 species considered in this study. The orthology assignments available in FOR are results of computational synteny-asissted homology reconstruction (Wapinski *et al.*, 2007a) but also incorporate curated homology assignments from the yeast gene order browser YGOB (Byrne & Wolfe, 2006) and are therefore believed to be of high quality (see Figure 4.3). Genes are grouped into homologous clusters based on sequence similarity and synteny conservation and the gene tree for each cluster is built by recursively traversing the species tree from leaves to root. At each node, the relationships for the species below that node are resolved using a Neigbor-Joining approach until the root is reached. Orthologous genes (one-to-one orthologies, but also one-to-many relationships) are grouped together in so-called orthogroups (Wapinski *et al.*, 2007a; see Chapter 4.1.1 for a full discussion of the underlying method). I screened FOR for orthogroups with exactly one representative one-to-one orthologs in each of the 14 species studied that were included in FOR, resulting in an initial list of 1148 orthogroups.

For the purpose of mapping those orthogroups to the remaining four species that were not included in FOR, I used the amino acid sequence of the representative *S. cerevisiae* protein for each orthogroup to search against the remaining four genome sequences using tblastn (Altschul *et al.*, 1997) with the standard genetic code table for *Saccharomyces kudriavzevii* and *Saccharomyces kluyveri* and the alternative yeast nuclear code for *Lodderomyces elongisporus* and *Pichia stipitis* In order for an orthogroup to be considered for further analysis I required

it to be complete, i.e. containing all 18 species, as well as sufficiently divergent in order to avoid the possibility of unknowingly analysing paralogous sequences and thereby introducing bias. The last point was addressed by further tblastn searches against the respective protein annotations using the *S. cerevisiae* member of each orthogroup. If each of the respective orthogroup members was found to be the best hit in its genome and its blast score was at least twice that of the next match the orthogroup was considered sufficiently divergent. This filtering step led to a reduction in the number of orthogroups to 629.

Within each orthogroup, the amino acid sequences were mapped to the corresponding genomes and their nucleotide sequences were extracted if I could find an exact match to the respective genome sequence, taking into account the differences in the genetic code in the CTG species. The orthogroups were again filtered by requiring nucleotide sequences for all 18 species to be present, further reducing the set of orthogroups to 357. This represents a large reduction in the number orthogroups and was mainly due to the *Candida albicans* genome release used in this study (Assembly 20) which I later discovered was a superposed "mosaic" haploid assembly of the previous diploid assembly (Assembly 19). A final six orthogroups were removed due to convergence problems in phylogenetic analyses (see below).

Upon release of the second version of FOR in January 2009, which included updated annotations for *Saccharomyces kluyveri* and *Lodderomyces elongisporus*, I re-examined the orthogroups collected for analysis. Overall, I found changes in orthology assignment in 14 orthogroups which were updated accordingly, and a further 11 orthogroups that were no longer in agreement with the conditions outlined above and hence removed from analysis.

The amino acid orthogroups were aligned using Mafft version 6.24 (Katoh & Toh, 2008). Regions of potentially low quality in the alignments were removed using Gblocks version 0.91b (Castresana, 2000), using default parameters apart from the minimum number of sequences for a flank position which was set to 10; the minimum block length, set to 5; and gaps were allowed for half the sequences. It is arguable that trimming alignments in this manner is also discarding potentially useful information and that this is done more or less arbitrarily. While this is true to some extent, I argue that not trimming poorly aligned regions

will more likely than not introduce conflicting signal and that the nature of this conflicting signal is different to the one that I set out to explore. As such this is a reasonable, if not necessary step to take. BLAT (Kent, 2002) was used to map the trimmed alignments to their respective genomic location, to create the nucleotide alignments of the coding sequences used in all further analyses.

These data collection and filtering procedures are considered to be very stringent, and were used in order to avoid introducing any confounding sources of error whilst trying to be as inclusive as possible.

2.2.1 Evolutionary Models

Choosing an appropriate evolutionary model for phylogenetic analysis is of great importance as it is well-known that overly simple models can give misleading results (e.g. Yang *et al.*, 1994, Sullivan & Swofford, 1997). Especially in genome-wide datasets such as the one presented here, where data from diverse regions of the genome are being analysed, between-gene heterogeneities are to be expected in addition to within-gene heterogeneities. I tested an array of combinations of evolutionary models (with varying degrees of complexity) and partitionings of the data, as summarised in Table 2.1. The Jukes-Cantor model (JC; Jukes & Cantor, 1969), being the simplest of all those evolutionary models, assumes equal nucleotide frequencies and no difference in rate between transitions and transversions. The Hasegawa-Kishino-Yano model (HKY; Hasegawa *et al.*, 1985) parameterises the different nucleotide frequencies (π) and includes a rate ratio parameter (κ) for the ratio of the rates of transition and transversion substitutions. Finally, the general time reversible model (REV; Rodríguez *et al.*, 1990; Tavaré, 1986) includes parameters for the different nucleotide frequencies (π), as for HKY, as well as exchangeability parameters (s_{ij}; Goldman & Whelan, 2002) for every possible type of substitution (also referred to as the exchangeability parameters *a*-*f* in the PAML package [Yang, 2009]). Among-site heterogeneity in the rate of evolution was modelled using a gamma distribution (Γ; Yang, 1994).

Apart from testing different evolutionary models, I also investigated the impact on model fit when partitioning the data, allowing some or all parameters to be estimated separately for individual partitions to account for between-partition

heterogeneities. Partitions represented different codon positions, different genes, or both (see following).

Partitioning of codon positions (see Note a in Table 2.1) was carried out using the "Mgene" options implemented in baseml from the PAML package (Yang, 2007). G_0 is the simplest of those options, introducing parameters for different rates at each codon position in the form of branch length scaling factors (c). G_2 and G_3 additionally estimate separate nucleotide frequencies (π^{123}) or separate exchangeability parameters (s_{ij}^{123}) for each codon position, respectively; G_4 estimates separate nucleotide frequencies (π^{123}), exchangeability parameters (s_{ij}^{123}) and different rates (c). Finally G_1, the most general option, estimates separate nucleotide frequencies, rate ratio parameters and different non-proportional branch lengths (bl^{123}) for each position.

Between-gene partitioning was performed with both unpartitioned genes (rows 1 and 2 in note b; Table 2.1), and in addition to partitioning by codon positions (rows 3 to 7 in note b; Table 2.1). Due to the capabilities of available software and the computational cost involved, I could only carry out the partitioning between genes by calculating likelihoods individually for each gene and then summing over the trees tested, treating genes entirely independently: this equates to Mgene option G_1. The most complicated model thus estimates separate branch lengths, nucleotide frequencies and exchangeability parameters for each codon position in every gene as well as a separate α for each gene in the dataset.

2.2.2 Model and tree comparison

The likelihood ratio test (LRT) is widely-used and is utilised to perform hypothesis testing between two nested models (Felsenstein, 2004). The LRT statistic is defined as:

$$\lambda = 2[lnL(\hat{\Theta}_1) - lnL(\hat{\Theta}_0)] \tag{2.1}$$

where $lnL(\hat{\Theta}_1)$ and $lnL(\hat{\Theta}_0)$ are the maximised log-likelihoods of the alternative and the null models with k_1 and k_0 being the number of estimated parameters in Θ_1 and Θ_0, respectively. Under the null hypothesis, this statistic is asymptotically distributed as a χ^2 distribution with $k_1 - k_0$ degrees of freedom.

	Model	Parameters	No. of parameters	
			Single-gene/ Concatenated	Partitioned
Single genes	JC+Γ	bl, α	35	n/a
	HKY+Γ	bl, π, κ, α	39	n/a
	REV+Γ	bl, π, s_{ij}, α	43	n/a
Supermatrix	REV	bl, π, s_{ij}	42	[b] 14406
	REV+Γ	bl, π, s_{ij}, α	43	14749
	REV+Γ+G_0	bl, π, s_{ij}, α, c	[a] 45	15435
	REV+Γ+G_2	bl, π^k, s_{ij}, α, c	51	17493
	REV+Γ+G_3	bl, π, s_{ij}^k, α, c	55	18865
	REV+Γ+G_4	bl, π^k, s_{ij}^k, α, c	64	21952
	REV+Γ+G_1	bl^k, π^k, s_{ij}^k, α	129	44247
	WAG+Γ	bl, π, α	55	18865
	LG+Γ	bl, π, α	55	18865

[a] Partitioning by codon positions

[b] Partitioning by genes

Table 2.1: Evolutionary models used and the number of parameters estimated per model. The number of parameters for partitioned analyses is based on the dataset size of 343 genes, where each parameter is estimated separately for each gene. α is the shape parameter of the gamma distribution, bl are the branch lengths, κ and s_{ij} are the rate ratio parameters and π are the nucleotide frequencies. When the Mgene options in baseml are used, c represents two scaling factors for proportional branch lengths at codon positions 2 and 3 and the "k" superscript indicates that in these models, those parameters are estimated separately for each codon position. The area shaded in orange (note a) indicates partitioning by codon positions, blue (note b) indicates partitioning by genes and purple indicates partitioning by both codon positions and genes.

In addition to LRTs I also calculated the AIC scores for each model. The AIC score has information theoretic foundations and can be used to compare several models, not necessarily nested, at the same time (Posada & Buckley, 2004). Because sample size in my dataset was small in comparison to the number of free parameters ($n/k < 40$) AIC_c, the second order approximation of the AIC score was used for model testing (Burnham & Anderson, 2004). Furthermore, as additional parameters for partitioned models are added for each partition rather than the entire dataset, I calculated the penalty term on a per-partition basis:

$$AIC_c = -2lnL + 2k + \sum_{\text{partitions } i} \frac{2k_i(k_i+1)}{n_i - k_i - 1} \tag{2.2}$$

where lnL is the maximised log-likelihood of the data, k_i is the number of parameters estimated for partition i with $k = \sum_i k_i$, and n_i is the sample size (number of alignment positions) for partition i with $n = \sum_i n_i$. The smaller the AIC_c score, the better the fit of the model to the data.

The AIC score is known to favour parameter-rich models under some conditions while the BIC is generally considered to be more conservative (Burnham & Anderson, 2004; Weaklim, 1999). In order to obtain a conservative estimate of model-fit I additionally calculated BIC scores (Schwarz, 1978) for each of the comparisons carried out. As for the AIC_c, I calculated the penalty term on a per-partition basis. The BIC is then defined as:

$$BIC = -2lnL + \sum_{\text{partitions } i} k_i \log(n_i) \tag{2.3}$$

where lnL is the maximised log-likelihood of the data, k_i is the number of parameters estimated for partition i with $k = \sum_i k_i$, and n_i is the sample size (number of alignment positions) for partition i with $n = \sum_i n_i$.

In order to gauge the heterogeneity inherent in the collected dataset, I examined the distributions of estimates of two parameters. I consider these distributions to be diagnostic of the amount of variation of the evolutionary process experienced by the individual genes. These estimates were taken from the single-gene analyses under the REV + Γ model of evolution, which was found to be

optimal in single-gene analyses (see below). The average transition/transversion (ts/tv) ratio R (Yang, 2006) for a gene is given by:

$$R = \frac{(s_{TC}\pi_T\pi_C + \pi_A\pi_G)}{(s_{TA}\pi_T\pi_A + s_{TG}\pi_T\pi_G + s_{CA}\pi_C\pi_A + s_{CG}\pi_C\pi_G)} \tag{2.4}$$

where π_i are the nucleotide frequencies and s_{ij} are the exchangeability parameters as defined by the REV model (Goldman & Whelan, 2002). Considering the distribution of R over all genes presents a way to display the diversity of the parameterisation of the same model for different genes and hence heterogeneity between them. Similarly α, the shape parameter of the gamma distribution that is used to model rate variation amongst sites (Yang, 1994), is a measure of the within-gene rate heterogeneity of the individual genes and observing its distribution across the single genes gives an indication of the observed between-gene heterogeneity of within-gene rate heterogeneity.

The difference between estimated trees was measured using the normalised version of the Robinson-Foulds (RF) distance (Robinson & Foulds, 1981). Based on the number of bipartitions induced by either tree that are not induced by the other, the normalised version of the RF distance (RF_N) is calculated by dividing this number by the the maximum possible value for the number of sequences in the alignment, $2N - 6$ where N is the number of sequences, so that the RF distances assume a range from 0 to 1. Only identical trees have $\mathrm{RF}_N = 0$; trees with $\mathrm{RF}_N = 1$ are as different as possible.

2.3 Single-gene Analyses — 343 Genes - 336 Trees

The congruence of reconstructed gene trees with the species tree they evolve in and their informativeness for determination of the latter depends on a number of factors including the accuracy of phylogenetic reconstruction (which in turn is known to be influenced by model specification [e.g. Ripplinger & Sullivan, 2008; Sullivan & Swofford, 1997] or stochastic error due to small sample size [i.e. short alignments]) and biological processes leading to gene trees different from the species tree (e.g. independent lineage sorting, horizontal gene transfer or

recombination, reviewed in Degnan & Rosenberg, 2009). The relative influence of these factors is not well known, yet the nature of the signal causing discordance between gene trees and the species tree is important to consider when choosing methods of species tree reconstruction as this might have large effects on downstream analysis, especially for methods that utilise and combine gene tree estimates such as supertree and some of the coalescent-based methods. In order to address some of these issues I explored the amount of variation between gene tree estimates encountered in my phylogenomic dataset and the statistical support for this variation when different tree reconstruction software and models of evolution were used as well as the variation between estimates from different genes.

ML phylogenetic methods were used for all of the analyses presented here due to their power and ability to explicitly model different evolutionary patterns both within and between genes (Yang, 2006). Gene trees were built using Leaphy 1.0 (Whelan, 2007) and PhyML 3.0 (Guindon & Gascuel, 2003), both ML programs but employing different tree searching strategies. Leaphy utilises a sophisticated tree searching strategy implementing algorithmic improvements in the refinement and resampling stages of tree search and novel stopping rules. PhyML was run using both the SPR and NNI algorithms. I estimated the trees using the JC, HKY and REV nucleotide models of evolution (Table 2.1, "Single genes" rows). Rate heterogeneity among sequence positions was modelled using a discrete approximation of a gamma distribution (Γ; Yang, 1994) with six rate categories. In order to assess confidence in the individual nodes of the tree I performed non-parametric bootstrap analyses with 100 replicates each (Felsenstein, 1985).

The baseml program (Yang, 2007) was subsequently used to evaluate the efficiency of the two tree-searching programs with respect to each other and their overall performance. Because Leaphy and PhyML differ slightly in their algorithms to calculate the likelihood of a tree, likelihoods calculated using baseml were used to provide a standardised measure of fit. The direct comparison was carried out by evaluating on a gene-by-gene basis which of the trees produced by Leaphy or PhyML was of higher likelihood. The overall performance would ideally have been addressed by scoring the ability of the respective softwares to find the highest likelihood gene tree. Since baseml does not implement sophisticated

tree building methods, guaranteeing finding this tree would require baseml to be presented with the entire set of possible topologies. However, since there are 1.92×10^{17} possible unrooted trees for 18 species this was not feasible. Instead, I considered all trees ever inferred by any of the single-gene analyses in this study, totalling 1668 different topologies, as a reasonable snapshot of the "tree space" of feasible single-gene trees and used this as the reference tree set for baseml. This reference tree set is referred to as "candidate tree set 1" (CTS_1).

2.3.1 Large Amounts of Variation Between Tree Reconstruction Methods

Phylogenetic analysis of the 343 nucleotide single-gene data sets was performed using the JC + Γ, HKY + Γ and REV + Γ models of evolution and using both PhyML and Leaphy. PhyML analyses were run using both implemented tree searching algorithms, NNI and SPR, and the best of those trees used. I found that Leaphy generally outperformed PhyML in finding the best tree (Table 2.2). (To ensure that those results were not biased in favour of Leaphy because of similarities in the likelihood calculation procedure I repeated this comparison using PhyML to obtain comparative likelihoods. This gave nearly identical results [not shown]). When the ML trees obtained by PhyML and Leaphy were compared to each other on a gene-by-gene basis, both methods found the same tree for approximately 70% to 75% of genes. For the remaining 25% to 30%, found uniquely by either software, Leaphy clearly showed a better performance. Overall Leaphy found the best tree for around 95% of the genes tested, compared to 75% found by PhyML (see Table 2.2, "Single-gene trees"). (Interestingly, PhyML's SPR tree search, generally considered better, did not always outperform NNI. Including SPR however considerably improved the accuracy of PhyML in this comparison compared to using NNI alone [results not shown].)

This comparison only showed how frequently PhyML and Leaphy equal or exceed each other at finding high likelihood trees. In order to gauge their performance with respect to finding ML trees, but in the absence of knowledge of the correct ML tree for each single-gene dataset, I instead used baseml to calculate

	Single gene trees					CTS_1				
Model	No. genes[a]	PhyML alone	Leaphy alone	SPR >= NNI	Both	No. genes[a]	PhyML alone	Leaphy alone	Both	Neither
JC	343	22(6.4%)	82(23.9%)	239(69.7%)	308(89.8%)	343	22(6.4%)	85(24.8%)	232(67.4%)	3(0.9%)
HKY	343	8(2.3%)	83(24.2%)	252(73.5%)	293(85.4%)	343	14(4.8%)	84(24.5%)	240(70.0%)	5(1.5%)
REV	343	15(4.3%)	74(21.6%)	254(74.1%)	304(88.6%)	342	8(2.3%)	71(20.8%)	245(71.6%)	18(5.2%)

[a] Number of genes where ML analysis converged

Table 2.2: Performance of PhyML and Leaphy when compared directly or with respect to a larger set of plausible trees. The numbers indicate how often the ML topology for a given gene found by PhyML and Leaphy respectively coincided with the ML topology determined using baseml to assess either just the PhyML and Leaphy ML trees ("Single-gene trees") or the CTS_1 ("CTS_1"; see Methods).

the ML trees for each of the genes from amongst the entire CTS_1 set of plausible trees (see above) and determined how often PhyML and Leaphy found the best tree. This procedure guaranteed that the best trees found by PhyML and Leaphy for a particular gene were compared to many other potentially superior trees, namely those found in any single-gene analysis performed in this study (since an exhaustive comparison against all trees was not feasible). Overall, the outcome was similar to that when PhyML and Leaphy results were compared directly (see Table 2.2, "CTS_1"). PhyML found the best tree for around 75% of the genes whereas Leaphy found the best tree for 90% to 94% of the genes. This showed that Leaphy not only outperformed PhyML in direct comparison, but also performed very well overall. I was thus confident to use Leaphy alone for all subsequent analyses.

Such differences between tree reconstruction methods can have important implications for understanding evolutionary relationships and in downstream analyses. While both programs tested here agree on a ML topology for the majority of alignments, about 25% to 30% of proposed gene trees differ for any given model of evolution. To investigate possible causes for the disagreement between PhyML and Leaphy, I examined whether convergence on the same ML tree was associated with alignment length, as a proxy for the amount of information available for tree estimation (Figure 2.2). If all variability was due to the lack of strong phylogenetic signal, we would expect a bias for short alignments to yield disagreeing gene tree estimates. Indeed, there was a weak, albeit significant, association between convergence and alignment length with genes where Leaphy and PhyML disagree (Mann-Whitney U test; $P < 0.01$). Nevertheless, I found numerous short alignments where both methods agree and long alignments where both methods disagree, suggesting that some but not all discrepancies might be attributable to the lack of strong signal. Arguably, there are situations where alignment length will be a poor proxy for the amount of information available, e.g. when a large number of alignment positions contain identical residues in all sequences. The choice of a measure for column diversity however is non-trivial and the search for an appropriate measure was not attempted here.

In order to explore how far-reaching the discrepancies between PhyML and Leaphy were, I investigated the statistical support for the differences between

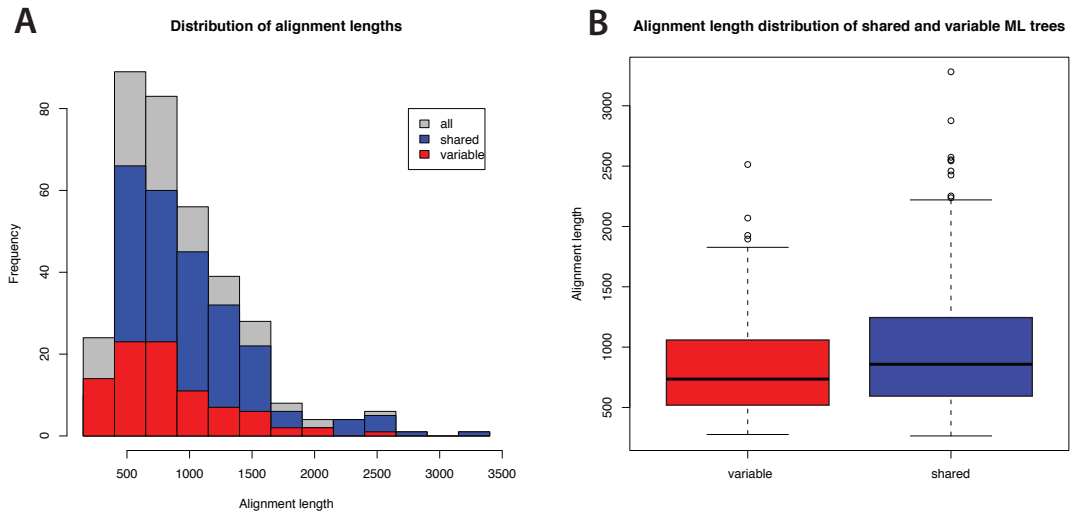


Figure 2.2: Distributions of the lengths of alignments yielding congruent "shared" (blue) and incongruent "variable" (red) ML topologies when analysed using PhyML and Leaphy and the REV + Γ.

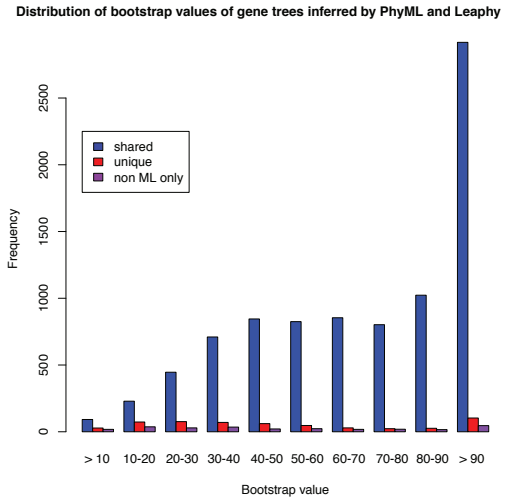


Figure 2.3: Distributions of the bootstrap values of nodes that were shared (blue) or unique (red) between the respective ML topologies proposed by PhyML and Leaphy for each of the 343 genes. "non-ML only" shows the distribution of unique nodes in the gene tree that was found to be of lower likelihood only.

the alternative topologies generated by the two different programs for each gene. For this, I focussed on REV + Γ results only (this being the best model for all 343 genes studied: see below). This was done by examining the bootstrap values obtained for conflicting nodes in the two ML trees proposed for a gene by PhyML and Leaphy. Figure 2.3 shows the distribution of the non-parametric bootstrap values for shared (red) and variable (blue) nodes between ML trees proposed by either Leaphy or PhyML using 100 replicates, and also the distribution of bootstrap values when only "inferior" ML trees (as determined by baseml) were considered (purple). Averaged over genes, the trees reconstructed by each program share about 80% of internal branches. Generally, bootstrap values were low for branches that differed between the alternative ML trees. There was, however, some overlap of the distributions of bootstrap values for branches that differed between trees and branches that were shared, and a number of cases where bootstrap support for regions of conflict between the two topologies was over 70%, often a boundary above which nodes are considered reliable and included for downstream analyses. Removing the trees that were found non-optimal in pairwise comparison did not visibly alter the distribution of bootstrap values, showing that the high bootstrap values for shared nodes occur in both "optimal" and "non-optimal" trees. The use of good tree searching methods such as implemented in Leaphy is thus vital for circumventing such otherwise undetectable non-optimal inferences.

2.3.2 The Influence of Model Choice on Gene Tree Reconstruction

To investigate the impact of model choice on tree reconstruction I analysed the 343 gene nucleotide dataset using the JC + Γ, HKY + Γ and REV + Γ models of evolution. Here, only results for models including gamma-distributed intragene rate heterogeneity are presented since these models were always significantly preferred (results not shown). When I examined how often the ML trees for a gene using the different models were identical, topologies recovered using the HKY + Γ and REV + Γ models were the same for about 47% of the genes. In contrast, the number of identical topologies recovered when REV + Γ results were

compared to JC + Γ results was much smaller, with the same ML tree obtained for just 8% of genes. Similarly, JC + Γ and HKY + Γ analyses resulted in 9% of shared trees. This discrepancy was also detected when I investigated the degree of differences between the trees recovered by the different models for a given gene. On average, the mean normalised RF distance (RF_N) between the REV + Γ and the HKY + Γ topologies was 0.08 whereas mean RF_N between REV + Γ and JC topologies was 0.27.

As above, I examined the bootstrap distributions for ML trees reconstructed using different models of evolution to determine the statistical support for the discrepancies between alternative topologies proposed for the same gene. Gene trees were reconstructed using Leaphy with 100 bootstrap replicates each and the JC + Γ, HKY + Γ and REV + Γ models of evolution. Figure 2.4 shows the distributions of shared (red) and variable (blue) nodes in pairwise comparison between models. The support for discrepancies between JC + Γ, the simplest and most unrealistic model used, and the two more complex models is mostly below 70%, but there is however a considerable amount of well-supported conflict (Fig. 2.4A, B). The comparison between HKY + Γ and REV + Γ shows fewer conflicting nodes overall and only very little highly-supported conflict, indicating an improvement in model fit (Fig. 2.4C). The increase in congruence and the decrease in well-supported conflict when more complex models are used both show that choosing a better model improves results and that it is important to choose a model that best fits the data.

2.3.3 Best-fit Models

Hierarchical LRTs were used to determine the optimal model for each gene. HKY + Γ was always found to be better than JC + Γ and REV + Γ was found to be the best-fitting model for all but one of the 343 genes studied, where HKY + Γ was favoured instead. I consequently focussed on REV + Γ as the model for analysis of the nucleotide data, omitting simpler models of evolution. Nevertheless, even with the best available models there remain discrepancies that can affect downstream analyses, as is shown next.

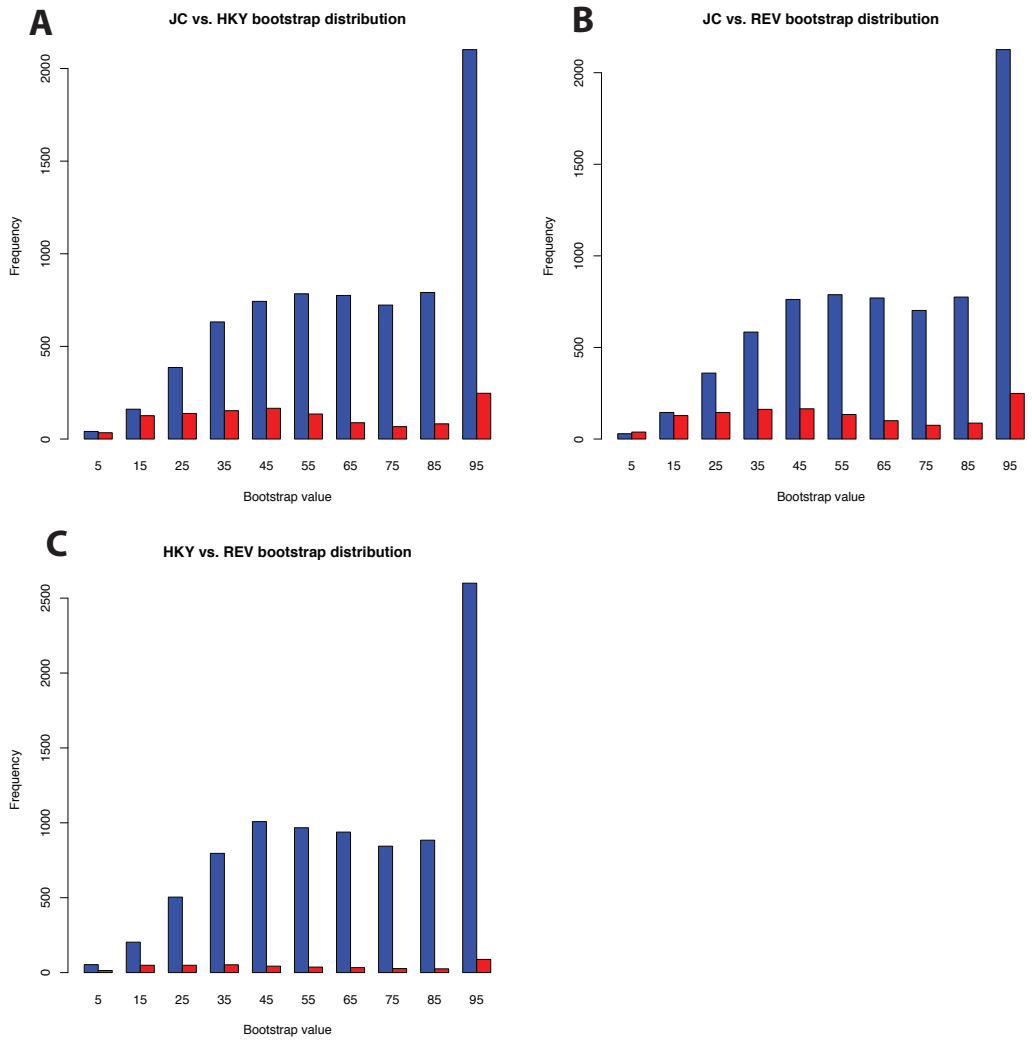


Figure 2.4: Distributions of the bootstrap values of nodes that were shared (blue) or unique (red) between the respective ML topologies proposed by Leaphy for each of the 343 genes, using different models of evolution.

2.3.4 Large Amounts of Incongruence Among the Single-gene Datasets

I will now concentrate on the results of analysing the single-gene datasets using the optimal model found for each gene (see above). I found very large amounts of incongruence within the set of ML trees recovered. In total I obtained 336 distinct ML topologies for the 343 genes: in other words, a different phylogeny of the 18 yeasts for almost every individual gene. The mean pairwise RF_N distance among the 336 topologies was 0.54. So whilst they were clearly more similar to each other than 336 randomly drawn trees of the same size ($P < 0.0001$ from 10000 simulations; mean RF_N distance = 0.93, std. dev. = 0.0006), it still meant that on average the gene trees for any two genes differed by 16 unique bipartitions. Pruning of all species not included by Rokas *et al.* (2003) from the 336 trees resulted in 49 distinct trees with a mean RF_N distance of 0.55 ($P < 0.0001$ from 10000 simulations; mean = 0.93, std. dev. = 0.0039) indicating similar levels of divergence.

The analysis of single genes thus proved inconclusive in this case and I was unable to derive a species phylogeny supported by the individual gene phylogenies. Incongruence among gene trees is not specific to this dataset but instead is found in a large fraction of studies comparing single-gene phylogenies (Rokas & Chatzimanolis, 2008). Even though it is as yet unclear how much incongruence between gene trees in a genome is "normal", the level of incongruence encountered here (i.e. 336 phylogenies from 343 genes) was surprising. In order to confidently resolve inter-species relationships there was thus the need for a way of combining the data in a sensible manner. Supermatrix analysis promised resolution of conflict where individual analyses failed.

To illustrate the heterogeneity across the single-gene datasets and thereby speculate about the validity of a concatenation approach for further analysis I investigated the transition/transversion (ts/tv) ratio for each of the genes (Fig. 2.5A). I also examined the distribution of α, the shape parameter of the Γ distribution which indicates the extent of among-gene rate variation for a given gene (Figure 2.5B). Each of these measures demonstrates a potential source of inter-gene

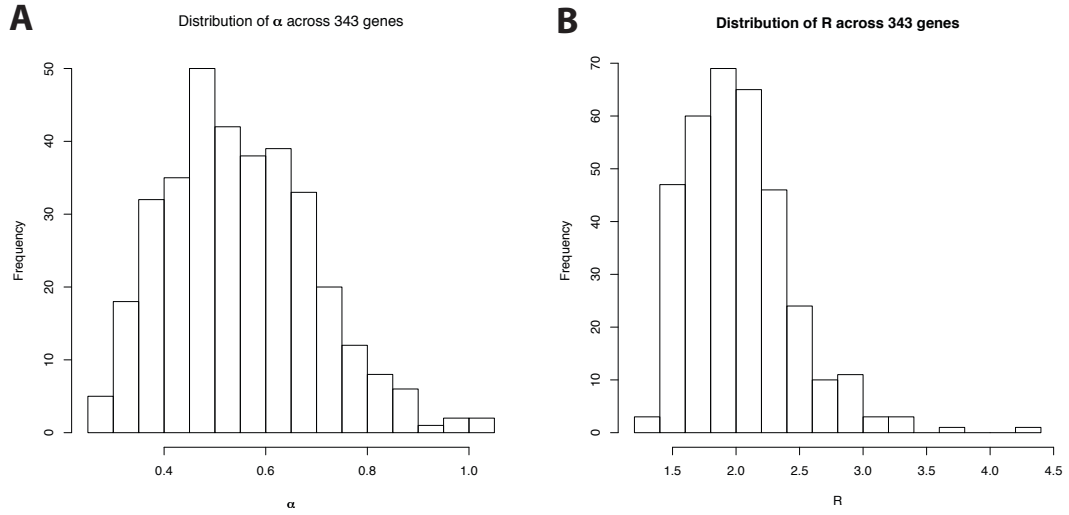


Figure 2.5: Distributions of **A:** ts/tv ratio (R) and **B:** gamma-distribution shape parameter α, estimated on the ML topology for each of 343 genes

variation in evolutionary dynamics. I used the maximised parameter estimates of each gene under the REV + Γ model of evolution as calculated by baseml.

The range of ts/tv ratios was large, extending from 1.30 to 4.22, and indicated that there was a lot of heterogeneity in the evolutionary process between the individual genes. These values were similar to previously compiled distributions of ts/tv ratios over a range of different genes (Whelan *et al.*, 2003). Similarly, α ranged from 0.25 to 1.03, showing that the distributions of rates across sites of the different genes were diverse. Again, this between-gene variation in ts/tv rates and α could have been the result of poor parameter estimates due to lack of phylogenetic signal in short alignments. In order to assess whether the more extreme values of those distributions were the result of noisy parameter estimation I examined whether the distance from the mean of their respective distributions for each estimate was associated with alignment length and the standard error (SE) of the parameter estimates (Fig. 2.6A and B for α and R [ts/tv] respectively; SE estimates were available for α only, as the calculation of SEs for ts/tv is complex due to the incorporation of multiple parameter estimates and requires information about covariance between those estimates for accurate calculation, which is not included in the standard PAML output).

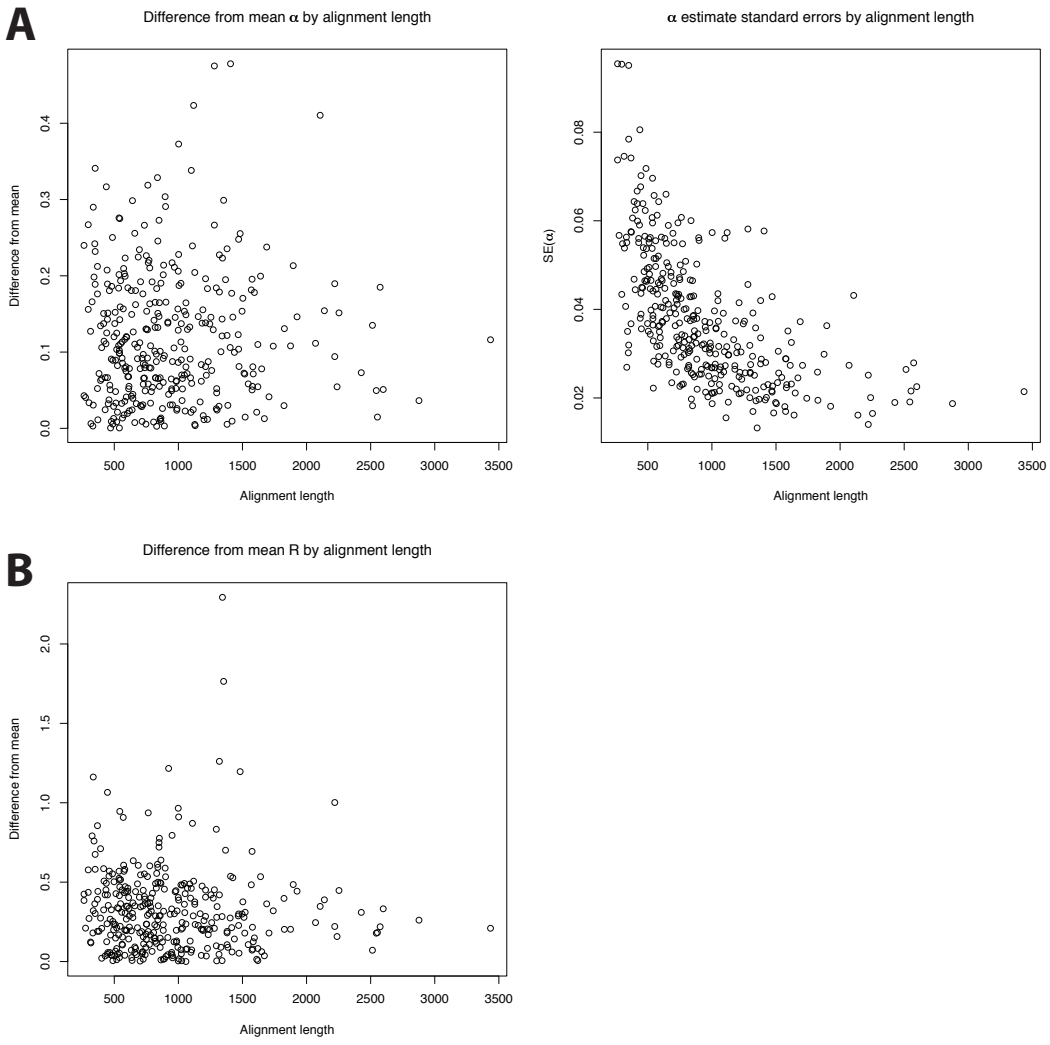


Figure 2.6: **A**:Distributions of parameter estimates and standard error of α, when considered by alignment length. **B**:Distributions of parameter estimates of R, when considered by alignment length. Values are plotted as the distance from the population mean.

The distribution of the standard error of α (Fig. 2.6A, right plot) showed that there was a clear relationship between gene length and the accuracy of parameter estimation, as would be expected. In contrast, the distribution of the parameter itself (Fig. 2.6A, left plot), although arguably somewhat associated with gene length, showed numerous estimates close to the population mean for very short alignments and, *vice versa*, plenty of estimates on long alignments that were far from the mean. This underlined that the variation of α that I encountered across the 343 genes was not solely due to accuracy of parameter estimates but was capturing the heterogeneity inherent in the data. Similarly, the difference from the mean estimate of R was weakly associated with alignment length but again the shape of the distribution suggested that alignment length alone is not explaining the variation encountered.

The number of distinct topologies recovered in single-gene analyses underlined the need for phylogenomic methods to distil the shared "historical" signal between the genes analysed. At the same time, the level of heterogeneity encountered between the genes when examining two key parameters of the optimal evolutionary model suggested that a simple concatenation supermatrix approach might not be valid and complex models were needed to analyse a dataset containing this much variation.

2.4 Supermatrix Analysis

I chose baseml (Yang, 2007) to perform analyses of the supermatrix dataset due to the wealth of models it implements. However, as mentioned above, baseml is best used with a predefined set of trees for input. To obtain such a plausible set of trees, I compiled a second candidate tree set (CTS_2) based on prior knowledge about the evolutionary relationships in some parts of the phylogeny. The three major clades shaded in different colours in Figure 2.1A are stably recovered in this configuration and in addition to the evidence from sequence analysis, other features of these genomes support these groupings. Namely those are the "post-WGD" species that underwent a whole genome duplication (green) and show a 2:1 syntenic relationship to their close relatives the "pre-WGD" species (blue; Byrne & Wolfe, 2006) and finally the lineage leading to the *Candida* clade that

translate the CTG codon into a Serine instead of a Leucine (red; Santos & Tuite, 1995; Sugita & Nakase, 1999).

Figure 2.1A shows the "backbone" species tree of the 18 yeasts I studied. Inter-species relationships that are well-accepted are shown fully resolved, whereas regions of uncertainty in the phylogeny are collapsed into polytomies. The CTS_2 initially included all trees found by resolving all the polytomies shown in the tree into all possible arrangements. In order to keep the number of topologies manageable, I first resolved the relationship between *Candida glabrata* and *Saccharomyces castellii* as shown in Figure 2.1B, resulting in 1575 remaining topologies. I then ran an initial analysis using the partitioned REV + Γ + G_0 model (Table 1.1) that, whilst computationally very feasible, was found to be a good model for supermatrix analysis of nucleotide data (as described below). From this analysis I identified the 500 best-scoring topologies. Those were subsequently modified so that there would be a copy of each of those topologies, but incorporating either of the alternative resolutions in Figure 2.1C and D, totalling 1000 additional trees. Based on further ML analysis using the same model, I excluded all instances of *C. glabrata* and *S. castellii* branching as sister species (Figure 2.1D) due to consistently very low likelihoods for those trees (results not shown). The remaining 500 novel trees were added to CTS_2, now containing 2075 topologies.

Both nucleotide and amino acid analyses were carried out on a concatenated as well as a partitioned form of the data using the CTS_2 and the evolutionary models described in Table 2.1 ("Supermatrix" rows). Nucleotide analyses were performed using a range of nested models, each increase in complexity allowing for more amongst-genes heterogeneity. Amino acid analyses were carried out using the WAG and LG models of evolution (Le & Gascuel, 2008; Whelan & Goldman, 2001) with a gamma distribution (again using six rate categories) to allow for different rates at different positions. Partitioned analysis was achieved by analysing each gene separately (see Evolutionary Models, above) and total likelihoods thus calculated for each member of the CTS_2 to determine the ML tree. Bootstrap analyses were performed using the RELL method (Kishino & Hasegawa, 1989) with 1000 samples for each of the above analyses.

2.4.1 Complex Data Require Complex Models

Supermatrix analyses were performed using a range of different models and partitions (Tables 2.1, 2.3; see above). The basic model of evolution used was REV, as found to be the most appropriate in single-gene analyses (see above). Building on this, I introduced gamma-distributed rates across sites (+ Γ) and partitioning into either codon positions (Mgene options), genes or both, where some or all parameters are estimated separately for each partition thus allowing for heterogeneity between them. The optimal model was determined according to the AIC_c and BIC criterion as well as LRTs where applicable. The order in which LRTs for the different Mgene options were performed was as follows; G_0 - (G_2,G_3) - G_4 - G_1. G_2 and G_3 are not nested and thus could not be tested against each other in an LRT. The results are shown in Table 2.3. The most comprehensive model, partitioned REV + Γ + G_1, was found to be optimal for the supermatrix by both AIC_c and LRT — reflecting the complexity of the signal encoded in these data.

BIC results for the nucleotide models were very similar to the ones obtained using LRTs or AIC_c (see Table 2.3 and Figures 2.7 and 2.8). All three tests support the use of complex models, treating each codon position separately. Furthermore, partitioned models were always preferred over concatenated models showing that even a conservative test supports partitioning despite the large number of additional parameters. The BIC, however, selected REV + Γ + G_4 as the best model as opposed to REV + Γ + G_1 selected by the other two tests. Examination of the AIC_c results in Figure 2.7 shows that the gain of information with respect to the number of parameters added between those models is not as great as for other comparisons, suggesting that the model is nearing an optimal level of complexity.

Model testing and the choice of an appropriate model proved to be important in this case, because the ML tree again changed depending on which model was used (Table 2.3). Both different rates across sites (+ Γ) as well as different rates at each codon position (+ G_0) influenced which tree was found to be the ML tree. Interestingly, as more parameters were free to vary at different codon positions the likelihood increased significantly but the ML topology remained the same, reinforcing my confidence in having found a good species tree (see Table 2.3).

	Concatenated				Partitioned			
Model	**ML tree**	**ΔAIC**	**ΔBIC**	**P(LRT)**	**ML tree**	**ΔAIC**	**ΔBIC**	**P(LRT)**
REV	A	990359	886041	-	A	872007	832326	- (0*)
REV+Γ	B	431437	327128	0*	B	315551	277374	0*(0*)
REV+Γ+G_0	C	283288	179000	0*	C	155540	120364	0*(0*)
REV+Γ+G_2	C	279792	175568	0*	C	146131	119900	0*(0*)
REV+Γ+G_3	C	162718	58536	0*	C	30691	10374	0*(0*)
REV+Γ+G_4	C	144273	40188	0*/ 0*[a]	C	7156	0	0*/ 0*(0*)[a]
REV+Γ+G_1	C	123756	20364	0*	C	0	80015	0*(0*)
WAG+Γ	C	74186	21951	-	C	25394	27007	- (0*)
LG+Γ	C	52235	0	-	C	0	1612	- (0*)

[a] Models with Mgene option G_4 were tested against both G_2 and G_3 models

Table 2.3: ML trees and test statistics for model tests performed on the supermatrix dataset. Models used are as in Table 2.1. LRTs were performed between the model considered and the next-smallest nested model. ΔAIC is the difference in AIC_c between a model and the best-fitting model; ΔBIC accordingly. Partitioned models were also tested against their concatenated version (in brackets). Significant P-values (< 0.001) are indicated by a star. Trees A and B are shown in Figure 2.10; tree C as in Fig. 2.9.

It is also noteworthy that the bootstrap support across the ML trees recovered using the overly-simple REV and REV + Γ models (see below) was high and thus was not acting as a reliable indicator when trying to assess the confidence for trees recovered in my analyses. In general, large amounts of data can give over-confidence in the result obtained, notably when there is model mis-specification and, in particular, in cases of over-simple models (Yang *et al.*, 1994). Great care should be taken to find the best available model for any given dataset.

2.4.2 Partitioned Analysis Outperforms Concatenated Analysis

All models tested were implemented both using a conventional concatenation approach where a single parameterisation of the evolutionary model is used to analyse the entire dataset, as well as more sophisticated partitioned analysis where parameters were estimated for each partition (in this case genes), thus allowing for more amongst-genes heterogeneity.

Partitioned analysis consistently outperformed concatenated analysis independently of the model of evolution used (Figures 2.7 and 2.8; Table 2.3). The ΔAIC between partitioned and concatenated analyses was substantial for all models, representing a major improvement in model fit. In contrast to results obtained with a mammalian dataset (Nishihara *et al.*, 2007), the use of partitioned versus concatenated analysis had no effect on which topology was found to be optimal in the nucleotide analyses — a poor choice of model (e.g. REV or REV + Γ) still lead to a sub-optimal tree, even if partitioned analysis was used.

2.4.3 Amino Acid Analyses

Nucleotide data are more prone to mutational saturation than amino acid data, and this can present a source of stochastic error. Also, the effects of base composition variation between different genome sequences are likely to be less pronounced in amino acid data. Therefore, in addition to analysing the nucleotide dataset, I also investigated supermatrix analyses on the translated amino acid sequences. I used the WAG + Γ and LG + Γ models (six discrete gamma rate categories) to perform concatenated and partitioned analyses of CTS_2 using the codeml program from the PAML package (Yang, 2007). Again I found the partitioned model to outperform the concatenated model when tested with AIC_c (Figure 2.7). Furthermore, the ML topology calculated by partitioned analysis was found to be identical to the one obtained using the best nucleotide model (Fig. 2.9), reinforcing the results obtained with the nucleotide dataset.

Results obtained for amino acid analysis using BIC differed, in that BIC selected concatenated models over partitioned ones. The interpretation of this is somewhat unclear, seeing that the AIC_c is often considered too liberal and the BIC as too conservative, and further study is needed to determine which of the tests applied here is most appropriate for these kind of data. Nevertheless, in this instance the choice of optimal tree was not affected.

2.4.4 Species Phylogeny of 18 Ascomycetous Yeasts

The species phylogeny that was obtained using supermatrix analysis is shown in Figure 2.9. It is in good agreement with previously published studies for clades

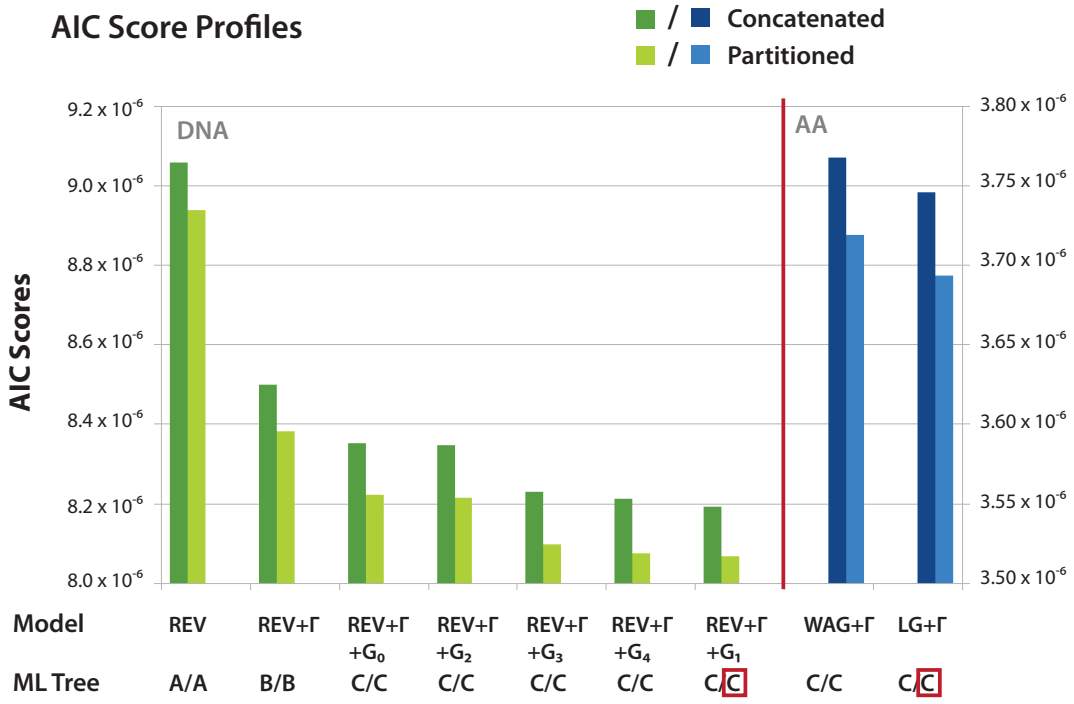


Figure 2.7: AIC_c score profiles for supermatrix analyses, differentiated by evolutionary model and type of analysis. Partitioned analysis (light colours) consistently outperformed concatenated analysis (dark colours). The choice of a partitioned vs. concatenated model did not affect which tree was found to be optimal when analysing nucleotide data (green) as well as amino acids (blue). The ML topology obtained using the optimal model for both nucleotide and amino acid data is the same (red boxes) and is depicted in Figure 2.9. Trees A and B are depicted in Figure 2.10

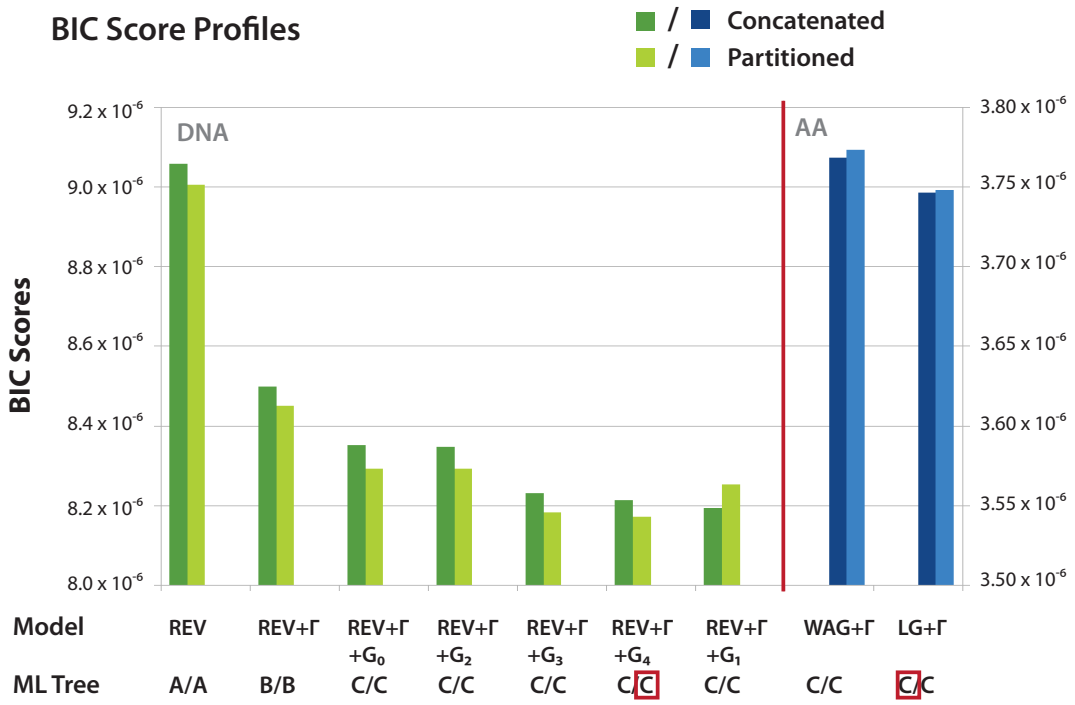


Figure 2.8: BIC score profiles for supermatrix analyses. The results of BIC testing were very similar to results obtained using AIC_c ,with partitioned analysis (light colours) consistently outperforming concatenated analysis (dark colours) for the nucleotide data. The tests however differed in their preferred models which was found to be REV + Γ + G_4. In the analysis of amino acid data (blue), AIC_c and BIC differed however, with BIC favouring concatenated over partitioned models (discussed in the main text).

that were shared between them (e.g. Fitzpatrick *et al.*, 2006; Jeffroy *et al.*, 2006; Phillips *et al.*, 2004; Rokas *et al.*, 2003) and received good bootstrap support across the tree. The pre-WGD species were recovered with high bootstrap support as a monophyletic clade where *Saccharomyces kluyveri* and *Kluyveromyces waltii*, and *Ashbya gossypii* and *K. lactis*, respectively, are sister species that form sister clades to one other. This agrees with what has been found in the more recent studies (Fitzpatrick *et al.*, 2006; Jeffroy *et al.*, 2006). I inferred *Pichia stipitis* to be basal to the *Candida* species and *Lodderomyces elongisporus*, albeit with somewhat lower bootstrap support suggesting the presence of either conflicting signal (e.g. due to incomplete lineage sorting) or stochastic effects influencing the resolution of this node.

The branching order at the base of the WGD clade in the ML tree recovered saw *C. glabrata* splitting off before *S. castellii* and is thus in disagreement with the branching order suggested by synteny data (Scannell *et al.*, 2006). When I examined the support for either of three possible branching orders (*C. glabrata* first and *S. castellii* second; *S. castellii* first and *C. glabrata* second; and *S. castellii* and *C. glabrata* as sister species) in the single-gene dataset (REV + Γ) I found the first branching order to be most strongly represented amongst the genes trees (150 occurrences) in relation to the other two branchings (66 and 58 occurrences, respectively). The study of substitution patterns hence consistently suggested *C. glabrata* as the species at the base of the WGD clade. Disagreement with synteny information could be due to the fact that, as yet, we have limited knowledge of how to statistically and reliably estimate phylogenies from chromosome rearrangement data. Alternatively, it might just as well be a result of biological processes such as independent lineage sorting, especially seeing that the divergence time between *S. castellii* and *C. glabrata* is relatively short (Scannell *et al.*, 2006). Because the data in support of a branching order where *S. castellii* is basal to the remaining post-WGD species is very strong, I settled for this resolution in the species phylogeny for use in downstream analyses described in the following chapters.

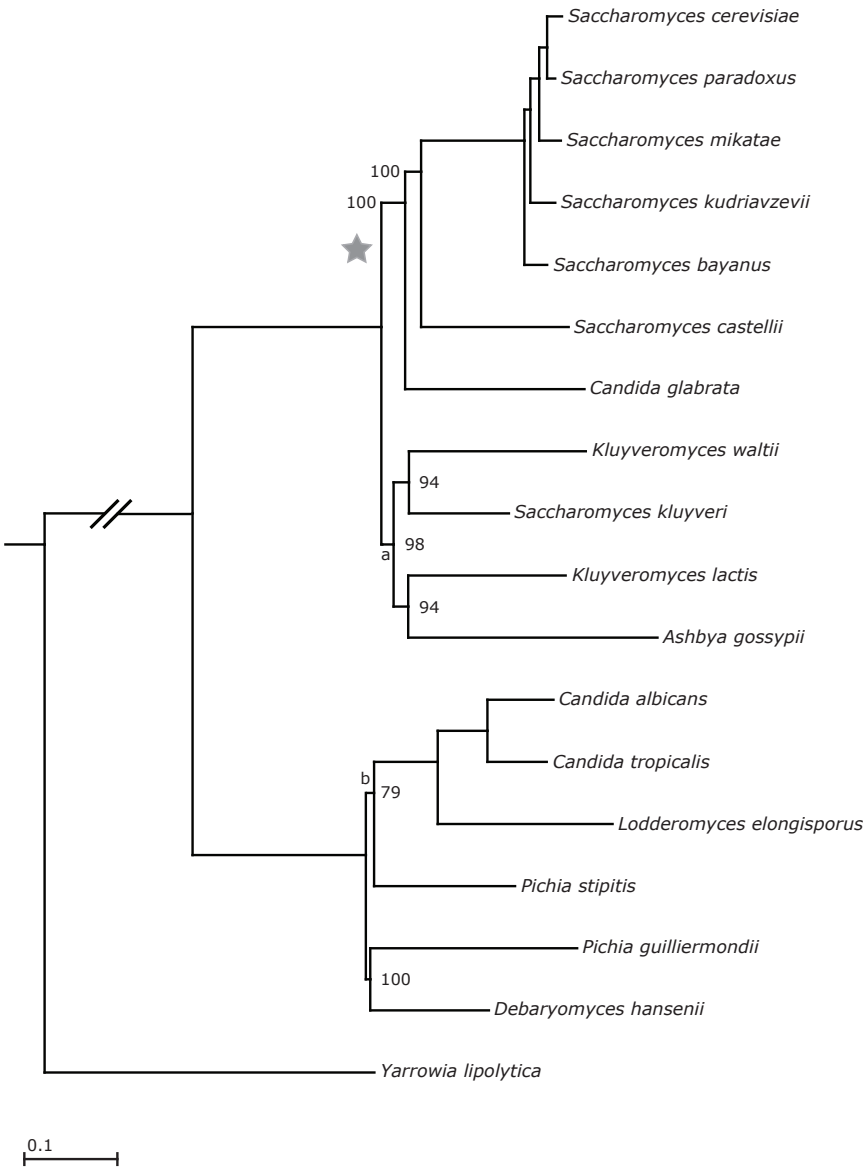


Figure 2.9: ML tree obtained using the optimal nucleotide and amino acid models of evolution. Bootstrap values were calculated using 1000 iterations of RELL resampling. Branch lengths, in expected number of substitutions per nucleotide, were calculated as the weighted mean of individual estimates in partitioned analysis of the 343 genes of the nucleotide datasets. The branches marked by lowercase letters were extended for the purpose of visualisation. The WGD event is marked by star.

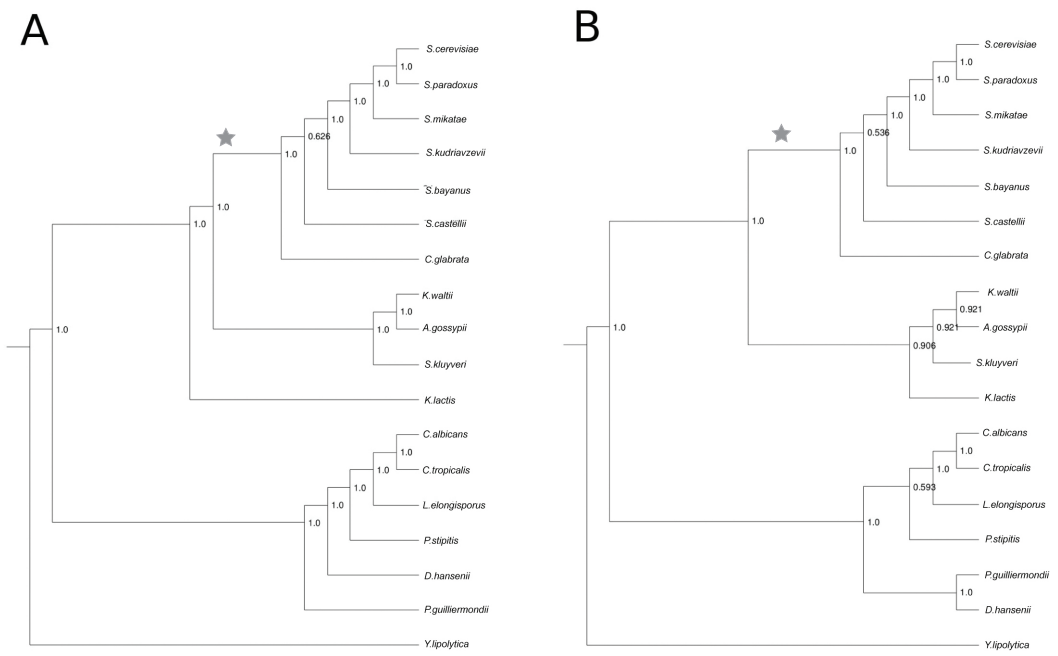


Figure 2.10: ML trees recovered using non-optimal substitution models. As above, bootstrap values were calculated using 1000 iterations of RELL resampling and are indicated as the proportion of the total number of samples supporting this node. The WGD event is marked by a star.

2.5 Conclusions

The power of phylogenetic reconstruction is heavily dependent on the evolutionary models being utilised. This is already well-known (e.g. Ripplinger & Sullivan, 2008; Sullivan & Swofford, 1997; Yang *et al.*, 1994) and it was thus not surprising to find model choice having a large impact in phylogenomics. Furthermore, I also found the methods used to search tree space to be of considerable importance. Leaphy, a relatively new software, outperformed the more established software PhyML both in direct and more comprehensive comparisons and is hence the preferred method for tree reconstruction.

My datasets deliver another example where single-gene phylogenetics failed to find a congruent solution when trying to resolve the species tree that underlies the evolution of those genes. In fact, in this difficult phylogenetic problem the number of proposed ML trees increased almost linearly to the number of genes studied. Phylogenomics approaches are well-suited to address such questions and we would like to see them deliver resolution even as datasets increase in size and diversity. This has been demonstrated in the past, largely on smaller and easier studies (e.g. Diezmann *et al.*, 2004; Kurtzman & Robnett, 2003; Rokas *et al.*, 2003). Given my results, it appeared that this remained true for a dataset of my size and complexity, but with the qualification that it is vital that analyses appropriate to the complexity of the data are used.

Whilst in my nucleotide analyses the identity of the ML topology was not affected by the choice of a partitioned over a concatenated model, it was however highly dependent on the evolutionary model employed, resulting in a different ML tree for all three types of model (no heterogeneity; among-site heterogeneity; and among-site heterogeneity plus individual treatment of codon positions) used. This indicated that the heterogeneity of the evolutionary process affecting substitution patterns within a single protein-coding gene resulted in stronger (non-phylogenetic) signal than the differences between such processes acting on different loci across the genomes of 18 yeast species. While this was surprising, it possibly reflects the biological conditions governing the yeast nuclear environment.

I found a considerable improvement in likelihood when the data were partitioned on a gene by gene basis. This confirmed the significance of inter-gene variation of evolutionary dynamics. This in turn argues for the importance of using appropriate models when analysing complex datasets, as do the results obtained from amino acid analyses. This conclusion stands in contrast to the conclusions drawn by Rokas *et al.* (2003) who suggested simple concatenation of genes would be sufficient for the resolution of species trees.

The use of different model testing frameworks (LRT, AIC_c and BIC) furthermore highlighted interesting questions about the interpretation of results from different tests. While all three agreed on a complex, partitioned model for the nucleotide dataset, results for the amino acid data were conflicting. Indeed, model testing in phylogenomics is often performed using a "black box" approach and there currently is a lack of discussion of which tests are most appropriate to use, especially in the case where both the number of observations and the number of parameters estimated are large and composite pieces of data are combined such as is the case in partitioned analysis. A more thorough discussion of this subject is needed, that would benefit from and be of benefit to other areas of quantitative biology dealing with similar-sized datasets and models.

I obtained a fully-resolved and well-supported species tree for 18 ascomycetous yeasts. This placed *Pichia stipitis* at the bottom of the branch containing the *Candida* species. Furthermore, I was able to confirm the relationships in the pre-WGD species. The analysis presented here suggested *C. glabrata* to be at the base of the WGD clade. This is in contrast with the branching order inferred from synteny data and future work is needed in order to resolve this conflict. Recent developments in coalescent-based methods such as the ones reviewed in Degnan & Rosenberg (2009) provide a promising perspective of resolving short internal branches and will probably provide valuable complementary results to the classic phylogenetic methods such as the ones used in this study. Nevertheless, overall the study presented here lent confidence to the notion that phylogenomics methods, given the right evolutionary models, can give robust answers to a number of yet-to-be-resolved branches of the Tree of Life and has provided a trustworthy phylogeny for downstream analyses.

Chapter 3

Transcription Factor Repertoires in the *Saccharomycotina*

3.1 Introduction

Transcription factors (TFs) are the main protein-coding components of the transcriptional machinery acting in a gene-specific manner (see Chapter 1.2 for a review of transcriptional initiation and the players involved). They are characterised by the presence of a sequence-specific DNA-binding domain (DBD) and often contain numerous additional regions mediating a TF's activity, including protein-protein interaction (PPI) domains, ligand-binding regions, activation domains or sites conferring post-translational regulatory control such as nuclear import or export signals and phosphorylation sites. To date, ∼140 distinct DBDs have been characterised (Charoensawan *et al.*, 2010a,b) falling into several broad structural classes and differ in their DNA-binding modes (for a brief overview see Chapter 1, Table 1.1). Genome-wide analyses of TF repertoires in diverse species have revealed that the number of TFs encoded in a genome scales approximately with genome size, with a tendency for larger genomes and more complex organisms to contain a relatively higher percentage of transcriptional regulators (e.g. Aravind *et al.*, 2005; Babu *et al.*, 2006a; Charoensawan *et al.*, 2010a; Pérez-Rueda *et al.*, 2004; Riechmann *et al.*, 2000; Shiu *et al.*, 2005; van Nimwegen, 2003; Vaquerizas *et al.*, 2009). TFs are usually classified by their DBD and frequently occur in large families. With the exception of plants, TF repertoires have been

found to be highly asymmetric in their composition with few, very large DBD families providing the majority of TFs (Charoensawan *et al.*, 2010b; Nowick & Stubbs, 2010; Pérez-Rueda *et al.*, 2004; Riechmann *et al.*, 2000; Shelest, 2008; Vaquerizas *et al.*, 2009). Plant TF repertoires are usually more balanced, containing more DBD families that contribute a significant portion of TFs (e.g. Mitsuda & Ohme-Takagi, 2009; Riechmann *et al.*, 2000; Shiu *et al.*, 2005).

Compared to the body of work already present and steadily growing on *cis*-regulatory evolution (see Chapter 1.4.1), relatively little is known about evolutionary dynamics in TF repertoires. Analysis has mostly focussed on the presence, absence, expansion or contraction of DBD families (e.g. Bussereau *et al.*, 2006; Charoensawan *et al.*, 2010b; Nowick & Stubbs, 2010; Pérez-Rueda & Janga, 2010; Pérez-Rueda *et al.*, 2004; Riechmann *et al.*, 2000; Shelest, 2008; Vaquerizas *et al.*, 2009) and has rarely included a more in-depth phylogenetic analysis (but see Haerty *et al.*, 2008, for example). From those studies it has become clear that lineage-specific amplification of different DBD families (at the broad taxonomic level) and different domain architectures (in the more shallow levels of classification) play an important role in TF repertoire evolution. Distantly related phyla are characterised by amplifications of different DBD families, e.g. the helix-turn-helix family in bacteria and archaea (Aravind & Koonin, 1999; Charoensawan *et al.*, 2010a; Pérez-Rueda & Janga, 2010; Pérez-Rueda *et al.*, 2004) or various types of zinc-coordinating DBDs in eukaryotic organisms (Charoensawan *et al.*, 2010b; Nowick & Stubbs, 2010; Riechmann *et al.*, 2000; Shelest, 2008; Vaquerizas *et al.*, 2009). Between more closely-related species, differences are often more fine-grained, e.g. C_2H_2 zinc fingers (ZFs) are amplified in many animal TF repertoires, but occur in different domain architectures that are specific to certain clades such as C_2H_2 - ZAD ZFs in insects, C_2H_2 - FAX ZFs in amphibians and C_2H_2 - KRAB ZFs in vertebrates (reviewed in Nowick & Stubbs, 2010). Such novel domain combinations can confer distinct roles to TFs containing the same DBD, e.g. the KRAB domain mediates interactions with the Kap1 cofactor which in turn recruits a histone deacetylase to modify local chromatin structure and act as a repressor (Vissing *et al.*, 1995). Besides different types of domains, the number of occurrences of the same domain in different TFs can span a large range. The C_2H_2 - KRAB ZFs in humans, for example, can encode over 30 repeats of

the C_2H_2 zinc finger motif (Tadepally *et al.*, 2008; Thomas & Emerson, 2009). Moreover, the absolute numbers of TFs with certain domain architectures can also vary greatly between closely related species, as is the case for the C_2H_2 - KRAB ZFs which show high copy number variation in different species of mammals and show signs of independent lineage-specific amplifications in different species (e.g. Tadepally *et al.*, 2008; Thomas & Emerson, 2009), indicating great evolutionary plasticity of TF repertoires at various taxonomic levels.

Recent work in yeasts, published after my study outlined below was begun, examined DBD family distribution across the fungi, using data available from the DBD database (hereafter referred to as DBD-DB to avoid confusion; Wilson *et al.*, 2008a). Shelest (2008) collected TF repertoires in 62 genomes spanning the entire kingdom and characterised those according to the types of DBD found. Overall, the study revealed matches for 37 different Pfam (Finn *et al.*, 2010) DBD families, five of which were found to be fungi-specific. Those included the $Zn(II)_2Cys_6$ DBD family, which also represents the largest TF family in the ascomycetous yeasts, the fungal-specific TF domain which was almost exclusively found in combination with a $Zn(II)_2Cys_6$ domain, the APSES DBD as well as the mating type factor MATα1 DBD (Shelest, 2008).

Here, I will present a comparative analysis of TF repertoires in 15 species of ascomycetous yeasts belonging to the *Saccharomycotina* (see Chapter 1, Fig. 1.9 for a full list of species). In this chapter I will mainly focus on the collection of my dataset, which also formed the basis for subsequent evolutionary analyses discussed in the following chapters, and the overall composition of TF repertoires within the species and clades studied with respect to distribution of different DBDs and domain architectures. I will discuss relative changes in the copy numbers of different DBDs and domain architectures and how those relate to genome evolution, e.g. the whole-genome duplication (WGD) event (see Chapter 1.5.1), and differences in life style in the three major clades studied. The pipeline used to assemble TF repertoires was distinct from the approach used by Shelest (2008) and together with the more detailed comparative analysis provided an extension to their work. A detailed comparison of the results obtained by Shelest (2008) and my study will be discussed below.

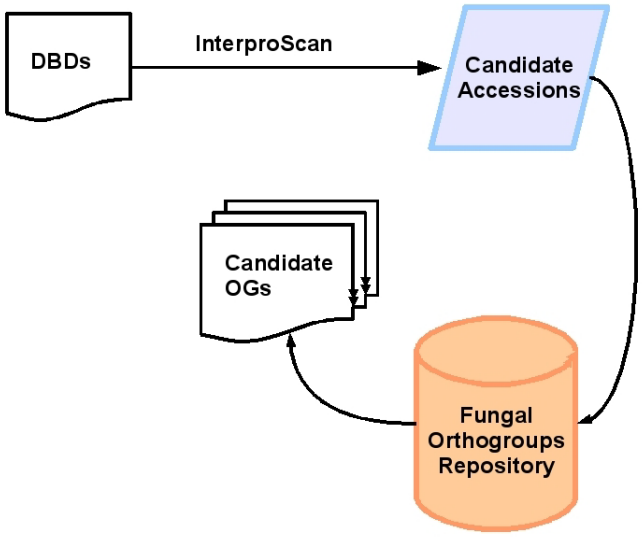


Figure 3.1: The DBD pipeline. Fungal proteomes were first screened for the presence of a DBD using InterProScan. Candidate accessions were then retrieved from the Fungal Orthogroups Repository and manually filtered for false positive matches to obtain the final dataset.

3.2 Genome-wide Screen for Transcriptional Regulators

3.2.1 Assembling Transcription Factor Repertoires: The DBD Pipeline

In order to obtain an unbiased view of the DBDs present in the *Saccharomycotina* I used a manually curated list of 168 InterPro signatures of DBDs from all domains of life to screen the full proteomes of all species studied. This list had originally been assembled by Juanma Vaquerizas for use in a screen for DBD-containing proteins in human (Vaquerizas *et al.*, 2009) and was subsequently augmented by me with missing fungal DBD signatures that were derived from domain annotations of known transcriptional regulators retrieved from the Saccharomyces Genome Database (SGD). An overview of the pipeline developed for the collection of fungal TF repertoires is shown in Figure 3.1.

InterProScan v.15 (Hunter *et al.*, 2009) was run on the full proteomes of all 15

species. InterPro is an integrated database of a number of different resources cataloguing known protein domains based on sequence similarity (ProDom), position-specific weight matrices (PRINTS) and hidden Markov models (HMMs; PANTHER, PIRSF, Pfam, SMART, TIGRFAMs, Gene3D and SUPERFAMILY) as well as short sequence motifs (PROSITE and HAMAP) that are consolidated under common InterPro domain and family identifiers (see Hunter *et al.*, 2009, and references therein). InterProScan uses the native search tools of each of those consolidated databases to identify domain or family signatures in query sequences and here scans were performed using all available search tools, keeping the intrinsic cutoff values of each of the member databases' screening methods fixed as those are optimised for each individual program. InterProScan results were then filtered for matches to the DBD signatures. Overall, I detected 4028 sequences that contained one or more matches to 77 of the 168 DBDs screened for.

The sequence identifiers of those were subsequently used to retrieve groups of orthologous proteins from the Fungal Orthogroups Repository (FOR; Wapinski *et al.*, 2007a) which is discussed in more detail in Chapter 4.2. Briefly, the FOR contains sequences and their inferred orthologs rooted at the most recent common ancestor of 23 ascomycetous fungi, the most divergent with respect to *Saccharomyces cerevisiae* being *Shizosaccharomyces pombe*. Such assemblages of orthologs are referred to as orthogroups and will be an important concept throughout this thesis. By definition, each orthogroup is derived from a single homolog in the common ancestor of the species included and members can either be in a one-to-one orthology relationship where no duplication events have occurred since the last common ancestor, or in a one-to-many relationship with respect to the root where duplications have happened since (see Figure 3.2 for examples). A special case are "orphan" orthogroups where only a subset of the species considered is present. These can have arisen either through duplications that could not be accurately mapped to their homologs, or through losses in other species and clades.

Lodderomyces elongisporus and *Saccharomyces kluyveri* were not included in the initial release of FOR (1.0) and homologs were mapped into existing orthogroups based on sequence similarity. Each orthogroup was aligned using Mafft 6.24 (Katoh & Toh, 2008) and alignment regions of potential low quality were

trimmed using Gblocks 0.91b (Castresana, 2000) with the same settings as described in Chapter 2.2. I used HMMER 2.3.2 (see Durbin *et al.*, 1998 and Eddy, 2009) to build global HMM profiles for each of the alignments using hmmbuild. Profiles were then searched against the respective proteomes and sequences were added to an orthogroup if a significant match to the orthogroup HMM was found. If the sequence resulted in significant hits to more than one orthogroup it was added to the one yielding the highest score. If no significant matches were found, the sequences were retained as a single-species orphan orthogroups. This approach had obvious shortcomings as despite using global HMM building and search settings, often alignable regions of orthogroups were limited to the DBD and regions of very strong local homology that can result in misassignment even when phylogenetic information is taken into account (see Chapter 4.4) and was later abandoned when an updated version of FOR became available that included *L. elongisporus* and *S. kluyveri* (see below).

By taking this approach I hoped to increase the coverage of my dataset in accounting for instances where the DBD motif, if present, did not produce a strong enough match to be included in InterProScan results as well as instances where the DBD has been lost completely. Such loss events are of considerable interest seeing that they provide examples of functional turnover, e.g. novel mechanisms of combinatorial regulation by tethering existing interaction partners without the ability to bind DNA thereby preventing them to perform other functions, or the loss of TF function all together. This increased the size of the dataset to 6327 sequences in total.

DNA-binding domains occur promiscuously in a large number of proteins, many of which are not directly involved in transcriptional regulation, e.g. proteins involved in splicing, DNA repair or telomere maintenance. Furthermore, known TFs such as Abf1 (see Chapter 3.6.1) are known to perform several functions in addition to TF activity, in this case including DNA replication and repair. This necessitates extensive filtering to remove false positives. All FOR orthogroups retrieved using the pipeline above were annotated in a two-step approach to ensure maximum sensitivity. An initial search against the Pfam database (Finn *et al.*, 2010) was used to determine significantly matching domains at the gathering threshold used by Pfam internally to build full alignments of a domain family.

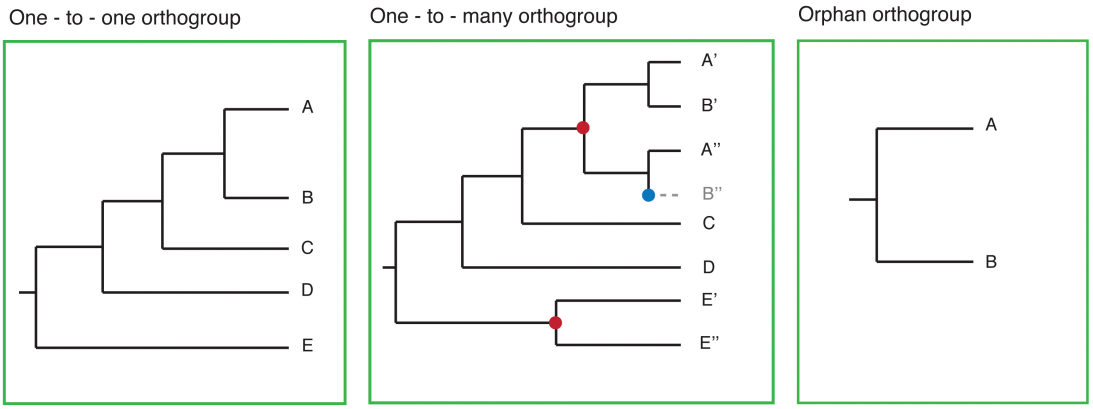


Figure 3.2: The orthogroup concept as defined by Wapinski *et al.* (2007a). One-to-one orthogroups contain single-copy orthologs of an inferred ancestral protein in the most recent common ancestor of A,B,C,D and E [A,B,C,D,E]. In this case the one-to-one orthogroup is said to be complete seeing that it contains descendants in every species. One-to-many orthogroups have experienced duplication events since the [A,B,C,D,E] ancestor. Here two duplication events (red) and one species-specific loss (blue) have occurred. Orphan orthogroups contain a subset of the species studied (one or several). The orphan orthogroup here is said to be rooted at [A,B] and can have arisen either through losses in C, D and E or been acquired on the branch leading to A and B.

This was followed by a second Pfam search with the E-value cutoff set to 10. If a domain was found only at the loosened threshold (E-value cutoff 10) in a protein but was detected at the more conservative gathering threshold in another member of the orthogroup it belonged to, the annotation was transferred.

Orthogroups were then manually filtered and examined for the presence of accessory domains that were indicative of roles other than transcriptional regulation. The *S. cerevisiae* ADP-ribosylation factor *AGE2*, for example, contains a GATA ZF-like domain but is not involved in transcriptional regulation as is underlined by the other domain it was annotated with, ArfGap, which mediates GTPase hydrolysis. If in doubt about a protein's functional role, I considered SGD annotation if available. Any orthogroups showing evidence against TF activity of at least a subset of members, either through their domain architectures or consultation of the literature, were removed from the dataset.

Whilst this strategy removed obvious false positives, it is likely that there remained sequences containing DBDs that are not involved in transcriptional regulation. Based on comparison with a collection of *Saccharomyces cerevisiae* TFs assembled from a range of different biochemical and ChIP-chip experiments the incidence of this seemed to be relatively low however (see below).

On release of an updated version of the FOR (FOR 1.1) in January 2009, the second part of the pipeline was rerun excluding sequences that were found to be false positives in the previous filtering step. The updated dataset was then filtered again resulting in a final dataset of 3222 putative transcription factors.

Table 3.1 shows the number of candidate TFs retrieved for each species in different steps of the pipeline. Inclusion of the FOR step did indeed improve coverage by over 35% compared to the raw InterProScan output. After all filtering steps were applied to the data, the percentage of sequences exclusively found through the FOR was still close to 20%, proving the utility of this approach. The discrepancy in percentages between the raw and filtered results however also indicated an increased inclusion of false positives. Inspection of multiple sequence alignments during later stages of analysis revealed that this was likely to be due to a bias in the SYNERGY algorithm underlying the FOR (Wapinski *et al.*, 2007a) that can lead to artificial grouping together of only marginally related sequences through shared local similarities. An in-depth discussion of this

Species	InterProScan	FOR 1.0	FOR 1.0 filtered	FOR 1.1	FOR 1.1 filtered	FOR only
Ashbya gossypii	238	399	170	212	179	18
Candida albicans	308	395	234	287	236	41
Candida tropicalis	175	496	237	290	235	117
Candida glabrata	276	441	194	243	208	22
Debaryomyces hansenii	314	541	239	280	239	24
Kluyveromyces lactis	250	424	188	231	193	18
Kluyveromyces waltii	251	436	175	240	196	32
Lodderomyces elongisporus	294	294∗	206∗	278	218	33
Pichia guilliermondii	329	532	249	307	254	24
Saccharomyces bayanus	289	459	184	239	205	36
Saccharomyces castellii	264	487	206	261	225	48
Saccharomyces cerevisiae	300	485	202	258	222	19
Saccharomyces kluyveri	124	124∗	182∗	213	187	121
Saccharomyces mikatae	309	472	192	254	216	28
Saccharomyces paradoxous	307	465	195	246	209	17
Total	**4028**	**6327**	**3102**	**3839**	**3222**	**598**

Table 3.1: Statistics for each step of the pipeline and the species considered. "InterProScan" refers to raw results from InterProScan. "FOR 1.x" results include InterProScan results as well as the additional homologs retrieved trough inclusion of different version of the FOR. "FOR 1.1 filtered" is the final dataset used. The species marked by an asterisk were not included in the FOR 1.0 release but were mapped into existing orthogroups based on sequence similarity (see above).

is provided in Chapter 4.2. Briefly, the reconstruction of orthogroups includes a BLAST-based clustering step which was implemented with a relatively liberal e-value cutoff that presumably was frequently surpassed by sequences sharing only short homologous regions but that are otherwise unrelated. This resulted in about a third of orthogroups containing such misclustered sequences (see Chapter 4.2). All affected orthogroups were split and manually reassessed for false positives. Despite these shortcomings of the algorithm however, the updated version of the FOR presented a considerable improvement and increased sensitivity with a further 120 putative transcription factors found compared to the filtered FOR 1.0 results. The species that benefitted most from the FOR update were *Saccharomyces kluyveri* and *Candida tropicalis* ("FOR only", Table 3.1). The improvement in coverage in those species was very clearly due to the release of a higher quality set of predicted proteins than the proteome used as initial input to the DBD pipeline.

Overall about 65% of matches found in the InterProScan step remained in the final dataset. This will have been partly be due to a large number of false positives

arising from the short and degenerate nature of some of those motifs, e.g. C_2H_2 ZFs, discussed further in consideration of results of a genome-wide resampling experiment of a subset of families (see below). However, this also underlined the promiscuity of DBDs in general. This promiscuity provides a huge evolutionary potential with the possibility for new regulatory proteins arising through *de novo* acquisition of DNA-binding domains or gain and/or loss of accessory domains after duplication of genes not previously involved in transcriptional regulation. An example of this is the Rap1 DNA-binding protein whose ancestral role is in telomere maintenance but which has been coopted as a transcription factor into the regulatory network controlling ribosomal regulation and glycolytic pathways in the *Saccharomycetaceae* (including post-WGD and pre-WGD clades but not the CTG clade; see Chapter 1, Fig. 1.9), facilitated by the gain of a transactivation domain (Tanay *et al.*, 2005).

3.2.2 Assessing Completeness of the Dataset

The fungal species considered in this study differ in their gene architectures, e.g. the number of intron-containing genes (see Chapter 1.5.1), and hence make it an elaborate and time-intensive task to perform *de novo* gene prediction. To bypass this problem I chose to search the predicted proteomes rather than the raw genome sequences for DBD-containing proteins based on the hope that individual genome sequencing projects will have established the best methods for predicting protein-coding genes in the respective genomes. While the *S. cerevisiae* proteome is likely to be of high quality, other species such as *Saccharomyces castellii* have been sequenced at low coverage and we would expect the proteomes of those to be of lesser quality. This has previously been shown to be problematic in a comparative study of C_2H_2 zinc fingers in human and chimpanzee: Tadepally *et al.* (2008) drew conclusions about certain TFs being human-specific based on their absence from the chimpanzee predicted proteome, but it was later shown by Thomas & Emerson (2009) that these absences were in fact often due to the low quality of the chimpanzee predicted proteome. In order to gauge the completeness of the dataset collected and to get a handle on how my dataset might be affected by annotation quality, I performed a genome resampling experiment of selected

DBDs in a subset of genomes as well as comparisons with previously published TF repertoires in *S. cerevisiae*, the results of which are outlined below.

3.2.2.1 Genome Resampling Experiment

I chose a subset of 15 DBD families to screen the full genomic sequence of four randomly sampled genomes per family. The genomes were assigned to one of the following categories: (1) trusted, (2) finished and (3) draft, based on the status of the genome sequencing. To ensure a representative sample from each category, one genome was chosen from categories 1 and 2 and two were sampled from category 3. Similarly, the 15 DBD families were chosen based on a broad coverage of different family sizes and phylogenetic distribution.

HMMs for each DBD selected were downloaded from Pfam (Finn *et al.*, 2010) and used as input for ESTwise (Birney *et al.*, 2004). ESTwise matches a protein HMM to a DNA sequence by scoring the HMM against protein models emitted from the DNA sequence, accounting for gaps and mismatches (but not introns). The algorithm used was "3:33L" which incorporates insertions and deletions throughout the the alignment and is in "looping mode", allowing for multiple local matches of the domain within an alignment.

ESTwise returns a list of genomic coordinates where the HMM motif has been found and their associated match scores. Pfam HMMs differ widely in their specificity for detecting a particular protein family and consequently results showed one of two patterns of score distributions (Figure 3.3). ESTwise results for different families either form a discrete (Figure 3.3A) or continuous distribution of scores (Figure 3.3B). In the latter case, definition of an appropriate detection threshold is a difficult problem and I decided to use the score of the lowest-scoring known true positive as a cutoff as this seemed to be the biologically most plausible solution even though a conservative one (see below for mapping of true positives).

ESTwise hits were matched with the protein sequences retrieved using the pipeline by their genomic location and were scored as either "matched", when the hit fell within 1000 bp of the mapped locus of one of the proteins found using the DBD pipeline; "non-matched", where a significant ESTwise match was found in the genome but I had not retrieved a protein mapping to the same genomic

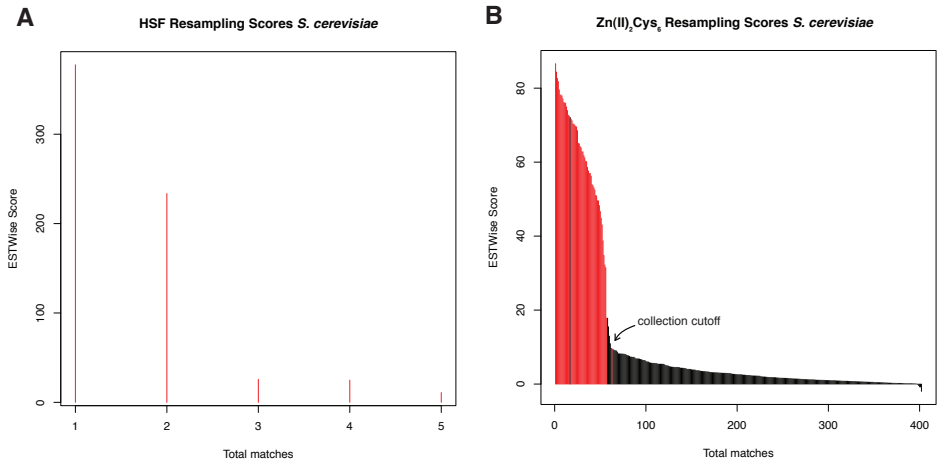


Figure 3.3: Illustration of the types of score distributions recovered for different HMMs in genome-wide screens using ESTWise. Known true positives are indicated in red, other scores in black. **A**: Score distribution of matches to the heat shock factor (HSF) DBD in *S. cerevisiae*. **B**: Score distribution of matches to the $Zn(II)_2Cys_6$ DBD in *S. cerevisiae*.

location; or "undetected", where there was a protein mapping to a particular location found by the pipeline but no ESTwise hit was found. If multiple matches mapped into the same open reading frame, they were considered to be repeats of the domain in a single protein. This should provide a general idea about (a) the sensitivity of the DBD pipeline in general, (b) the impact of annotation status on the completeness of results and (c) whether using ESTwise, which does not model introns, was appropriate for addressing (a) and (b). The results of the resampling experiment are shown in Tables 3.2 and 3.3.

Table 3.2 shows the consolidated coverage statistics from the genome resampling experiment. The overall agreement between the DBD pipeline and the ESTwise approach ("matched") was 86.11%. 2.92% of DBD proteins were exclusively found by the DBD pipeline, thus giving this approach an overall coverage of approximately 89%. Also, the small percentage of proteins that were missed by the resampling approach indicated that ESTwise's inability to model introns did not have a major influence on its performance in the species used in this study and therefore was suitable for the assessment of completeness. I examined

	Matched	Non-matched	Undetected	Total DBD	All found	Not annotated
All families	502	17	64	519	583	18
%	86.11	2.92	10.98	89.02		7.89

Table 3.2: Coverage statistics across all genomes and families analysed in the genome resampling experiment. The number of "matched", "non-matched" and "undetected" hits refer to hits detected by both ESTWise and the DBD pipeline, DBD pipeline only and ESTWise only respectively. "Total DBD" refers to the proportion of all hits found by either methods ("All found") that were detected by the DBD pipeline and "Not annotated" is counting the number of hits not found by the DBD pipeline that were due to missing protein predictions.

the remaining 10.98% of matches that were exclusively found by the resampling approach by searching the proteome for each species using BLASTx (Altschul *et al.*, 1990) and a 1kb region around the genomic location where the hit had been found to see whether these were due to missing annotations. Indeed, 18 out of the 64 DBD proteins that were found using the resampling approach but not in the DBD pipeline were not present in the annotated proteomes of the respective species.

A detailed breakdown of the different attributes of the genomes and TF families selected for resampling, such as genome annotation status, gene family size and whether or not ESTwise scores were distributed in a discrete or continuous way gave further insights into the performance of the pipeline on subsets of the data (Table 3.3). Overall sensitivity of the DBD pipeline ranged from 85% to 99% depending on which subset was considered. As expected, performance was worst in the low coverage genomes (Table 3.3; Annotation quality "draft") and reflected the fact that just under a third of hits detected by ESTWise only were due to missing predicted proteins in the proteome versions used (see Table 3.2). The best performance was seen for medium-sized DBD families (Table 3.3; Family size "2") where the DBD pipeline sensitivity was close to 99%. Presumably this was due to increased sensitivity of family HMMs for those seeing that they were likely to include a larger number of fungal representatives in Pfam (Finn *et al.*, 2010) due to the increased copy number. The two very large families (C_2H_2 ZFs and $Zn(II)_2Cys_6$ ZFs) did not benefit from this increased sensitivity however,

Factor	Category	Matched	Non-matched	Undetected	All found	DBD sensitivity
Score distribution	D	51	7	6	64	
		79.7%	10.9%	9.4%		90.6%
	C	451	10	58	519	
		86.9%	1.9%	11.2%		88.8%
Annotation quality	trusted	131	6	13	150	
		87.3%	4.0%	8.7%		91.3%
	finished	120	5	6	131	
		91.6%	3.8%	4.6%		95.4%
	draft	251	6	45	302	
		81.3%	2.0%	14.9%		85.1%
Family size	1	51	7	6	64	
		79.7%	10.9%	9.4%		90.6%
	2	83	5	6	89	
		93.3%	5.6%	6.7%		98.9%
	3	368	5	52	425	
		86.6%	1.2%	12.2%		87.8%

Table 3.3: Coverage statistics from the genome resampling experiment by sample category ("Factor"). Score distributions are either discrete ("D") or continuous ("C"); annotation quality is binned by 1: trusted, 2: finished and 3: draft (see text for details), gene family size distribution is binned according to average number of copies per genome, 1: 1-10, 2: 10-20, 3: >20. Percentages are calculated with respect to the to the total number of sequences found by both methods.

most likely due to their short and degenerate motifs and promiscuous occurrence in the genome (see Fig. 3.3B for example). Indeed, the C_2H_2 ZFs provided the largest proportion of hits missed by the DBD pipeline (∼50%). Nevertheless, whether or not scores were distributed discretely or continuously did not greatly alter the sensitivity of the DBD pipeline (Tab. 3.3, Score distribution) suggesting that missing protein annotation in the largest DBD families had the greatest impact on performance of the pipeline. It needs to be emphasised at this point, that some DBD families such as the C_2H_2 ZFs can encompass a number of subfamilies that are not functioning in transcriptional regulation (see Krishna *et al.*, 2003). As such the numbers presented here are likely a conservative estimate of missing regulators.

In summary, the DBD pipeline proved to be an adequate approach for assembling TF repertoires in yeasts without entirely reannotating parts of individual genomes which was beyond the scope of this study. Overall coverage was close to 90% which provided a reasonable overview of the transcriptional regulators encoded within the genomes of the species studied.

3.2.2.2 Comparison to Published Datasets

In addition to the resampling approach, I compared the *S. cerevisiae* DBD proteins collected with a database of transcriptional regulatory proteins (Wilson *et al.*, 2008a) as well as a previously published dataset of transcriptional regulators collected from a series of experimental studies (Jothi *et al.*, 2009).

The DBD Transcription Factor Prediction Database (DBD-DB; Wilson *et al.*, 2008a), is a database of automatically predicted transcription factors based on the presence of a sequence-specific DBD motif. DBD-DB uses Pfam (Finn *et al.*, 2010) and SUPERFAMILY (Gough *et al.*, 2001) for detection of DBD motifs, a subset of the methods implemented in InterProScan (Hunter *et al.*, 2009). Despite the availability of other species studied here in DBD-DB, I decided to focus on *S. cerevisiae* for an in-depth comparison (Figure 3.4). This was due to the extensive availability of functional annotation and consequent relatively easy assessment of false positives. Statistics for other species available in DBD-DB closely mirrored those obtained for *S. cerevisiae* and can be found in the Appendix, Table A.1.

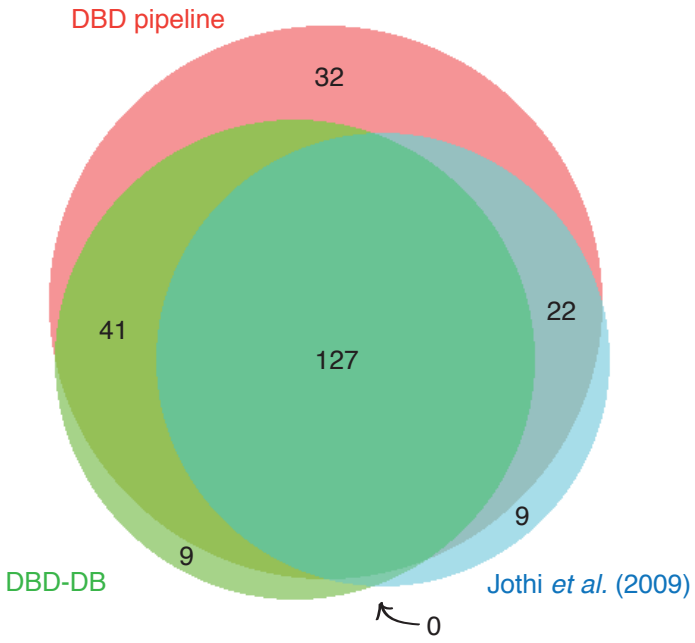


Figure 3.4: Comparison of the collected TF repertoire for *S. cerevisiae* "DBD pipeline" with the datasets retrieved from DBD-DB (Wilson *et al.*, 2008a) and Jothi *et al.* (2009).

DBD-DB contains 177 DBD-containing proteins for *S. cerevisiae* (as of June 2010), 168 of which were shared with the dataset I collected using the DBD pipeline. Inspection of the nine missing proteins revealed that seven of those are not involved in transcriptional regulation and/or do not bind DNA in a sequence-specific manner and are thus false positives. In contrast, 54 DBD proteins were uniquely found by my DBD pipeline and included well-characterised TFs such as MSS11, MET4 or OPI1 as well as uncharacterised open reading frames, underlining the value of using the combined InterProScan and FOR approach instead of relying on previously published datasets for increased sensitivity.

Jothi *et al.* (2009) assembled the *S. cerevisiae* transcriptional regulatory network from an array of biochemical and ChIP-chip experiments focussing on interactions involving sequence-specific DBD proteins to the exclusion of chromatin remodelling factors. Overall, they retrieved 158 DBD proteins, 149 of which were shared with my dataset. Out of the nine missing proteins, four were false positive hits. Conversely, 73 DBD proteins were exclusively found by the DBD pipeline. The remaining five proteins not found by the DBD pipeline were subsequently retrieved from the FOR along with their homologs in other species and added to the final dataset. Inspection of the domain architecture of those revealed that none of the sequences had significant matches to any of the InterPro signatures screened for and had thus not been detected in the DBD pipeline. Finally, I found no TFs that were represented in both DBD-DBD and the Jothi *et al.* (2009) dataset but not detected by the DBD pipeline further underlining the good coverage of my dataset.

Overall, I am thus confident that the final version of the data (as of June 2010) is a comprehensive collection of TF repertoires in the species studied. The DBD pipeline approach proved to be of good coverage, reaching approximately 90% sensitivity when averaged across different families and genomes. Specificity however was relatively low, seeing that as little as 65% of original InterPro hits were retained in the final dataset. This was due to both the promiscuity of DBDs as well as the sometimes short and degenerate motifs that are likely to result in false positives and false negatives alike. The C_2H_2 ZF motif is 23 amino acids long and only seven of those are strongly constrained (Finn *et al.*, 2010). In line with this, ESTWise reported over 33,000 hits of this motif in the *C. albicans*

genome which, although generally low-scoring, is in stark contrast with the small number of hits retrieved for e.g. the heat shock factor DBD (Fig. 3.3A). Furthermore, C_2H_2 ZFs can also bind to RNA and other proteins (e.g. Brayer & Segal, 2008) thus highlighting not only the difficulty in initial detection of C_2H_2 ZF proteins but also the ambiguities about their functional role. Here, I have overcome these issues through extensive manual filtering and consultation of the literature. While this approach was feasible for organisms with small genomes and sufficient experimental data, higher eukaryotes can encode thousands of TFs, very few of which have been experimentally confirmed (Vaquerizas *et al.*, 2009), and so the approach is not likely to scale well. Studies such as the one I presented here, however could be used to derive rules for more automated annotation of TFs in other organisms, although this is beyond the scope of this thesis.

3.3 Transcription Factor Repertoires in the *Saccharomycotina*: A Parts List

Table 3.4 shows a summary of the numbers of transcription factors retrieved in relation to the total number of protein-coding genes in the species studied. Overall, transcription factor repertoires were relatively stable in size with respect to the numbers of annotated genes in each genome, making up between 3.5 to 4.5% of protein-coding content which is in line with previous estimates (e.g. Riechmann *et al.*, 2000). Noticeably, the percentages in *Saccharomyces paradoxous*, *Saccharomyces mikatae*, *Saccharomyces bayanus* and *Saccharomyces castellii* are slightly higher compared to the closely related *S. cerevisiae* and *Candida glabrata*. This excess in relative contribution of TFs might be due to the comparatively low quality of annotation in those species: The number of protein-coding genes predicted in *S. mikatae*, *S. bayanus* and *S. castellii* is lower overall than in *S. cerevisiae* and *C. glabrata* and it is conceivable that DBD-containing proteins would be easier to detect, given their occurence in big and often well-defined families as well as their inclusion in the Yeast Gene Order Browser (YGOB; Byrne & Wolfe, 2005) incorporated into the Synergy algorithm allowing for easier detection of homologs in those species. *S. kluyveri*, conversely, has the lowest percentage of

Species	Total Numbers of TFs	DBD Families	Annotated Genes	%TFs
S. cerevisiae	226	45	5807	3.89
S. paradoxus	215	45	4788	4.49
S. mikatae	221	44	4525	4.88
S. bayanus	211	45	4492	4.70
S. castellii	231	45	4677	4.94
C. glabrata	214	45	5283	4.05
K. lactis	201	46	5329	3.77
A. gossypii	186	44	4718	3.94
K. waltii	186	44	5230	3.56
S. kluyveri	187	45	5321	3.51
C. albicans	241	46	6354	3.79
C. tropicalis	234	45	6258	3.74
L. elongisporus	217	45	5802	3.74
P. guilliermondii	254	44	5920	4.29
D. hansenii	240	46	6564	3.66

Table 3.4: Numbers of TFs collected in each species and relationship to the total number of protein-coding genes. Rows are coloured by the three major clades: (yellow) post-WGD; (green) pre-WGD; and (blue) CTG.

transcription factors relative to the size of its predicted proteome. This probably reflects that the initial InterProScan search has been performed on an old, lower quality set of annotations than the one included in FOR 1.1 and consequently might be missing a number of singleton TFs that had not been included in the original annotation or remained undetected by InterProScan.

Similarly, the number of families found in each genome is fairly stable, ranging from 44 to 46 families each. Interestingly, none of the genomes encodes TFs in all 48 families detected. A detailed breakup into contribution of DBD families is shown in Figure 3.5. Here every TF is categorised by the presence of one

type of DBD. Overall, the composition of TF repertoires is very similar across all 15 species. The majority of DBD families are present in low copy number, with ∼68% of families accounting for ∼26% of DBDs encoded in the collected TF repertoires. A further 18% of families are predominantly present in intermediate numbers (4 to 12 copies each) contributing another 22%, whilst the remaining four families have been heavily amplified, making up over 50% of the TF repertoires. The main contributors to TF repertoires in the *Saccharomycotina* include two families of zinc-coordinating DBDs, the fungal-specific binuclear cluster ($Zn(II)_2Cys_6$) and C_2H_2 zinc fingers as well as the less well-characterised, also fungal-specific Fungal_trans domains. These findings mirror previous reports of transcription factor repertoires in ascomycetous fungi (Babu *et al.*, 2004; Riechmann *et al.*, 2000; Shelest, 2008). In the following I will discuss some of the most significant trends based on the distributions of DBDs as well as the different domain architectures they occur in, and comment on the biological implications of these findings.

3.4 DNA-binding Domain Distribution

As mentioned above, I found the contribution of different DBDs to the TF repertoires to be highly asymmetric. The $Zn(II)_2Cys_6$ zinc fingers ("Zn_clus" in Fig. 3.5) form the largest family across all species, ranging between 39 and 91 members in *Ashbya gossypii* and *Pichia guilliermondii* respectively. $Zn(II)_2Cys_6$ zinc fingers predominantly belong to the ascomycete family although one member has been found in a basidiomycete species (see Chapter 1, Fig. 1.9 for phylogenetic relationships), suggesting either an older origin and subsequent large-scale loss or horizontal gene transfer (MacPherson *et al.*, 2006). $Zn(II)_2Cys_6$ zinc fingers frequently occur in combination with a Fungal_trans domain, the third biggest contributor to TF repertoires, which is almost exclusively found in this arrangement, where the $Zn(II)_2Cys_6$ domain is located at the N-terminal and the Fungal_trans domain at the C-terminal end of the protein. Although I found several occurrences of the Fungal_trans domain in isolation, these were small in number and in orthogroups containing few homologous sequences suggesting its dependency on the $Zn(II)_2Cys_6$ domain for canonical functionality. The $Zn(II)_2Cys_6$

domain however was frequently found to occur in isolation as is evident from the difference in abundance between those domains (Figure 3.5).

The second largest contributors to the TF repertoires were the C_2H_2 zinc fingers. This DBD is conserved across all of eukaryotic life and is amplified heavily in several taxonomic groups including human and *Drosophila melanogaster* (Emerson & Thomas, 2009). Less heavily amplified, but still in substantial copy number were proteins containing one or more Myb-like DBDs. These are also conserved across all of the eukaryotes but show most diversity in plants (Jiang *et al.*, 2004; Riechmann *et al.*, 2000) and as a family remain relatively understudied, especially in fungi.

None of the families showed a substantial increase in copy number following the WGD (marked by the vertical dashed line in Figure 3.5) compared to their immediate relatives. This is in line with the findings that most duplicate genes were lost quickly after the WGD event (Scannell *et al.*, 2006) and argues for the lack of a large-scale effect of the WGD on the overall composition of TF repertoires. This was despite an overall increase in numbers of regulators and a larger proportion of TFs with respect to the total number of protein-coding genes (Table 3.4), suggesting that gains through the WGD were spread among several families.

An example of lineage-specific gain is the BAF1_ABF1 DBD, found in the multifunctional global regulator *ABF1* that is thought to be involved in a variety of functions including gene activation and repression, silencing, DNA replication and chromatin remodelling (Miyake *et al.*, 2004). This domain seemed to have appeared on the lineage leading to the clade spanning *S. cerevisiae* and *A. gossypii* and is generally not found outside this group of species, although the InterPro database (Hunter *et al.*, 2009) suggested the presence of a single copy in one of the *Drosophila* species as well as barley which, if correct, would be most parsimoniously explained by horizontal gene transfer, although the origins of the fungal domain itself remain unclear. Horizontal gene transfer, the transmission of genetic material other than through the germline, is a common mechanism for the gain of new TFs in bacteria (e.g. Price *et al.*, 2008). Although horizontal gene transfer between eukaryotes has classically been thought of as rare (see Keeling & Palmer, 2008 for review), it has recently been shown that transposable

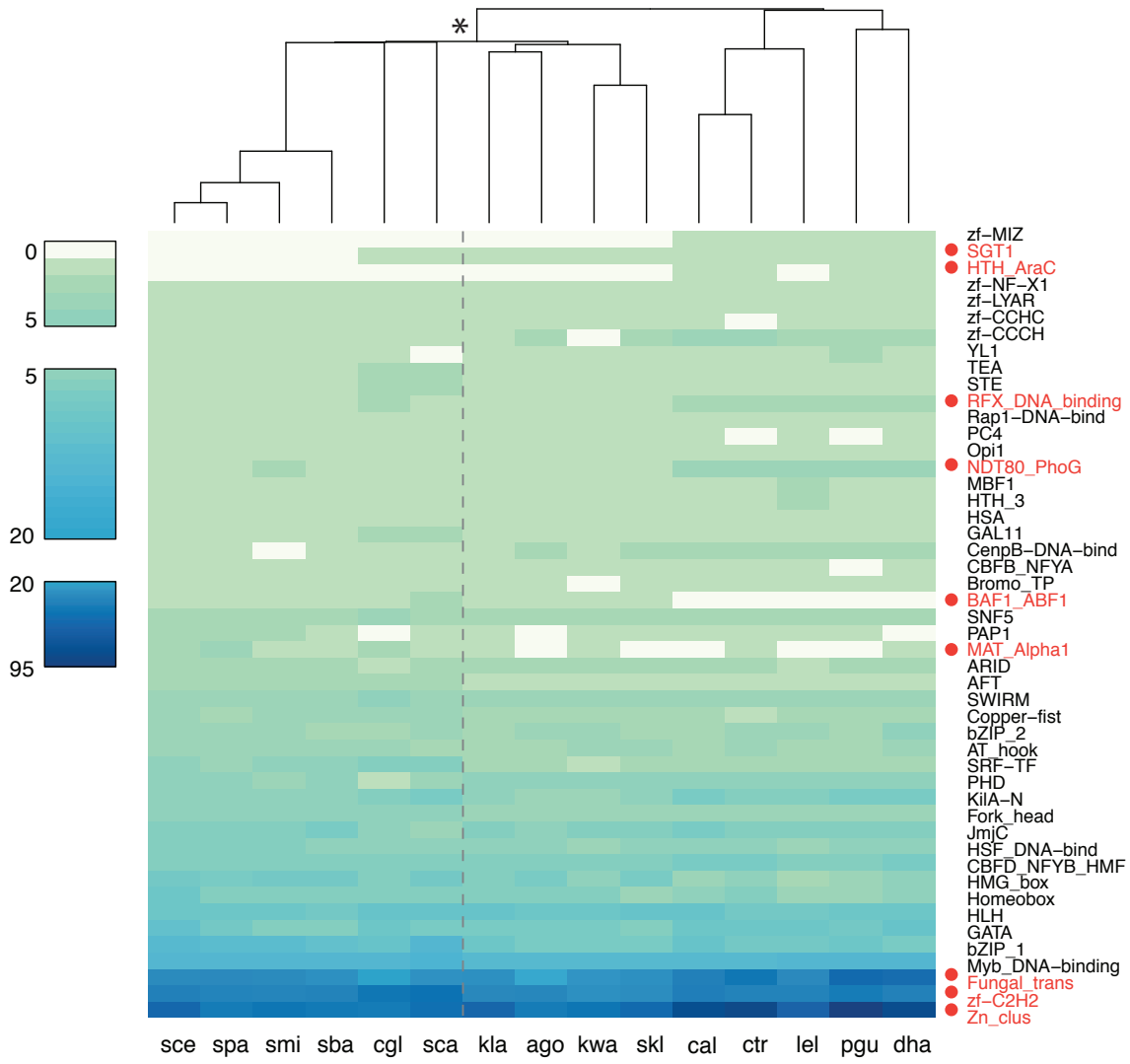


Figure 3.5: Distribution of different DBDs in TF repertoires across the *Saccharomycotina*. Fields are coloured depending on the number of family members found in each genome and sorted by the number of family members in *S. cerevisiae*. The WGD event is indicated by a dashed line and a star. Families marked in red are discussed in more detail in the main text.

elements (TEs) have the potential to be transferred between higher eukaryotes, e.g. (Pace *et al.*, 2008). Conceivably they could carry other genetic material, including parts of or full DBDs, during transmission. The recent TE colonialisations of higher eukaryote genomes also have implications for the establishment of new DBD families such as the WRKY-GCM1 zinc fingers which derive from the DBD of a transposase (Babu *et al.*, 2006c). By this mechanisms related DBD families could have been established independently in different clades, although currently there is lack of evidence supporting any of the horizontal transmission scenarios in this case.

Apart from the BAF1_ABF1 domain, all remaining DBDs found were ancestral to the group of species studied here based on information obtained from InterPro (Hunter *et al.*, 2009). In consideration of this, I found lineage-specific losses in several of the 48 DBD families. The larger patterns of absence across entire clades included the zf-MIZ, SGT1 and HTH_AraC domain families. In the case of the zf-MIZ DBD, which has been lost in the clade spanning *S. cerevisiae* and *A. gossypii*, there were still instances of this domain in the proteomes of those species, but these were not found in proteins directly involved in transcriptional regulation. There was no direct evidence for involvement of the remaining homologs in transcriptional regulation either but they lack the accessory domains associated with the SUMO ligase functionality found in the proteins removed as false positives. The SGT1 family appeared to have been lost from the *Saccharomyces sensu stricto* species. These domains occur across the eukaryotes, albeit in low copy number, and are well-conserved (Kainou *et al.*, 2006). SGT1 homologs have been shown to function, possibly as coactivators, in carbohydrate metabolism and this finding was thus of considerable interest and will be discussed in detail below. HTH_AraC is a predominantly bacterial DBD and is found across the *Ascomycota* in combination with an Ada domain which is also of bacterial origin, suggesting an ancient horizontal gene transfer into this clade. This DBD has been lost from most species, being retained only in *C. albicans*, *Debaryomyces hansenii*, *Candida tropicalis* and *Pichia guilliermondii* (Fig. 3.5).

Other examples of losses included the MATα1 (MAT_Alpha1 in Fig. 3.5), mating regulators which were also of interest as the yeasts studied here have evolved very different strategies for the control of mating loci (e.g. Butler *et al.*, 2004; Lee

et al., 2010b; Tsong *et al.*, 2006; see Chapter 1.4.2.2 for discussion). The data collected suggested the absence of MATα1 factors from *A. gossypii*, *S. kluyveri*, *C. albicans*, *Lodderomyces elongisporous* and *P. gulliermondii*. A literature search however revealed that *C. albicans*, for example, does indeed encode an ortholog of *S. cerevisiae* MATα1 that was previously undetected (Soll *et al.*, 2003). This was due to not being in the genome of the strain sequenced and had thus also not been included in the proteome. Unlike most of the *Saccharomyceraceae*, species in the CTG clade do not encode silent mating type loci, coding for silenced copies of both mating types (Butler *et al.*, 2004; Génolevures Consortium *et al.*, 2009; Lee *et al.*, 2010b; Soll *et al.*, 2009). It is possible that the sequenced genomes thus only contain orthologs for one of the two mating types commonly found, depending on which strain has been sequenced, thus necessitating more thorough investigation. Similarly, there appeared to be several species-specific loss events (such as for the PC4 and YL1 domains). These also need additional investigation to exclude the possibility that they were merely missed during data collection and will not be discussed here in further detail.

The results I obtained here were qualitatively similar to the study by Shelest (2008) reviewed in the introduction to this chapter, who also found the predominance of $Zn(II)_2Cys_6$ zinc fingers. Overall they had found 37 families of Pfam DBDs, 30 of which were also included in my dataset. The remaining six were either not present in the *Saccharomycotina* (CP2, HTH_psq) or found in proteins involved in other functions in this clade (TFIIH C1, DDT, Not1, SART-1,zf-C5HC2). Conversely, there were 17 families in my dataset not found by Shelest (2008), including BAF1_ABF1 mentioned above and Rap1 which is also referred to throughout this thesis due to its central role in the rewiring of ribosomal proteins (e.g. Chapters 1.4.2.1 and 6.5.1). The gain and divergence of both those TFs proved to be of considerable biological importance (see below). Closer inspection of four of the additional DBD families in my dataset, despite a role in transcriptional regulation, however also indicated lack of evidence for TF roles in the *Saccharomycotina* and were excluded from downstream analyses in the following chapters (GAL11, zf-CCCH, HSA and SNF5).

3.5 Domain Architectures

Next, I focussed on the different domain arrangements that these DBDs occurred in. An overview of the domain architectures recovered and their relative abundance is shown in Figure 3.6. Compared to considering DBDs in isolation, we begin to see more detailed dynamics of family evolution that were previously concealed. The C_2H_2 zinc fingers for example occur in several domain architectures that are amplified to different extents in different clades (e.g. TFs containing either two or three C_2H_2 fingers, highlighted in blue in Fig. 3.6). One of the most interesting global patterns emerges in *C. glabrata* and *S. castellii*, the species that diverged in quick succession just after the WGD. In contrast to the loss of duplicate copies experienced by the *Saccharomyces sensu stricto* species, those differ in their retention and loss patterns and *S. castellii* has retained a larger proportion of the duplicated TFs (cf. Table 3.4). This has been previously described by Scannell *et al.* (2006) in an analysis of patterns of reciprocal gene-loss after the WGD and includes major transcriptional regulators such as *STE12*, *TEC1*, *TUP1* and *MCM1* suggesting greater regulatory complexity in those species.

The increase in the number of transcriptional regulators in the CTG clade (see Table 3.4) is to the greatest part accommodated by large lineage-specific expansions of TFs with $Zn(II)_2Cys_6$ domains, both in isolation and in combination with an N-terminal Fungal_trans domain (Fig. 3.6, purple). The abundance patterns of TFs containing an isolated Fungal_trans DBD in contrast are somewhat erratic and in low copy number, further arguing for its dependence on the $Zn(II)_2Cys_6$ domain to carry out its role as a transcriptional regulator. The exact function of this domain is not well-understood and although annotated as DNA-binding in Pfam, it has been implicated in regulating transcriptional activity and ligand-binding (MacPherson *et al.*, 2006). Its functionality as "stand-alone" TF is thus unclear, although even without a $Zn(II)_2Cys_6$ zinc finger it might function as a regulatory cofactor, given that TFs containing both domains often act as homo- or heterodimers (MacPherson *et al.*, 2006). Interestingly, the third most abundant class of TFs, C_2H_2 zinc fingers containing a tandem repeat of the zinc finger domain, have remained relatively stable in their abundance. The main increase in C_2H_2 zinc finger-containing TFs in *Candida* species and their close relatives is

seen in proteins containing three C_2H_2 repeats (Fig. 3.6; zf-C2H2x3). One might speculate that this is a feature of greater regulatory complexity seeing that the binding motifs of tandem C_2H_2 (Fig. 3.6; zf-C2H2:zf-C2H2) are short with each zinc finger recognising a three nucletotide stretch of DNA and that in order to sustain sufficient regulatory specificity, amplification beyond a certain size is limited to TFs with higher DNA-binding recognition specificities. Indeed genomes encoding large numbers of C_2H_2-containing TFs tend to have larger number of zinc finger repeats (Emerson & Thomas, 2009). The TF repertoires collected also include C_2H_2 zinc finger proteins with larger numbers of zinc finger repeats (four, six and seven) but those have remained comparatively stable in number, raising the possibility of an "upper bound" to the optimal length of DNA-binding motif for families displaying large evolutionary plasticity that is in some way related to the rate of *cis*-regulatory DNA evolution.

Beside these larger amplifications, there are a number of smaller amplifications specific to the *Candida* species cluster (Fig. 3.6, purple), seeing copy number increases in several of the smaller TF repertoire contributors. Examples of those are NDT80 / PhoG like DNA-binding proteins, the duplicate of which has been found to be involved in drug resistance in *C. albicans* (Chen *et al.*, 2009) and *RFX2*, a *C. albicans* homolog of *S. cerevisiae RFX1*, encoding a RFX DNA-binding domain that is also thought to play a role in virulence (Hao *et al.*, 2009).

With respect to the domain architectures found there is some evidence of lineage-specificity, for example the zf-C3HC4:Zn_clus:Fungal_trans domain arrangement which includes a zf-C3HC4 domain to the N-terminal of the binuclear cluster zinc finger appears to be specific to the *Candida* lineage where it has subsequently been amplified. Another example are the homologs of *S. cerevisiae IXR1* that contains two high mobility group box DBDs (HMG_box:HMG_box; Fig. 3.6, orange) ,and can bind DNA containing intrastrand cross-links (Lambert *et al.*, 1994). This architecture appears to have been lost from *Candida* and related species. Furthermore I find incidences of potential domain shuffling, that must have occurred before the divergence of the species considered here based on their phylogenetic distribution, e.g. the ZZ:Myb_DNA-binding:SWIRM and SWIRM:ZZ:Myb_DNA-binding TFs where the SWIRM domain has moved from one end of the protein to the other.

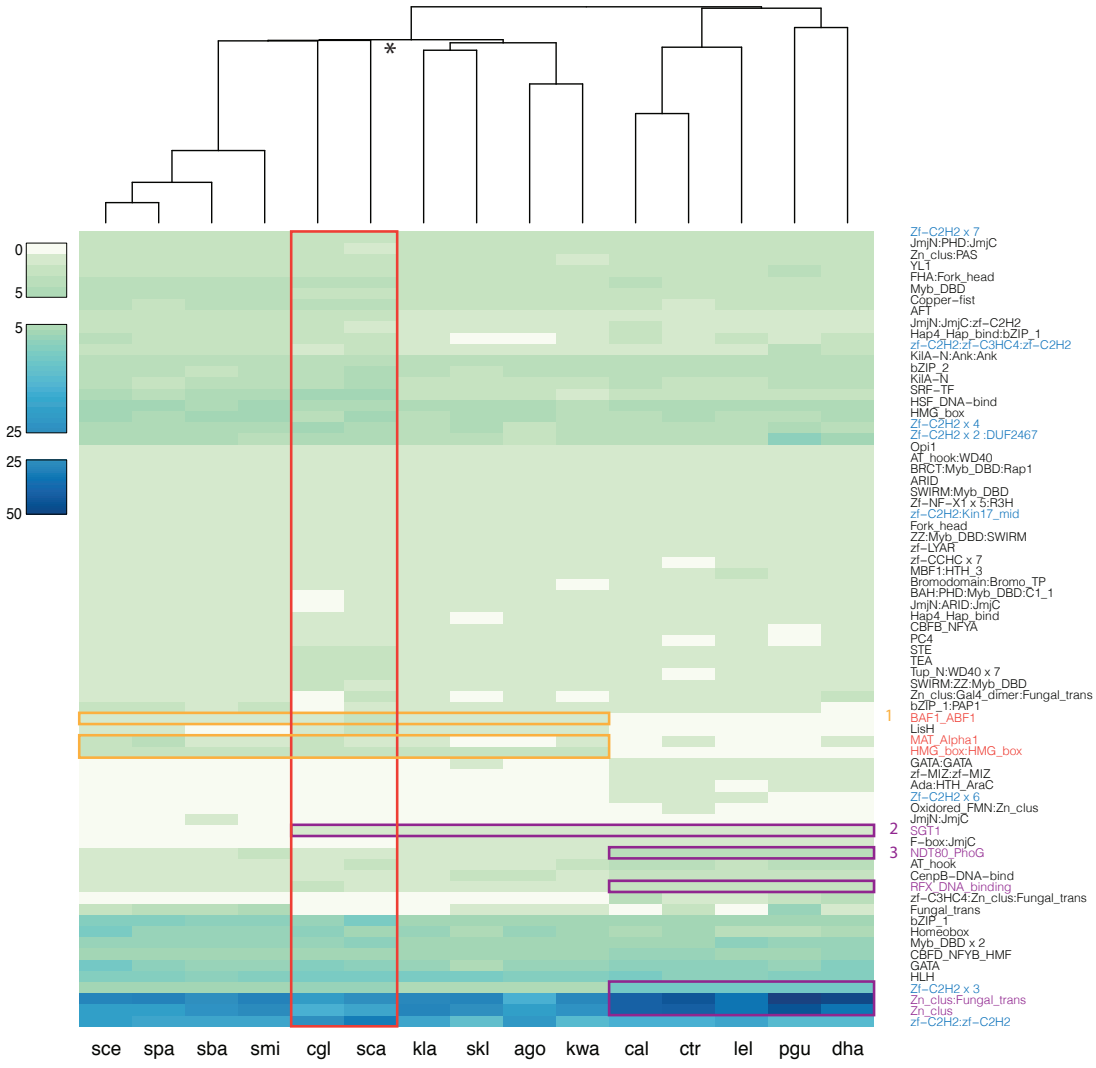


Figure 3.6: Domain architectures recovered in *Saccharomycotina* TF repertoires. Cases of lineage-specific expansion or loss are highlighted in orange (*Saccharomycetaceae*) or purple (CTG clade). The whole-genome duplication is marked by an asterisk. Both *C.glabrata* and *S. castellii*, the two species that diverged just after the WGD show distinct patterns of retention of WGD duplicates (red box). Numbered rows are referred to in the main text.

3.6 Functional implications

The genome-wide analysis of transcriptional regulators has not only produced an overview of the global patterns describing TF repertoire evolution, but is also detailed enough to provide information about individual evolutionary histories of DBD families that are relevant to the biology of the species studied. In the following section I will discuss selected examples of TFs showing lineage-specific patterns of copy number abundance that are of functional importance considering the biology of the yeasts studied here. These will include the multifunctional transcriptional regulator *ABF1*, the *SGT1* regulatory protein that has been lost in the *Saccharomyces sensu stricto* species, some of the regulators that were retained in duplicate after WGD in *S. castellii* or *C. glabrata* but subsequently lost in other species and an overview of TFs showing lineage-specific amplification in the CTG clade.

3.6.1 Evidence for gain of function in *ABF1* after the WGD and a possible role in efficient establishment of petite morphs

ABF1 encodes a multifunctional regulatory protein that is essential in *S. cerevisiae*. It performs a wide range of functions, acting as a transcriptional activator as well as repressor (Miyake *et al.*, 2004) in a variety of cellular processes such as carbon source regulation, meiosis, sporulation, mitochondrial and ribosomal regulation (see Miyake *et al.*, 2004, and references therein). Furthermore, *ABF1* is involved in DNA replication, the role in which it had been originally identified (Rhode *et al.*, 1989), and nucleotide excision repair (Reed *et al.*, 1999), chromatin remodelling (Venditti *et al.*, 1994) and gene silencing (Rusche *et al.*, 2003). Its high abundance and ubiquity in the regulatory network (Lee *et al.*, 2002; Luscombe *et al.*, 2004) underline the essential role of this protein in *S. cerevisae*.

Figure 3.7 shows an overview of the domain architecture of *ABF1* in *S. cerevisiae* and the corresponding sections of the multiple sequence alignments in the species encoding *ABF1* homologs found in this study. The protein contains a bipartite DBD with an abnormal zinc finger domain at the N-terminal end (DBD1)

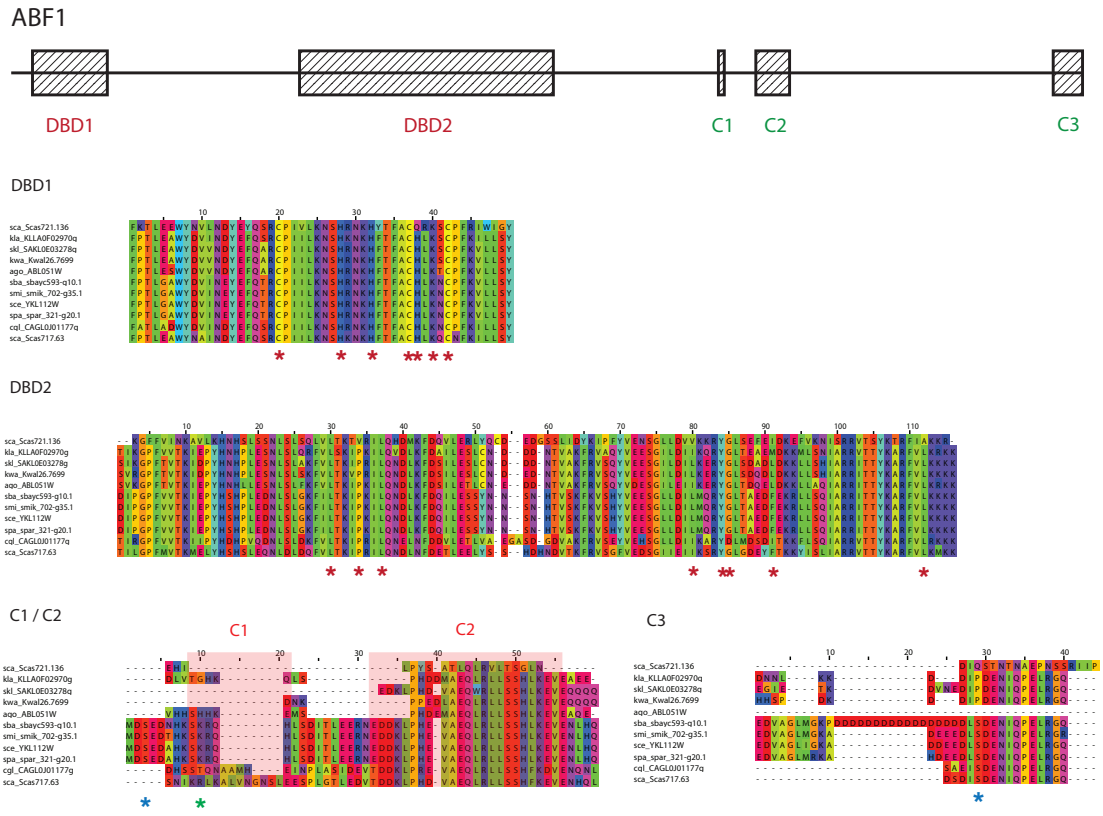


Figure 3.7: Mutliple sequence alignment of the conserved regions of *ABF1*. Stars indicate experimentally characterised phosphorylation sites and are discussed in more detail in the main text.

and a second part, resembling a helix-loop-helix domain (DBD2) further downstream seperated by a conserved linker region. Whilst both parts are important for DNA-binding activity, DBD2 has been implicated for determining binding specificity (Cho *et al.*, 1995). The linker region is absent in species that diverged before the WGD and has been shown to be dispensable for DNA-binding activity in *S. cerevisiae* through deletion studies (Cho *et al.*, 1995) as well as the rescue of an inviable *ScABF1* knockout mutant by replacement with the homologous protein from *K. lactis* (Gonçalves *et al.*, 1992).

ABF1 also contains a C-terminal activation domain that can be separated into several individual components that perform distinct roles, contributing to

the great functional modularity of this protein (Loch *et al.*, 2004; Miyake *et al.*, 2002). The C1 region (see Figure 3.7) spans a stretch of four amino acids in *S. cerevisiae* and has been determined to be important for nuclear localisation of Abf1p (Loch *et al.*, 2004). A temperature sensitive mutation in *ScABF1*, K625I, marked by the green asterisk in Figure 3.7, disrupts nuclear localisation of Abf1p at high temperatures that was found to be rescued by the addition of a heterologous nuclear localisation sequence (NLS) (Loch *et al.*, 2004). Interestingly, these authors identified a second region capable of mediating moderate levels of nuclear import upstream of C1 indicating the presence of several different pathways for nuclear import. The C2 region just downstream of the C1 region has been implicated to have a role in chromatin remodelling and consequently DNA replication as well as transcriptional activation (Miyake *et al.*, 2002). Indeed, the *ScABF1* C2 region was found to be sufficient for chromosomal activation and stimulation of transcriptional regulation when fused to a heterologous DBD. The function of the region just at the C-terminal of the protein, referred to as "C3" from this point, is less well-understood. Despite its considerable conservation, this region has usually been found to be "dispensable" in functional studies of the C-terminal end of *S. cerevisiae ABF1* (Loch *et al.*, 2004; Miyake *et al.*, 2002). Given the strength of the phenotypes resulting from mutations in C1 and C2, however and the functional modularity of the protein, it is likely that any potential effects of this region have thus far been masked by the penetrance of other mutations in the conditions tested. Indeed, C3 has been found to contain a casein kinase II phosphorylation site (blue asterisk, Figure 3.7) arguing for a functional role of the conserved C-terminal end of *ABF1* (Upton *et al.*, 1995).

Analysis of the abundance pattern and multiple sequence alignments of *ABF1* in the TF repertoires I collected (see Figure 3.6, note 1; Figure 3.7) revealed several interesting points with respect to evolutionary mechanisms generating such multifunctional proteins as well as the biology of the species studied. Firstly, *ABF1* has been retained in duplicate after the WGD in *S. castellii* only, whilst WGD duplicates were lost in all other post-WGD species studied here. Examination of the alignments of all collected homologs gave evidence for sub-or neofunctionalisation of one of the two copies. Scas721.136 (top line Fig. 3.7) is diverging faster than its WGD paralog Scas717.63 (bottom line, Fig. 3.7), having

accumulated mutations in both parts of the DBD that had previously been determined to be important for DNA-binding activity in *S. cerevisiae* (red asterisks, Figure 3.7). The most striking of these are possibly the substitution of an otherwise highly conserved histidine to a glutamine in DBD1 (pos. 38; Fig. 3.7), which forms an important part of the zinc finger motif (Cho *et al.*, 1995), and substitution of an equally well-conserved proline residue to a valine in DBD2 (pos. 34; Fig. 3.7). Indeed, both parts of the Scas721.136 DBD harbour numerous other substitutions, suggesting changes in DNA-binding ability and/or specificity. The C-terminal conserved regions of Scas721.136 also show greater divergence than its paralog, with C1 essentially missing. It needs to be noted that alignment around the C1 region is difficult however and it is not clear if either the second *S. castellii* or the *C. glabrata* homologs contain a functional NLS or a phosporylation site, both of which are clearly conserved in the *Saccharomyces sensu stricto* species.

Similarly, the C3 region shows less conservation in Scas721.136 compared to the other homologs. Notably, the serine which is phosphorylated by casein kinase II (Upton *et al.*, 1995) is only conserved in the post-WGD species (blue asterisk, Figure 3.7). The C2 region in turn is conserved across all homologs found, reflecting the importance of chromatin remodelling to most if not all of the functions known to be carried out by Abf1p. Given these results it thus appears that whilst the Scas717.63 homolog of *ABF1* is under selective pressure to retain most of the functionality of its counterpart in other species, its WGD paralog Scas721.136 is diverging in both its DBD and regulatory regions whilst retaining at least part of the functionality as evidenced by the retention of the C2 region.

Secondly, the phylogenetic distribution of *ABF1*, which seems to have appeared in a common ancestor to the *Saccharomycetacae* according to this study and supported by the absence of this domain outside this clade in the InterPro database (Hunter *et al.*, 2009) is striking, given the extent of integration of this protein into the regulatory network as well as numerous other cellular processes in *S. cerevisiae*, providing a great example of the evolutionary plasticity in gene regulatory networks. Furthermore, the appearance of *ABF1* coincides with several interesting biological processes, exclusive to that clade, in which Abf1 has been found to play a role in *S. cerevisiae*.

Saccharomyces cerevisiae is known as a Crabtree-positive yeast, meaning even in aerobic conditions it preferentially ferments glucose and fructose to ethanol instead of utilising the more efficient respiratory pathways when those sugars are present in high concentrations (see Chapter 1.5.2; Postma *et al.*, 1989). This property is shared with other species that diverged after the WGD, along with the ability to grow under strict anaerobic conditions and the generation of stable respiratory-deficient mutants called "petites" (Merico *et al.*, 2007). Petite mutants are characterised by deletions of all or parts of the mitochondrial genome, disrupting aerobic metabolism(Piskur, 1994). The ability to form and maintain petite mutants is largely thought to be restricted to species within the *Saccharomycetacae* (Bulder, 1964), although *Candida albicans* has also been reported to give rise to petite mutant colonies more recently (e.g. Roth-Ben Arie *et al.*, 1998). The ability to form petites has often been suggested to be related to the metabolic capability to grow under strictly anaerobic conditions. A recent study examining the coincidence of those capabilities has indeed found that the post-WGD species, displaying the most efficient anaerobic metabolisms, were all (with the exception of three species that diverged soon after the WGD) able to generate petite mutants (Merico *et al.*, 2007). The pre-WGD species in turn display a mosaic pattern with respect to being able to grow anaerobically and generation of petites, suggesting that the basic metabolic prerequisites for both these processes were present in the last common ancestor of the *Saccharomycetaceae* but have been fine-tuned after the WGD (Merico *et al.*, 2007). Indeed, it has been suggested that fine-tuning of the anaerobic metabolism through an increased ratio of glycolytic enzymes after the WGD and a resulting increase in concentrations has led to an increase in glycolytic flux in the post-WGD species, increasing the efficiency of fermentation to an energetically favourable level (Conant & Wolfe, 2007).

Petite formation and the biochemical and physiological processes behind it however remain less well-understood. The fact that amongst the pre-WGD species examined by Merico *et al.* (2007), there were species that could not grow under anaerobic conditions, even on rich media, but were still found to be petite-positive such as *Kluyveromyces wickerhamii*, or in reverse were petite-negative but could grow anaerobically, such as *S. kluyveri*, suggests that while the ability

to grow anaerobically even on minimal medium probably aids the maintenance of petites there are additional mechanisms in place controlling the response to loss of all or parts of the mitochondrial genome. Abf1p has recently been implicated as an important regulator in intergenomic signalling between the mitochondrial and nuclear genomes in *S. cerevisiae* (Woo *et al.*, 2009). Cells lacking mitochondrial DNA experienced downregulation of genes involved in mitochondrial respiration and oxidative phosphorylation, TCA cycle, amino acid metabolism and the transcriptional regulator *CAT8*, responsible for derepression of genes required under non-fermentable growth conditions (Woo *et al.*, 2009). Woo *et al.* (2009) specifically linked the downregulation of *COX6*, a subunit of the cytochrome c oxidase, to regulation by Abf1p. Indeed, *COX6* expression levels have previously been shown to be dependent on the phosphorylation status of Abf1p (Silve *et al.*, 1992). Out of the 66 genes that were found downregulated through intergenomic signalling, 26 contained a binding motif for Abf1p in their promoters suggesting a broader role amongst those (Woo *et al.*, 2009).

Considering these results in light of the phylogenetic distribution of *ABF1* (Fig. 3.8) one can hypothesise about the involvement of intergenomic signalling and the resulting downregulation of genes involved in respiratory metabolism in the ability to generate stable petite mutant colonies. Repression of genes superfluous to respiratory metabolism conceivably confers an energetic advantage to the cell, being able to use those resources elsewhere, and might thus aid the success of respiratory-deficient mutants. The presence of *ABF1* in the pre-WGD species, yet petite-negative status of all pre-WGD species included in my dataset (see Merico *et al.*, 2007) suggests that if intergenomic signalling mediated through Abf1 plays a role in formation and maintenance of petites, such mechanisms have arisen after the WGD. Although the pathways involved in intergenomic signalling are not well understood at this stage, the fact that *COX6* expression level in *S. cerevisiae* is dependent on phosphorylation status of Abf1 raises the possibility of this being part of the communication pathway between mitochondrion and nucleus.

Examination of the alignments of the conserved C-terminal domains in Figure 3.7 revealed two candidate regions that are phosphorylated in *S. cerevisiae* and conserved amongst the post-WGD species to varying extent but absent from the

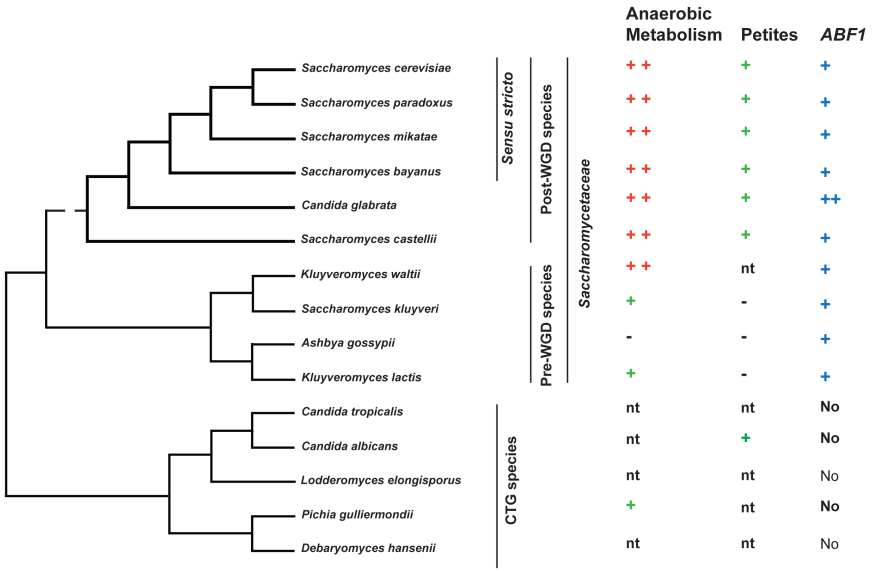


Figure 3.8: Phylogenetic distribution of *ABF1* (+: single copy; ++: two copies) and the ability of different species to grow under strict anaerobic conditions (++: minimal media; + rich media / low growth rate) and to form petite mutants. "nt" means not tested. See text for references. The WGD event is indicated by a star.

pre-WGD species (Fig. 3.7, blue asterisks). The first phosphorylation site is found just upstream of C1 and is the target of a DNA-damage kinase (Smolka *et al.*, 2007). The site is conserved across all the *Saccharomyces sensu stricto* species but is either too divergent to align or absent in *C. glabrata* and *S. castellii* — although the region just downstream shows strong conservation and it is possible that one of the serines nearby is still phosphorylated in these species. C3 also harbours a phosphorylation site that is absent from pre-WGD species, although in this case the surrounding region is highly conserved across most of the alignment. In addition to that, several other sites in ScAbf1 are know to be phosphorylated and might contribute to intergenomic signalling. Merico *et al.* (2007) also found three species that diverged after the WGD and are not included in my study to be petite-negative and examining alignments of *ABF1* homologs in those, if any, would be interesting to focus hypotheses about involvement of regions of the protein in petite formation. Unfortunately no genome sequence is available for any of these species to date.

The example of *ABF1* thus clearly demonstrates how flexible regulatory networks in yeasts are and shows how a novel transcription factor has become an integral part of the gene regulatory network and cellular functions beyond within just 150 - 200 million years of evolution. Furthermore, the phylogenetic profile assembled here provided sufficient resolution to be able to form first hypotheses about the evolutionary significance of this pattern of presence and absence within the biological context of each species. Examination of sequence alignments gave an example of asymmetrical divergence after gene duplication with one of the copies evolving to perform new, or a subset of, functions compared to its ancestor. Also, I found ample evidence for potential neofunctionalisation through changes in posttranslational modification such as phosphorylation or nuclear localisation, reflecting the functional versatility of Abf1p in *S. cerevisiae*. Indeed, a recent study found that highly posttranslationally modified genes showed higher rates of survival after WGD (Amoutzias *et al.*, 2010), indicating that this potential divergence in posttranslational regulation might have been an important mechanism allowing for the retention of both paralogs in *C. glabrata*.

3.6.2 Loss of a Putative Carbohydrate Metabolism Regulator in the *Sensu stricto* Species

The regulation of nutrient utilisation is known to have been rewired extensively in the evolutionary history of the species studied here (e.g. Askew *et al.*, 2009; Hittinger & Carroll, 2007; Lavoie *et al.*, 2009). Besides the emergence of *ABF1*, which amongst other functions plays a role in regulation of glycolysis (see above), in the pre-WGD lineages there were numerous other examples that showed lineage-specific patterns of abundance such as *SGT1* (Fig. 3.6; note 2). *SGT1* is likely to be involved in regulation of carbohydrate metabolism Kainou *et al.* (2006); Sato *et al.* (1999a) but has not previosuly been implicated so in the species considered here.

Expression of glycolytic genes in *S. cerevisiae* is mediated by an activatory complex consisting of Gcr1 and Gcr2 that acts in cooperation with the general TF Rap1 (Uemura & Jigami, 1995). This activatory complex is present across the *Saccharomycetacae* but absent from the species that diverged before (Askew *et al.*, 2009; Haw *et al.*, 2001). Regulation of glycolysis in *C. albicans*, which lacks *GCR1* and *GCR2* homologs, has recently been shown to be regulated by the TFs Tye7 and Gal4 (Askew *et al.*, 2009) that are known to perform different roles in *S. cerevisiae*. Functional studies in *K. lactis*, a facultative aerobic species that diverged before the WGD, indicate that its homologs of *GCR1* and *GCR2* are indeed functional homologs of ScGcr1 and ScGcr2 and as such involved in regulation of glycolytic genes (Haw *et al.*, 2001; Neil *et al.*, 2004). Furthermore, it has been shown that the *K. lactis* homolog of Tye7, Sck1, is also required for high-level expression of glycolytic genes, suggesting the regulatory network controlling glycolysis to be in an intermediate state with respect to *S. cerevisiae* and *C. albicans* (Neil *et al.*, 2007). *S. cerevisiae* also encodes a homolog of Tye7 which has been found to be able to complement a *gcr2* null mutant and rescue a growth defect phenotype in the presence of glucose when overexpressed but not a *gcr1* null mutant (Sato *et al.*, 1999b). Its functional role in *S. cerevisiae* is otherwise not well-characterised, however.

Similar to the complementation of ScGcr2 by ScTye7, Sato *et al.* (1999a) isolated a human protein that is able to complement a *gcr2* null mutation. This

protein, HsSgt1 encodes a transcriptional coactivator that has no homologs in *S. cerevisiae* (Sato *et al.*, 1999a). (Note that there is a gene called *SGT1* in *S. cerevisiae* which encodes a cochaperone protein that is however not homologous to the *SGT1* referred to here). Although *S. cerevisiae* does not encode an *SGT1* homolog, my dataset includes homologs of *SGT1* in most other species and it appears to have been lost from the *Saccharomyces sensu stricto* lineage (Fig. 3.6; note 2). Experimental analysis of the *SGT1* homolog in the basal ascomycete *Shizosaccharomyces pombe* found *SpSGT1* to be essential for growth in haploid conditions and involved in regulation of carbohydrate metabolism on glucose-containing media (Kainou *et al.*, 2006). This argues for an ancestral role of *SGT1* in carbohydrate metabolism. Although there is currently a lack of experimental evidence, *SGT1* is a strong candidate to be an additional player in the regulation of nutrient utilisation in the *Saccharomycotina*. If this is indeed the case, the reconfiguration of the regulatory network controlling carbohydrate metabolism in the post-WGD species might have led to its redundancy and subsequent loss in the *sensu stricto* lineage.

Also of interest with regards to the rewiring of the carbohydrate metabolism network is the fact that *C. glabrata* and *S. castellii* have each retained duplicate copies of both *GCR2* and *TYE7* and it would be interesting to experimentally investigate the possible functional consequences of that.

3.6.3 Increased Retention of WGD-duplicate TFs in *S. castellii* and *C. glabrata*

It has previously been reported that genes retained in duplicate after the WGD in *S. castellii* and *C. glabrata* but subsequently lost in the *Saccharomyces sensu stricto* species are enriched for transcriptional regulators (Scannell *et al.*, 2006). Examination of both DBD family-level (Fig. 3.5) and domain architecture-level (Fig. 3.6; red box) abundance patterns supports this observation, with frequent cases of greater abundance of certain architectures in either *S. castellii*, *C. glabrata* or both compared to the pre-WGD or the *sensu stricto* species. Based on these observed enrichments, I have compiled a list of transcriptional regulators exhibiting these patterns and their functional roles. This is shown in Table 3.5.

S. castellii **A**	*S. castellii* **B**	*C. glabrata* **A**	*C. glabrata* **B**	*S. cerevisiae*	**Function**
Scas492.3	Scas484.3	CAGL0M12100g	CAGL0E05566g	*TYE7*	Transcriptional activator in Ty-mediated gene expression; Potential role in activation of glycolytic genes; Compensates for GCR2 null mutation
Scas718.27	Scas635.12	CAGL0M01716g	CAGL0F04081g	*TEC1*	Transcription factor required for full Ty1 expression, Ty1-mediated gene activation, and haploid invasive and diploid pseudohyphal growth
Scas629.12	Scas630.6	CAGL0E00561g	CAGL0C03608g	*TUP1*	General repressor of transcription
Scas584.13	Scas646.6	CAGL0M01254g	CAGL0H02145g	*STE12*	Transcription factor that is activated by a MAP kinase signaling cascade, activates genes involved in mating or pseudohyphal/invasive growth pathways; cooperates with Tec1p transcription factor to regulate genes specific for invasive growth
Scas719.24	Scas551.1	CAGL0M09955g	CAGL0D00682g	*SFP1*	Transcription factor that controls expression of ribosome biogenesis genes in response to nutrients and stress
Scas655.16	Scas677.7	CAGL0G05467g	CAGL0G05379g	*GCR2*	Transcriptional activator of genes involved in glycolysis
Scas668.25	Scas673.3	CAGL0M10087g	CAGL0K02585g	*YAP3*	unknown
Scas707.31	Scas602.9	CAGL0K02145g	CAGL0L06072g	YER130C	unknown
Scas643.23	Scas548.5	CAGL0H06215g	CAGL0F00803g	*GAL11*	General transcription factor
Scas488.2	—	CAGL0B04895g	CAGL0H03751g	*RFX1*	Major transcriptional repressor of DNA-damage-regulated genes, recruits repressors Tup1p and Cyc8p to their promoters; involved in DNA damage and replication checkpoint pathway
Scas573.4	—	CAGL0B03421g	CAGL0K05841g	*HAP1*	Zinc finger transcription factor involved in the complex regulation of gene expression in response to levels of heme and oxygen
Scas697.35	—	CAGL0I02838g	CAGL0L03916g	*AZF1*	Zinc-finger transcription factor, involved in induction of CLN3 transcription in response to glucose
Scas588.1	Scas686.10	CAGL0J10956g	—	*SUM1*	Transcriptional repressor required for mitotic repression of middle sporulation-specific genes; also acts as general replication initiation factor
Scas666.8	Scas697.43	CAGL0G06688g	—	*MET4*	Leucine-zipper transcriptional activator, responsible for the regulation of the sulfur amino acid pathway
Scas713.11	Scas718.67	CAGL0K01727g	—	*RPN4*	Transcription factor that stimulates expression of proteasome genes; Rpn4p levels are in turn regulated by the 26S proteasome in a negative feedback control mechanism; RPN4 is transcriptionally regulated by various stress responses
Scas709.65	Scas625.5	CAGL0L01903g	—	*RGT1*	Glucose-responsive transcription factor that regulates expression of several glucose transporter (HXT) genes in response to glucose
Scas613.8	Scas669.4d	CAGL0F07667g	—	*MSS11*	Transcription factor involved in regulation of invasive growth and starch degradation; controls the activation of MUC1 and STA2 in response to nutritional signals
Scas721.136	Scas717.63	CAGL0J01177g	—	*ABF1*	DNA binding protein with possible chromatin-reorganizing activity involved in transcriptional activation, gene silencing, and DNA replication and repair
Scas720.58	Scas699.7	—	—	*GAL4*	Transcriptional activator involved in control of galactose metabolism

Table 3.5: List of transcription factors retained in duplicate in either *S. castellii*, *C. glabrata* or both. Functional annotation was based on annotation of the corresponding *S. cerevisiae* gene from the Saccharomyces Genome Database (SGD).

Considering this list of duplicates, I found a striking number of TFs known to be directly involved in carbohydrate metabolism. Out of the 19 examples listed here, five of the genes (*TYE7*, *GAL4*, *GCR2*, *AZF1* and *RGT1*) are involved in the regulation of glycolytic genes or galactose metabolism. As mentioned above, *TYE7* and *GAL4* as well as *GCR2* are among the major players in control of glycolytic genes in *C. albicans* and *S. cerevisiae* respectively, with the pre-WGD species likely to display a "mosaic network" that includes regulatory pathways involving both the Gcr1/Gcr2 complex and the Tye7 homolog (Askew *et al.*, 2009; Haw *et al.*, 2001; Neil *et al.*, 2004). *AZF1* positively regulates *CLN3* in *S. cerevisiae* in high-glucose conditions which in turn influences the timing of cell-cycle progression leading to rapid cell growth in glucose medium (Newcomb *et al.*, 2002). *RGT1* is involved in regulation of hexose transporter genes, having a repressive role in low-glucose conditions but switching to an activating role when hyperphosphorylated by the glucose sensing pathways, ensuring high-levels of glucose transport into the cell (Kim & Johnston, 2006). While *TYE7* and *GCR2* are retained in duplicate in both species, *AZF1* is only present in two copies in *C. glabrata* and conversely *RGT1* and *GAL4* WGD duplicates have only been retained in *S. castellii*. Notably, *C. glabrata* is entirely lacking a *GAL4* homolog and indeed has lost the ability to metabolise galactose (Hittinger *et al.*, 2004). Together this suggests that the mechanisms for the control of carbohydrate metabolism might be quite different in those species.

The post-WGD yeasts distinguish themselves from their relatives that have not undergone a WGD by the ability to ferment glucose efficiently even in aerobic conditions (see Fig. 1.9; Merico *et al.*, 2007). This is thought to be facilitated partly by an increased glycolytic flux in the post-WGD species (Conant & Wolfe, 2007) as well as efficient glucose repression of genes involved in alternative resource utilisation pathways in high-glucose conditions (Carlson, 1999; Merico *et al.*, 2007). The duplicate retention of such a considerable number of important carbohydrate metabolism regulators in the two species that diverged just after the WGD possibly reflects an initial phase of "sorting out" where new regulatory connections were established by diverging duplicate copies of the TFs, followed by subsequent adaptation of the regulatory networks towards the more specialised aerobic-fermentative lifestyle and loss of functionally redundant TFs.

Other examples of retained WGD-duplicates include *TEC1* and *STE12*, which are both involved in regulation of haploid invasion and diploid pseudohyphal growth which represent distinct developmental stages in yeast (e.g. Köhler *et al.*, 2002; Zeitlinger *et al.*, 2003). Interestingly, *MSS11*, also in the list of retained duplicates in *S. castellii* has been shown to be important in regulating invasive growth, sharing regulatory targets with *TEC1* and *STE12* (van Dyk *et al.*, 2005). This "cluster-like" retention of functionally related TFs may be indicative of rewiring events when one considers the argument that in order to retain the duplicate pair sufficient levels of regulatory specificity (e.g. through differential expression of distinct sets of target genes) for either copy are needed. Alternatively, dosage effects between interacting TFs could be an explanation for coordinated retention (Papp *et al.*, 2003) although the non-uniform patterns of retention between *S. castellii* and *C. glabrata* as well as the subsequent loss of the duplicates argues against this theory and the functional significance, if any, of such clusterings remains to be determined.

The list also includes *SFP1* which is involved in upregulation of ribosome biogenesis in respiro-fermentative conditions and as such provides a link between the *AZF1* duplicate mentioned above and ribosome biogenesis in general which is also known to have undergone extensive rewiring between *C. albicans* and *S. cerevisiae* (Cipollina *et al.*, 2005; Lavoie *et al.*, 2009, 2010).

In addition to those, we find a number of general regulators *TUP1*, *GAL11* and *ABF1*, probably facilitating this initial increase in regulatory proteins in dosage dependent manner. Other examples include *RFX1*, a major repressor of DNA-damage regulated genes (Zaim *et al.*, 2005) and various others that I will not discuss here.

3.6.4 Lineage-specific Amplifications in the CTG-clade and the Evolution of Pathogenicity

Based on the enrichments observed in Figure 3.6, I investigated the domain families showing lineage-specific amplification in the CTG-clade encompassing the *Candida* species as well as *Lodderomyces elongisporus*, *Pichia guilliermondii* and *Debaryomyces hansenii*. Amongst the families that showed a net increase in size

were TFs containing an NDT80-like DBD, RFX DBD containing proteins, the zf-C_2H_2 TFs containing three repeats of the motif, as well as $Zn(II)_2Cys_6$ zinc finger-containing proteins. Although a large number of the TFs that I found to be lineage-specific across the CTG clade did not have functional annotation, all the annotated examples were of relevance to the evolution of pathogenicity in the *Candida* species and included regulators of hyphal development as well as TFs involved in drug resistance phenotypes. This is perhaps unsurprising seeing that there is a great study bias towards mechanisms of pathogenicity and drug resistance but probably also reflects the broadness of developmental and metabolic processes involved in allowing those yeasts to survive in a mammalian host (reviewed in Chapter 1.5.2.2).

Candida albicans is a commensal pathogen dwelling on the skin and mucosal surfaces of the gastrointestinal and urogenital tract of most humans. Whilst it usually represents a relatively harmless companion, serious *Candida* infections can have detrimental outcomes, especially in immunosupressed patients, showing greater than 40% morbidity (Eggimann *et al.*, 2003). Being a eukaryotic organism, the range of possible drug targets is relatively limited due to the similarity to its human host. Furthermore, *C. albicans* has evolved a variety of mechanisms conferring drug resistance, making effective treatment yet more difficult (Cowen & Steinbach, 2008). Consequently, drug resistance has been the subject of intensive study and I will briefly outline the main mechanisms implicated so far below.

The primary targets of antifungal drugs are ergosterol, the main sterol of fungal membranes, its biosynthesis pathway, or the biosynthesis of (1,3)-β-D-glucan cell wall components (reviewed in Cowen & Steinbach, 2008). The triazole-like compounds are the most widely-studied antifungal drugs in *C. albicans* and a lot of the current knowledge about drug resistance derives from the study of the pathways mediating this. Triazoles work by blocking Erg11p, a late enzyme in ergosterol biosynthesis, which leads to the accumulation of toxic intermediate products and eventual cell death. *C. albicans* has evolved several independent mechanisms to overcome the toxic effects of triazole treatment. These include modifications to Erg11p itself, compromising triazole-binding as well as upregulation of the enzyme and amplification of copy number to sequester most of the

drug whilst retaining a functional ergosterol biosynthesis pathway, and the up-regulation of different types of multidrug transporters increasing drug efflux out of the cell (Cowen & Steinbach, 2008). Furthermore, a variety of cell signalling processes have been shown to be important in triazole resistance, often in a complex and highly interconnected manner influencing the downstream expression of stress-response genes involved in small-molecule transport, cell wall integrity and vesicular trafficking (Cowen & Steinbach, 2008).

NDT80 (Figure 3.6; note 3) is found in single-copy in all of the *Saccharomycetaceae* and in *S. cerevisiae* is known to be a key modulator of progression of meiotic division (reviewed in Sellam *et al.*, 2010). I found the species in the CTG clade in turn to encode two paralogs of *NDT80*, arising before differentiation of this clade. This has also been reported recently by Sellam *et al.*, 2010. One of the *C. albicans* homologs, CaNdt80, is an important regulator of hyphal development responsible for both activation and repression of genes in the yeast-to-hypha switch with an *ndt80* null mutant being unable to form hyphae which results in attenuated virulence (Sellam *et al.*, 2010). CaNdt80p has been shown to bind to ∼23% of genes in *C. albicans* and is thought to be a major regulator in this species (Sellam *et al.*, 2009). Besides its role in hyphal development it has been found to bind several multidrug transporters as well as conferring highly significant upregulation of genes involved ergosterol metabolism, including *ERG11* itself, implicating its involvement in drug resistance (Sellam *et al.*, 2009). This gives CaNdt80p a central role in virulence as well as drug resistance and consequently this gene duplication might be an adaptation involved in allowing the *Candida* species to infect mammalian hosts.

When examining the domain architecture of the two paralogous copies of *NDT80*, Sellam *et al.* (2010) found that while one of the paralogs maintained the same order of domains as the homolog in *S. cerevisiae*, with an N-terminal DBD followed by a C-terminal activation domain (AD), the other copy shows a reversed domain order. Interestingly, the functionally characterised copy of the paralogs, *CaNDT80*, refers to the paralog bearing the "non-canonical" domain arrangement while the role of the TF containing the *ScNDT80*-like domain arrangement is unknown; attempts at finding growth conditions under which it is being expressed have been unsuccessful so far (Sellam *et al.*, 2010).

Furthermore, I investigated amplifications in the (zf-C_2H_2)x3 lineage. Although most *C. albicans* ORFs in this class did not contain functional annotation, those that did included *CRZ1* and *CRZ2*, paralogs of *CRZ1* in *S. cerevisiae*. CaCrz1 is a downstream target of several signalling pathways, including calcineurin signalling (Karababa *et al.*, 2006; Miyazaki *et al.*, 2010) and CK2 signalling (Bruno & Mitchell, 2005). Both of these have been shown to be important for triazole resistance in *C. albicans*. CaCrz2, in contrast, is not targeted by either pathway according to current knowledge but has been shown to be important in mediating response to changes in pH (Kullas *et al.*, 2007). Again, this duplication has occurred in a common ancestor to the species in the CTG clade suggesting early functional fine-tuning of these signal integrators that might have been important in facilitating the evolution of pathogenicity of the *Candida* species.

Lastly, the RFX DBD-containing proteins show increased abundance in the CTG species. Again, *S. cerevisiae* contains a single homolog containing this DBD, whereas the CTG species encode two copies. In contrast to the previous examples, however, this duplication is likely to have predated the divergence of the *Saccharomycotina*, followed by a loss in the *Saccharomycetaceae*. This inference derives from consideration of the conservation of the DBD in multiple sequence alignments of the two CTG paralogs and their homologs in the *Saccharomycetaceae* (not shown). ScRfx1 is a downstream target of the DNA damage response pathway and its phosphorylation results in derepression of the DNA damage response genes (Zaim *et al.*, 2005). The function of the more closely related *C. albicans* homolog CaRfx1 has not been determined to date but it has been shown that CaRfx2 shows some functional redundancy with ScRfx1 through complementation studies (Hao *et al.*, 2009). Apart from the shared role in DNA damage response, CaRfx2 has also been implicated in regulation of hyphal genes, with an *rfx* null mutant showing overexpression of hypha-specific genes under non-hypha inducing conditions as well as the inability to revert back to yeast morphology once hyphal growth has been induced. The null phenotype also showed attenuated virulence in mouse models, suggesting the importance of flexible control of morphological state in response to changing conditions for pathogenicity (Hao *et al.*, 2009). I found that *C. glabrata*, a post-WGD species that also shows

pathogenicity, also contains two copies of *RFX1*. These have arisen through the WGD according to my data as well as syntenic information (Byrne & Wolfe, 2005) and any significance of this for virulence is speculative.

Although there was a clear bias with respect to experimental information available for TFs in *C. albicans* due to the medical relevance of hyphal development and drug resistance and the associated research effort, it is still striking that all the examples I examined, i.e. those containing functional annotation, are in one way or other related to pathogenicity in the *Candida* species. In this light, it is also noteworthy that while the amplifications discussed above are shared among all species in the CTG clade, not all of them are pathogenic. Whereas *C. albicans* and *C. tropicalis* are highly pathogenic, *L. elongisporus* and *P. guilliermondii* display weak levels of pathogenicity and *D. hansenii* is mostly non-pathogenic (Butler *et al.*, 2009). This raises the possibility that changes in hyphal developmental control and stress signalling were probably not causative of pathogenicity but had a facilitating role.

3.7 Conclusions

I have developed and implemented a pipeline for the collection of TF repertoires in the *Saccharomycotina*. While sensitivity of the DBD pipeline was good, achieving approximately 90% of coverage when averaged across DBD families and genomes, specificity was relatively low with a high proportion of false positives. This was likely to have been partly due to the often short and degenerate sequence motifs of some DBDs, such as the C_2H_2 zinc fingers, where detection cutoffs were difficult to define. Also, domains often found in TFs (both DBDs and accessory domains) occur promiscuously and regions of strong homology are often confined to relatively short segments that can lead to false homology assignments (see Chapter 4.2 for further discussion). As such, the dataset needed extensive manual curation. Nevertheless, I have collected a high-quality dataset of transcriptional regulators in 15 species of yeasts.

The analysis of TF repertoire composition across the *Saccharomycotina* allowed me to gain an overview of TF evolution and the role of individual gene families in these species. The relative ratio of TF to protein-coding genes is fairly

stable in the species studied here, ranging from about 3.5 to 4.9%, and the overall contribution of different domain families is similar. However, even at this coarse resolution of evolutionary dynamics it was possible to identify numerous examples hinting at both the evolutionary plasticity of transcriptional regulatory networks as well as the functional importance of those changes. I have found evidence for mechanisms suggesting a slow, gradual turnover of regulatory "responsibilities" as illustrated by changes in the regulatory network controlling carbohydrate metabolism as well the potential for fast adaptation to abrupt changes such as the complete loss of *GAL4* in *C. glabrata* or the acquisition of many, highly interconnected regulatory interactions within relatively short evolutionary time as illustrated by *ABF1*.

More detailed analysis of the *ABF1* example furthermore suggested that evolutionary novelty, adaptive or not, is often created outside the DBD and while this is anecdotally reported frequently, it will be interesting to explore this systematically. An analysis of relative evolutionary rates between DBD and non-DBD regions in the data collected here is presented in Chapter 5.4.

Species that diverged after the WGD did not generally have noticeably larger TF repertoires than their pre-WGD relatives. Nevertheless, it was clearly visible that *C. glabrata* and *S. castellii*, the species that diverged very soon after the WGD, had the tendency to retain different numbers and types of WGD homologs. This differential retention of WGD duplicate TFs in those two species is not unexpected but rather follows genome-wide patterns of post-WGD gene retention, where the WGD was followed by series of independent gene loss in the lineages leading to the extant species (e.g. Scannell *et al.*, 2006). It however implies that all of the TFs that have remained in duplicate in any of the post-WGD species have been retained in duplicate until at least after the divergence of that species, followed by independent losses. This initial excess of TFs possibly allowed for rapid asymmetric divergence and acquisition of new regulatory interactions that led to subsequent loss of consequently redundant regulators and as such was facilitating rapid rewiring of the regulatory network. A putative example of this would be the loss of *SGT1* in the *Saccharomyce sensu stricto* species (discussed above) which might have become redundant in regulating carbohydrate metabolism after the WGD. It would thus be of great interest to study those two

species in detail seeing that their genomes might bear the hallmarks of this initial phase of "sorting" or regulatory connections and the way WGDs can facilitate regulatory evolution. This has however not been attempted here.

Especially in low-abundance DBD families, it was possible to examine changes in family size in context of the biology that defines the species in a particular clade even without inference of evolutionary histories of groups of homologous TFs. Regulation of carbohydrate metabolism, for example, is known to have changed drastically, mirroring the origin of Crabtree-positive yeasts, and I found much evidence for this in my data. Although the phylogenetic distribution of a lot of the important players had been previously described individually, considering TF repertoires in a systematic way allowed for the identification of a novel potential regulator of carbohydrate metabolism based on functional studies in other species and its phylogenetic distribution in relation to other known regulators. I also found numerous virulence-associated amplifications in the CTG clade, suggesting the importance of regulatory evolution in the establishment of pathogenicity.

Overall those results yielded a number of experimentally testable hypotheses, such as a possible involvement of *ABF1* in the ability to maintain stable petite mutants or the impact of the retention of two WGD copies of *RFX1* on virulence in *C. glabrata*. This gave a promising perspective for the more detailed evolutionary analyses of the TF repertoires as well as individual regulatory pathways described in the following chapters.

Chapter 4

Evolutionary Dynamics in Transcription Factor Repertoires

4.1 Introduction

The distribution of DNA-binding domain (DBD) families and their abundance patterns across different lineages and species have provided interesting insights into the evolution of transcription factor (TF) repertoires and underlined their important role in the evolution of the traits distinguishing the species of *Saccharomycotina* included in this study (see Chapter 3.6). This initial analysis also suggested that turnover can be rapid, as shown for example by the complete loss of the important metabolic regulator *GAL4* in *Candida glabrata*, a gene central to galactose metabolism in *Saccharomyces cerevisiae* (Martchenko *et al.*, 2007) and glucose metabolism in *Candida albicans* (Askew *et al.*, 2009). Consideration of abundance patterns alone however only reveals changes in absolute numbers of copies of regulators within each family rather than the number and history of events leading to the observed number of copies. A copy number profile that is the same in all species can be an example of a very stable gene family without any gain or loss events as well as of a family that has undergone constant turnover but maintained a stable number of copies in each genome over time. In order to examine to a fuller extent the evolutionary dynamics experienced by each of the DBD families, I inferred duplications and losses for each of the orthogroups

collected using the DBD pipeline (see Chapter 3.2.1) and tested for enrichment of these events in individual families and orthogroups.

4.1.1 Inference of Duplication and Losses

The number of gene duplications and losses for a given gene family can be estimated by comparison of its gene tree with the species tree relating the taxa being studied. The most commonly used methods for inference of duplications and losses employ the reconciliation of the proposed gene tree with the species tree (e.g. Chen *et al.*, 2000; Zmasek & Eddy, 2001) and are based on a parsimony principle, minimising the total number of duplication and loss events that need to be inferred in order to map the gene tree into the species phylogeny (see Figure 4.1A). These methods however strongly rely on the assumption that both gene tree and species tree are correct. While we can infer species trees with reasonable confidence based on the availability of data from many loci across the genome, estimation of gene trees is a challenging problem due to the relatively small amount of data available for their inference and the resulting artefacts in tree reconstruction that can lead to, sometimes highly supported, wrong trees (see Chapter 2 for an in-depth discussion). Besides these methodological issues, biological processes such as independent lineage sorting or horizontal gene transfer can lead to incongruences between gene and species trees (reviewed in Degnan & Rosenberg, 2009). Reconciliation of such incongruent trees forces the placing of duplications on deeper branches of the species phylogeny, resulting in an increased number of proposed duplications close to the root of the tree and consequently an over-inference of losses on branches near the tips (Figure 4.1B; Hahn, 2007). The same gain cannot happen twice independently, so a single gain is inferred at a deeper node in the tree. Multiple losses are possible however and are then inferred to correct the observed pattern of occurrences.

Currently available approaches for gene tree reconstruction and speciation-duplication inference (SDI) can be classified into three categories (see Figure 4.2). Firstly, reconciliation methods employ the *ad hoc* approach of reconciling a reconstructed gene tree with a species tree. The advantages of such methods

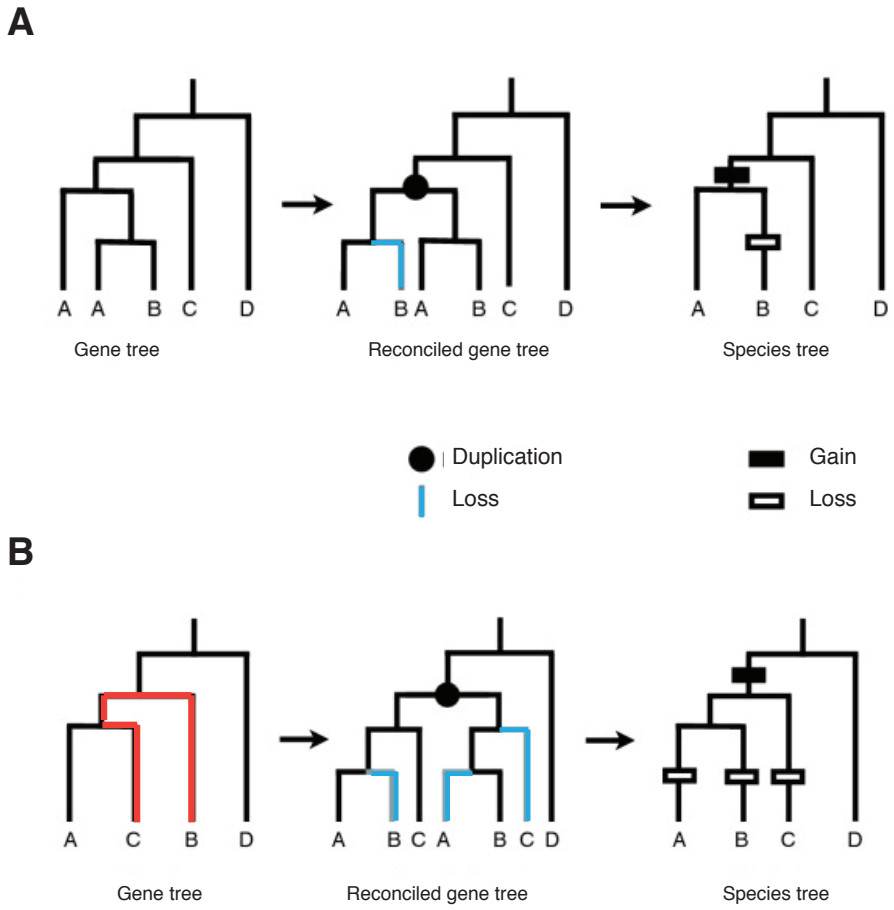


Figure 4.1: Reconciliation of gene trees and species trees. Figure modified from Hahn (2007). **A**: Mapping of gene tree into the species tree through the inference of duplications and losses. Here a duplication event has occurred in the ancestor of A and B, with subsequent loss of one of the duplicates in B. **B**: Inference bias arising from the reconstruction of incorrect gene trees. Here no duplications or loss events have happened, but in order to reconcile the erroneous gene tree (note red branches) a duplication needs to be inferred in the ancestor of A, B and C. This also forces the inference of three independent loss events. Duplications and losses on the middle trees indicate the actually inferred duplication and loss events, whereas symbols on the right ("gain" and "loss" indicate the directionality of change in that clade.

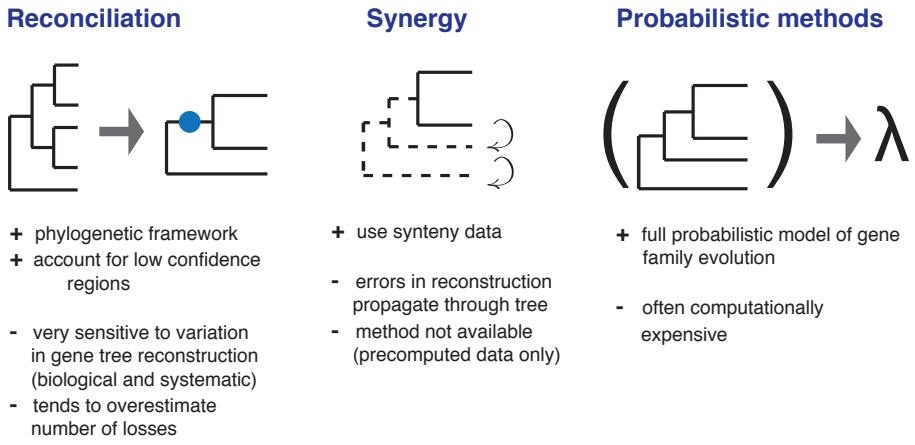


Figure 4.2: Gene tree reconstruction approaches.

clearly lie in the ability to use sophisticated phylogenetic methods such as Maximum Likelihood (ML) and a wealth of evolutionary models. Furthermore, these methods are generally fast and calculation of statistical support through bootstrapping can be easily parallelised. It is these *ad hoc* methods, however, that suffer most from the biases discussed above and even if bootstrap support is incorporated when reconciling gene and species tree, they are prone to over-inference of duplication and, especially, loss events (Hahn, 2007).

Alternative methods include SYNERGY, the algorithm behind the Fungal Orthogroups Repository (FOR) described in Chapter 3.2 (Wapinski *et al.*, 2007a). SYNERGY is a two-step procedure, performing both the initial homology assignment as well as gene tree reconstruction. Genes are grouped into a graph structure by sequence similarity and synteny conservation ("gene similarity graph") and, given a species tree, orthogroups (OGs) containing one-to-one or one-to-many orthologous genes with respect to the root (see Chapter 3, Fig. 3.2 for a definition) are reconstructed recursively from leaves to root (Fig. 4.3). This approach has the advantage of making use of synteny information, which especially in the *Saccharomycetacea*, is highly conserved (Dujon, 2010). The disadvantage of such recursive reconstruction, however, is the propagation of erroneous homology assignments during an early step in further iterations of the algorithm. As will be discussed below, I found this to be particularly problematic in clades where

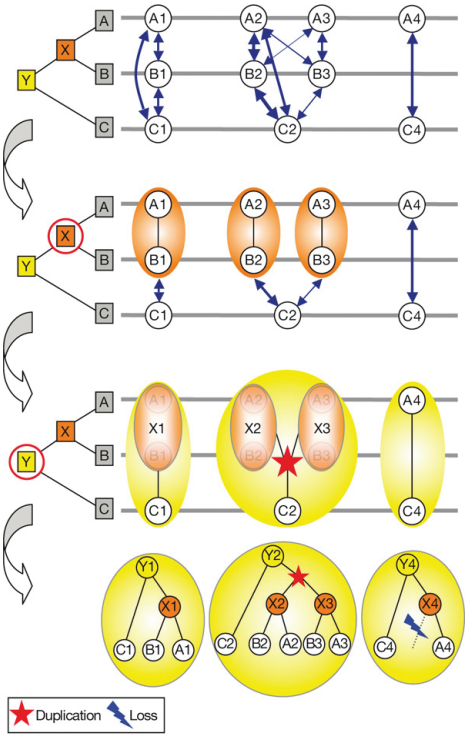


Figure 4.3: The SYNERGY pipeline. In the first step, genes are grouped into homologous clusters called "gene similarity graphs" by sequence similarity and synteny conservation. Subsequent steps recursively group connected homologs at each interior node of the species tree and infer the branchings at order at each level using neighbour-joining analysis if more than two intermediate OGs map to the same level. Figure modified from Wapinski *et al.* (2007b).

high-quality synteny information is lacking, leading to frequent misclustering of non-orthologous sequences due to local homologies.

Finally, more recently a number of full probabilistic methods for gene tree reconstruction have emerged (e.g. Akerborg *et al.*, 2009; Rasmussen & Kellis, 2010). Given a species phylogeny with branch lengths, these methods model a birth and death (BD) process of gain and loss of gene copies inside the species tree. Based on estimated rates of birth and death (which serve to define a prior on a given gene tree) and a substitution model to derive the phylogenetic relationships between the sequences, they find the most likely gene tree given the species tree and the data using Bayesian approaches. In an extension to the work by Akerborg *et al.* (2009), Rasmussen & Kellis (2010) further introduced gene- and species-specific evolutionary rate multipliers allowing for heterogeneity in the evolutionary process between genes and over time. Prime-GSR (Akerborg *et al.*, 2009) proved to be very computationally expensive and as such not suited for the analysis of the over 300 OGs collected here. For this reason I had initially concentrated on reconciliation approaches and SYNERGY for the reconstruction

of gene trees but later incorporated SPIMAP (Rasmussen & Kellis, 2010) when it was released earlier this year.

In the first part of this chapter I will perform an in-depth comparison of these three different types of tree reconstruction methods with respect to the consistency and accuracy of the inferences derived from the reconstructed gene trees as well as potential effects on downstream analyses. The second part is devoted to the analysis of results obtained using SPIMAP, the method that I found to perform best. In an extension of the work presented in the previous chapter, I will discuss patterns of duplications and losses in different clades and DBD families. Furthermore, I will relate those findings to the structure of the regulatory network and comment on the consequences of the whole-genome duplication (WGD) on network architecture.

4.2 Reassessing FOR

In order to assess the quality of the FOR dataset, I manually inspected the multiple sequence alignments of all 314 OGs which were realigned using a domain-anchored version of PRANK (version 081107; Löytynoja & Goldman, 2005, 2008). Briefly, PRANK was modified to use domain coordinates for shared conserved domains as softbound anchors for the alignment. This meant that corresponding positions in two subalignments ("anchors") had to be aligned within five columns of each other in the resulting alignment. I mapped the coordinates of Pfam domains found in the sequences to be aligned to each of the intermediate progressive alignments in each of PRANK's iterations. These coordinates were used to guide the alignment. If a domain was present more than once, homologous domains were determined based on sequence similarity of the domain-only region of the intermediate alignments. If domain coordinates for the two subalignments were inconsistent, e.g. overlapping or in a different order, anchors were dropped, defaulting to the standard PRANK algorithm. I had implemented this based on the observation that often only DBDs share sufficient homology to be confidently aligned and when using other alignment programs such as MUSCLE (Edgar, 2004) or Mafft (Katoh & Toh, 2008), domain boundaries were often not accurately aligned. A later release of PRANK (version 090707) however gave almost

identical results to the domain-guided version and as such the domain-guided version will not be discussed here further in the interest of brevity.

Assessment of the realigned OGs revealed numerous cases of "misclustered" OGs that contained sets of sequences, clearly related to each other but not sharing regions of extended homology between them. One such example is provided in Figure 4.4, showing parts of an alignment of the orthogroup containing homologs encoding an NDT80-like DBD. Examination of the alignment clearly revealed a "bipartite" structure where subsets of sequences are related within but not between each other (blue and green boxes in Fig. 4.4). Annotation of the two *Saccharomyces cerevisiae* proteins included in each boxed set further supported a different origin of these two sections with the green set containing *NDT80*, a TF described in more detail in the previous chapter, and the blue set containing *LSM2* which is involved in RNA processing and does not contain an NDT80-like DBD (reviewed in Beggs, 2005). Overall, I found similar incidences of misclusterings in 74 other orthogroups thus affecting almost 24%, a considerable portion, of the collected dataset.

These missasignments were due to shortcomings in the SYNERGY algorithm which will be discussed in more detail below and explain why inclusion of FOR in the DBD pipeline (see Chapter 3.2.1) resulted in an increased number of false positive results. Instances like the one highlighted here were removed during the manual filtering process however and do not contribute to the analyses of the final dataset in Chapter 3.

Based on these observations, I decided to split OGs showing such patterns of non-relatedness between groups of sequences and reassess the newly created OGs based on their domain annotation. This reassessment led to the removal of 321 sequences that had been misclustered into 40 of the 74 orthogroups. In those cases, neither the domain nor functional annotation of a misclustered subset supported their TF status, resulting in their removal from the final dataset. The example discussed above is one of these cases. In the remaining 34 OGs, although misclustered, there was sufficient evidence for the TF status of all unrelated subsets in the OG and those were thus retained as separate OGs.

In order to gain a better understanding of the causes behind the failure of the SYNERGY algorithm used to assemble the FOR, I examined the new splits

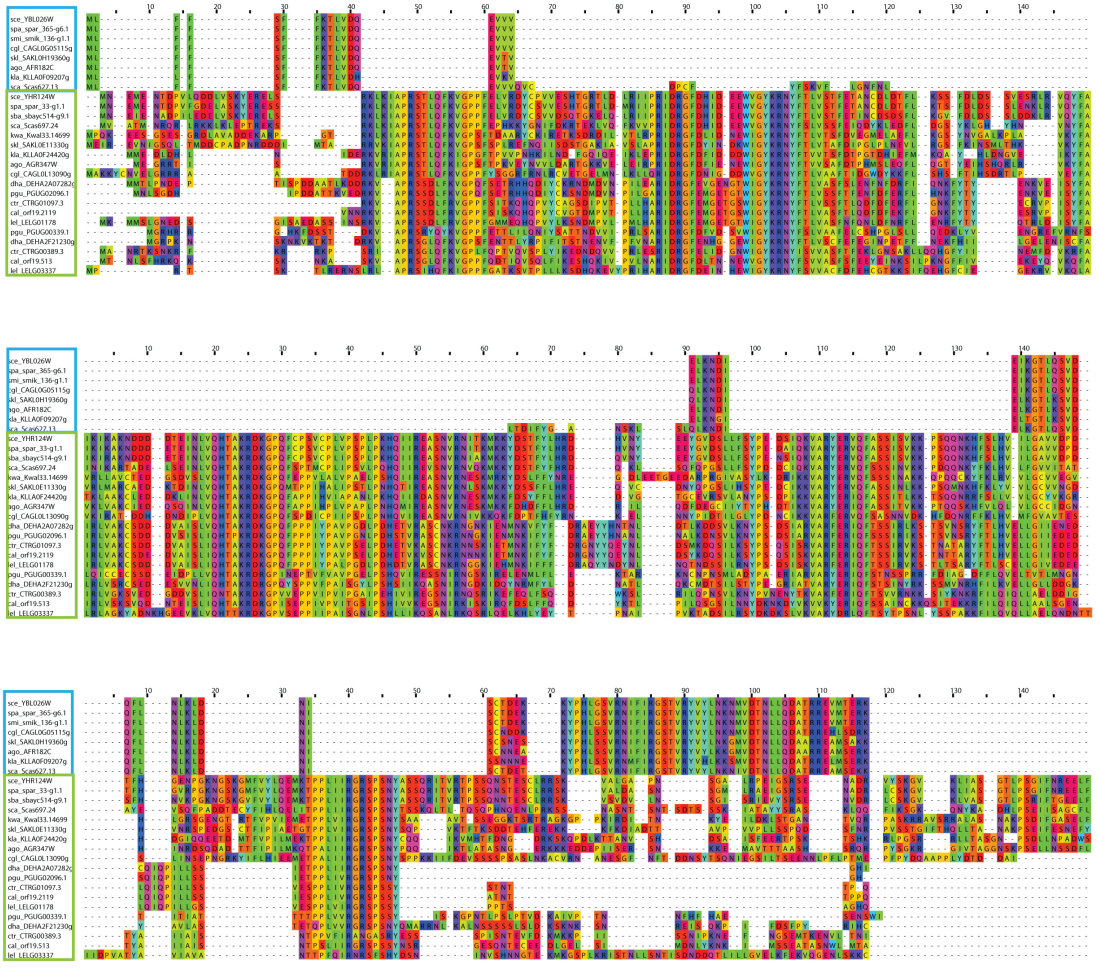


Figure 4.4: Multiple sequence alignment of FOR orthogroup encoding an NDT80-like DBD. Sequences were aligned using domain-guided PRANK and regions of low homology were trimmed manually for the purposes of visualisation. Blue and green sections highlight the misclustered sequences.

created with respect to their position on the phylogenetic tree. Figure 4.5 shows the (A) frequency of splits created by separating sequence clusters with questionable homology between them on the phylogenetic tree and (B) the corresponding node assignment. Generally, the bulk of newly created splits was concentrated at deep node levels (labelled N0, N1 and N11 in Fig. 4.5B) as would be expected given the recursive nature of the algorithm and the assumption that assignment over short evolutionary distances should be less error-prone due to longer alignments and less complicated evolutionary histories. By definition the number of gene duplication and loss events describing the pattern seen at the tips can only increase the further down we move into the tree.

The two most frequently occurring splits were between nodes N0 / N11 and N1 / N11 suggesting that the relationships between N11, which corresponds to the CTG clade (see Fig. 1.9), and other clades are problematic to assign for SYNERGY in general. This was also reflected by the fact that N11 was the most frequently created split. There are two plausible explanations for this phenomenon: (i) the divergence time between N11 and other nodes is relatively large and (ii) as opposed to the other clades, the CTG clade is not included in the Yeast Gene Order Browser (YGOB; Byrne & Wolfe, 2005) and assignments at this level are no longer guided by the high-quality synteny data available for the other species. This second theory is supported further by the fact that the following four most frequent splits (pgu, ctr, N13 and N15) all fall within the CTG clade, indicating that the synteny data contributed considerable amounts of information to the SYNERGY assignments. In line with this, the nodes below N1 showed relatively less frequent misclusterings, with most of the splits being created at very shallow node level and between nodes with large divergence time between them.

Given the above results, it appeared that the reason for these misclusterings was likely to be the sequence similarity threshold used to decide on inclusion or exclusion of an edge between two sequences from different species into the gene similarity graph that is used for reconstruction of orthogroups (see Fig. 4.3). Briefly, edges are placed between gene pairs from different species if their FASTA alignment E-value is below 0.1 and the corresponding sequence is the best FASTA hit in the other species or the percent identity between the two sequences is at least 50% of that between the sequence and the best FASTA hit in the other species.

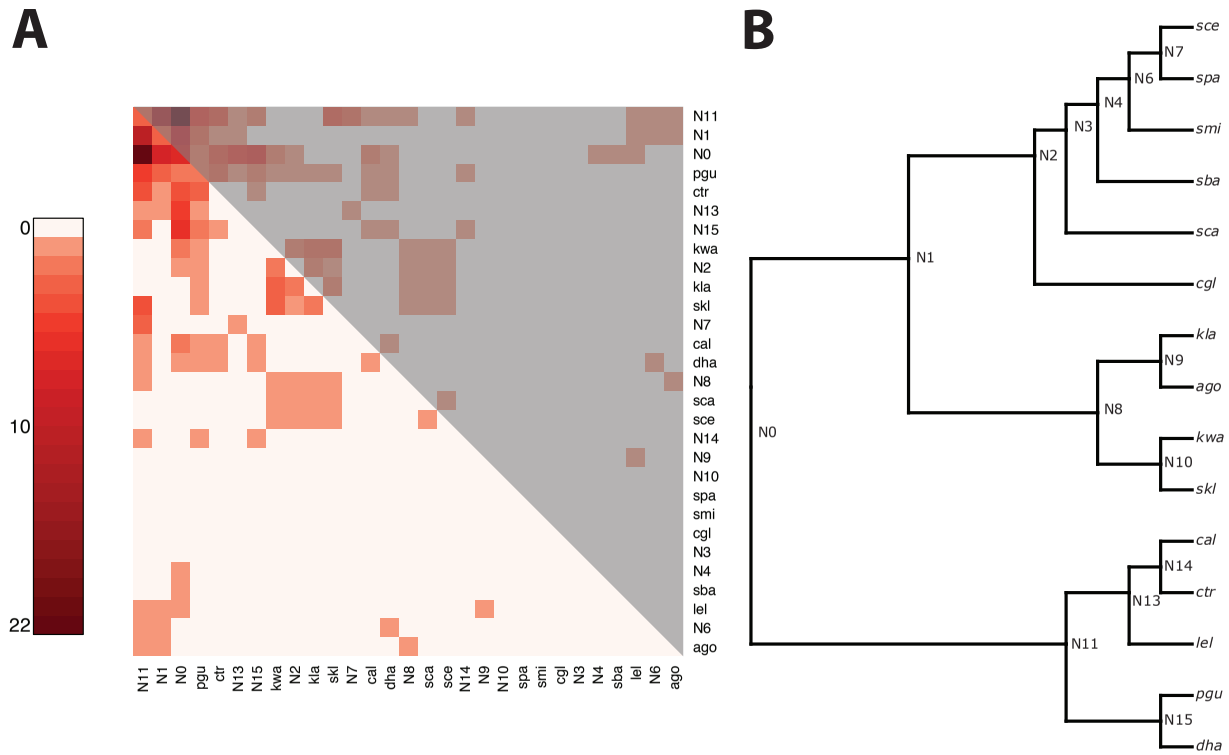


Figure 4.5: Node-level statistics of orthogroups that were separated into distinct subgroups based on examination of multiple sequence assignments. **A**: Pairwise node-level frequency of created splits, clustered by rows and columns. **B**: Node-level assignments on the species tree. Species are abbreviated in three-letter code e.g. *Saccharomyces cerevisiae*. Cells are shaded by the number of splits that were created between two different clades, e.g. there was one orthogroup that was split into a subset rooted at N11 and a subset containing a single *Ashbya gossypii* sequence resulting in a single (N11,ago) split (bottom left corner). This matrix is symmetrical by definition.

These edges are subsequently weighted using a composite distance score based on the ML distance and synteny similarity score between the pair of sequences (Wapinski *et al.*, 2007a). The authors themselves concede that this procedure is "lenient" and it is conceivable that erroneous edges are introduced through local similarities that will result in a significant E-value, especially when orthogroups (in the conceptual sense) are sparse (i.e. only contain members in few of the descendant species), as is reflected by the misclusterings at shallow node level. The synteny-based weighting of the edges also appears of great importance for the accurate resolution of the gene similarity graph into individual orthogroups as evidenced by the large number of misclusterings in the CTG clade for which no synteny information was available at the time. The authors however fail to give details about the relative weighting of peptide sequence similarity and synteny similarity score, making it difficult to further elaborate on this.

The splitting of the 74 misclustered orthogroups resulted in creation of numerous "fragmented" orthogroups rooted at a shallow node level. This could have led to a loss of information and over-inference of evolutionary events due to separation of true, but misclustered, orthologs into distinct orthogroups. To alleviate this as much as possible without introducing circularity through sequence similarity measures that appeared to be the initial reason for dubious homology assignment, I used an updated version of the YGOB (Gordon *et al.*, 2009) as well as the recently published Candida Gene Order Browser (CGOB; Fitzpatrick *et al.*, 2010) to screen for syntenic evidence supporting the reclustering of fragmented orthogroups. Both gene order browsers were collapsed into a single assignment of syntenic blocks based on shared *Saccharomyces cerevisiae* homologs, this being the only species within the *Saccharomycetacae* included in CGOB. I merged conserved syntenic regions from both gene order browsers if they contained the same *S. cerevisiae* sequence. Fragmented orthogroups were then searched for consistent mappings, requiring the agreement of all sequences in each fragmented orthogroup to map into the same, merged synteny block, and collapsed into a new orthogroup if this was the case. Although this approach is limited by the requirement of the presence of a syntenic ortholog between *S. cerevisiae* and the species in the CTG clade which are distantly related, this still resulted in 20 "reunited" orthogroups.

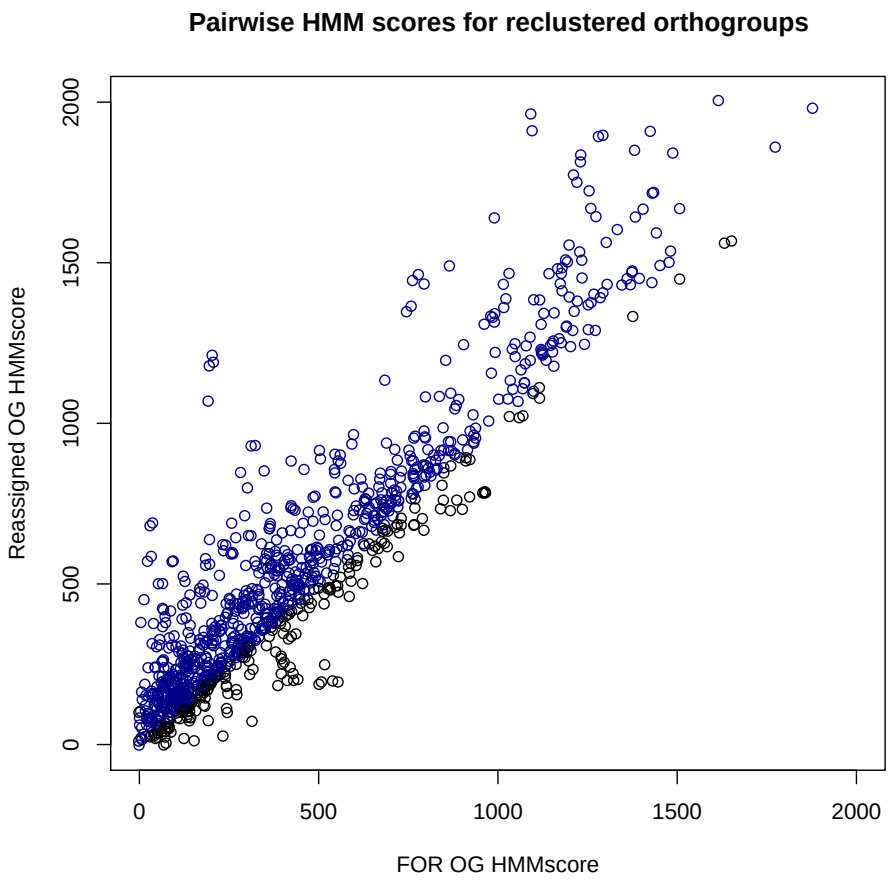


Figure 4.6: Pairwise HMM scores for each sequence belonging to one of the 75 reclustered orthogroups. Values in blue indicate a score ratio larger than one., in turn indicating a better fit to the new (reassigned) orthogroup.

Assessing the validity of reclustered orthogroups is a challenging problem since both alignments as well as trees have changed and as such are not comparable. To get an approximate measure of "relatedness" of a sequence to the orthogroup to which it is assigned, I chose to remove the sequence from the orthogroup in question, realign the remaining members and then score the sequence against the global, weighted HMM profile of the orthogroup. The HMM profile of an alignment is a statistical representation of the degree of conservation and consensus sequence at all positions in the alignment (Durbin *et al.*, 1998) and in theory a sequence should score better when globally aligned to a HMM containing true homologs rather than sequences sharing only short regions of local homology. HMM profiles of the realigned sequences were built using HMMER2 (Durbin *et al.*, 1998; Eddy, 1998) and the scores for the removed sequence aligned against the profile was determined using Profile Comparer (Madera, 2008). Figure 4.6 shows the pairwise scores for each sequence in its old and its newly assigned orthogroup for all members of all 75 modified orthogroups. Overall about 74% of sequences scored better in their newly assigned orthogroups. Surprisingly, this still left a quarter of sequences scoring worse however. Orthogroups containing sequences that scored lower in their new orthogroups included the example in Figure 4.4.

Here the sequences marked in blue were removed from the dataset due to a lack of evidence of those being TFs (see above). The removal of this unrelated subset from the alignment however also resulted in increased constraint on each position in the global HMM (due to the now higher percent identity/similarity). One of the *NDT80* paralogs in the CTG clade has a reversed domain order with respect to the other copies (see Chapter 3.6.4) and those "reversed" homologs score worse in the new OG due to the overall increased constraint. This highlights the difficulty of finding a suitable measure of "improvement".

The SYNERGY algorithm was previously found to perform similarly well on small and medium-sized orthogroups compared to computationally expensive probabilistic methods (Akerborg *et al.*, 2009) and in general in comparison to methods not incorporating phylogenetic information at all (Wapinski *et al.*, 2007a). Nevertheless, I found almost a quarter of the orthogroups in my dataset to suffer from erroneous homology assignments and inclusion of unrelated false

positives. Arguably this is due to the nature of the data itself, given the promiscuous occurrence of DBDs and other accessory domains commonly found in TFs in a range of different types of proteins. This in turn would have been equally likely to influence other methods relying on grouping homologs by sequence similarity such as best bidirectional hits (Fitch, 1970; Wall *et al.*, 2003) or OrthoMCL (Li *et al.*, 2003) and as such represents a particularly difficult case. After manual reassessment and splitting of dubious orthogroups I was confident that the dataset at hand contained few false homology assignments. Although it is likely that there is an increased occurrence of fragmented orthogroups which might lead to a slight over-inference of evolutionary events, due to having to infer losses to account for the "missing orthologs" which might in fact be present in another fragmented orthogroup, it is possible to account for this in downstream analyses and introduces less conflicting signal than analysis of non-orthologous proteins.

4.3 Widespread Disagreement between Speciation-Duplication Inference Methods

As discussed in the introduction to this Chapter, inference of gene trees and, implicitly, the inference of duplication and loss events is a difficult problem both due to the stochastic error inherent in phylogenetic reconstruction of relatively short sequences as well as the biological processes leading to incongruence between gene and species trees. When this work was begun there was little theoretical grounding on how to assess the performance of different gene tree reconstruction methods due to the lack of a sound statistical framework to model gene family evolution based on both gene duplication and loss as well as sequence evolution. Approaches such as PrimeGSR (Akerborg *et al.*, 2009) were beginning to address such questions; however they remained computationally infeasible to use on large datasets such as the one here. Very recently, SPIMAP (Rasmussen & Kellis, 2010), an implementation of similar theoretical principles as those PrimeGSR is based on (see above) was published and proved to be fast enough for large-scale analyses. Furthermore, Rasmussen & Kellis (2010) have developed a simulation

framework to assess Speciation-Duplication Inference (SDI) methods and I will discuss my performance measures in light of their empirical results.

To gain an insight into the performance of different SDI methods and the inferences they produce, I performed a comparative analysis between results obtained using four SDI methods employing different approaches for gene tree reconstruction. SYNERGY being one of these approaches, I decided to restrict the comparison of these methods to the 240 orthogroups that remained unchanged during reassessment of the dataset ("stable orthogroups") in order to avoid complication of the signal by possible false homology assignments. Of those stable OGs, 172 contained more than three members and were thus non-trivial to resolve. All comparative analyses were run on the 172 stable, non-trivial OGs. Reconstructed SYNERGY trees were taken directly from the FOR dataset. The species tree used in my analysis differs slightly from the one that was used for the FOR dataset, so in order to obtain comparable trees, nodes were rearranged to match the species tree used in my analysis. If there was more than one possibility for rearrangement, TreeBeST (see below) was used to resolve the subtree.

Secondly, I used Leaphy (Whelan, 2007) a ML software which I previously found to perform well in gene tree reconstruction compared to other popular phylogenetics software (see Chapter 2.3.1). Leaphy was run on the sequence alignments of the amino acid sequences calculated by the domain-guided version of PRANK (see above) using the WAG model of evolution (Whelan & Goldman, 2001) with four gamma rate categories and 100 bootstrap replicates each. The reconstructed gene trees were then reconciled with the species tree obtained from previous analyses (see Chapter 2) using NOTUNG 2.6 (Chen *et al.*, 2000), a software implementing the classic duplication-loss parsimony analysis outlined in the introduction. I used the default cost for duplications and losses (1.5 and 1.0 respectively). NOTUNG finds reconciliations by minimising the Duplication/Loss (D/L) score which is defined as $c_L L + c_D D + c_C C$ where L, D and C are the numbers of inferred losses, duplications and conditional duplications respectively, c_L is the cost assigned losses and c_D is the cost assigned to duplications. (C and c_C are defined as the number of inferred conditional duplications and the cost of conditional duplications, respectively. Those are however only relevant in the context of non-binary species trees which is not the case here and default to 0.)

Reconciled gene trees were rearranged using NOTUNG based on the bootstrap support at each node, with the support cutoff set to 70%. This means that nodes with bootstrap support of 70% or over are fixed irrespective of the cost they induce whereas edges with lower support are free to be rearranged to minimise the D/L score thereby mitigating the impact of low-confidence branchings in the reconstructed gene trees. This approach will be referred to as Leaphy$_N$ from now on.

In addition the two previously mentioned approaches, TreeBeST (Li *et al.*, 2006) was included in the comparison. Although TreeBeST also uses a duplication-loss parsimony reconciliation framework and as such falls into the same category as Leaphy$_N$, it employs a species-tree informed merging procedure of several input trees, calculated on both amino acid and nucleotide data, whilst minimising duplications and losses, thus making direct use of species relationships for the reconstruction of gene trees which represents an essential difference to Leaphy$_N$'s *ad hoc* approach. TreeBeST is widely used, being at the centre of the TreeFam database (Li *et al.*, 2006) and ENSEMBLCompara (Vilella *et al.*, 2009), and is generally considered a good gene tree reconstruction software, making it an interesting candidate to include into a comparative analysis. TreeBeST was run on nucleotide alignments, backtranslated from the amino acid alignments used in the Leaphy$_N$ analyses.

Finally, I ran SPIMAP 1.0 (Rasmussen & Kellis, 2010) on all stable, non-trivial orthogroups. SPIMAP is an empirical Bayes approach (Robbins, 1956) and requires training of species-specific evolutionary rates and the duplication and loss rate using a separate training script. I used the one-to-one ortholog phylogenomic dataset collected previously (see Chapter 2.2) as a training set for species-specific rates. Branch lengths to be used for training on each of the 343 training alignments were calculated using PAML (Yang, 2007) and the HKY model of evolution with four categories of Γ-distributed rates, this being the model of evolution implemented in SPIMAP. The duplication and loss rate parameters λ and μ were taken from Rasmussen & Kellis (2010), as estimated for their fungal dataset which spanned the same clades as my dataset. Values for λ and μ were 0.000732 and 0.000859, defined as the numbers of duplications and losses per unit branch length. Branch lengths are calculated in a relaxed clock model

that incorporates the gene- and species-specifc rate parameters mentioned above. Using the trained parameters, I ran SPIMAP with 500 iterations and 100 pre-screening iterations. On average, the algorithm converged on a topology after 186 iterations, indicating that this was sufficient.

4.3.1 Placement of Inferred Events

A rigorous comparison between the methods considered is difficult to achieve due to the lack of a "gold standard" dataset as well as the unavailability of the details of the SYNERGY method. The comparison presented here was mainly concentrated on the consistency of events inferred by the different approaches with respect to downstream analysis as well as the orthology assignments they induce. Table 4.1 shows a comparison of the total number of duplications inferred by each method. Since losses are placed as a consequence of where a duplication event has been inferred and as such are implicit, this first part of the comparison was focussed on duplication events only. Overall SYNERGY inferred about half as many duplication events as either the of two reconciliation methods. Approximately a third and a quarter of these events were shared with those inferred by Leaphy$_N$ and TreeBeST respectively and less than 15% have been inferred by all four methods. TreeBeST inferred most duplication events but showed less identically placed duplications than Leaphy$_N$ in all three comparisons, arguing for greater accuracy of the Leaphy$_N$ placement of duplications. Also noteworthy is the relatively large number of shared duplications inferred by Leaphy$_N$ and TreeBeST, two thirds of which were not recovered by SYNERGY. SPIMAP inferred similar numbers of duplication events to SYNERGY, but shared a larger proportion of identical duplications with both reconciliation methods. This higher consistency between the methods dwelling on full phylogenetic inference rather than recursive reconstruction might represent biases inherent to phylogenetic methods. Alternatively, this increased consistency might reflect a lack of sensitivity to more complex evolutionary scenarios by SYNERGY. Rasmussen & Kellis (2010) found such a lack of sensitivity in the case of gene conversion, presumably due to the relatively stronger weighting of synteny information over sequence similarity, which will be discussed further below.

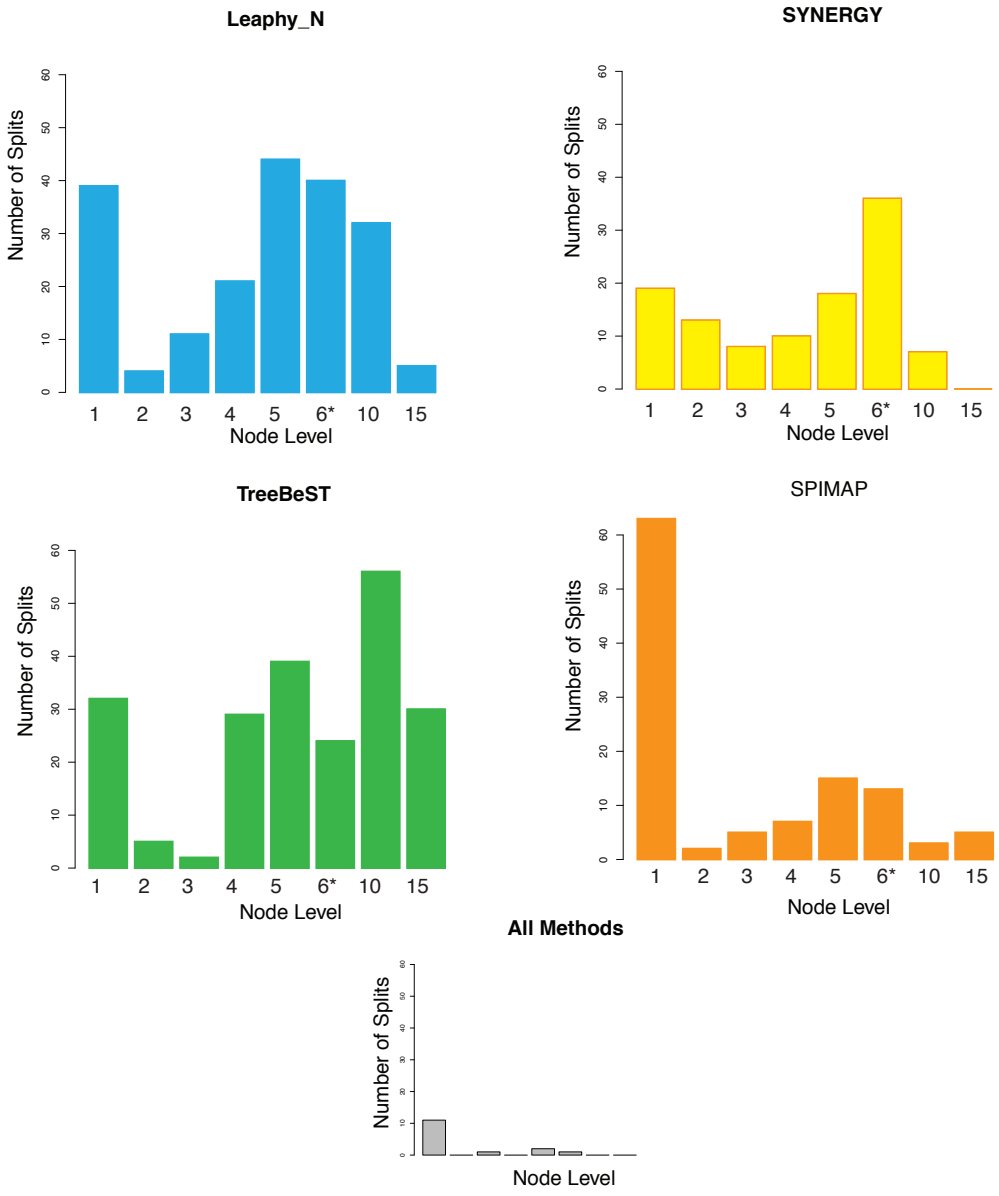


Figure 4.7: Placement of inferred duplications by Leaphy$_N$ (blue), SYNERGY (yellow), TreeBeST (green), SPIMAP (orange) and events shared by all methods (grey) according to the number of descendant species below that node. The category marked 6* only contains the WGD node.

		Number of Shared Duplications				
	Inf. Dups.	**SYNERGY**	**Leaphy$_N$**	**TreeBeST**	**SPIMAP**	**All**
SYNERGY	111	—	35(31.5%)	25(22.5%)	30(27.0%)	19(13.5%)
Leaphy$_N$	196	35(17.9%)	—	60(30.6%)	52(26.5%)	15(7.7%)
TreeBeST	217	25(11.5%)	60(27.6%)	—	35(16.1%)	15(6.9%)
SPIMAP	113	30(26.5%)	52(46.0%)	35(31.0%)	—	15(13.3%)

Table 4.1: Comparison between the numbers of duplications inferred using each of the four SDI methods tested.

It is known that reconciliation methods are biased towards over-inference of the number of duplications and frequent placement of those on deep branches whilst inducing excess losses towards the tips of the tree (see Chapter 4.1.1 and Hahn, 2007). To investigate whether this was the case here, I examined the exact placement of inferred duplications with respect to the species phylogeny. Figure 4.7 shows the number of duplications inferred by each method as well as those shared by all methods, according to the number of descendant species at each node level. Both Leaphy$_N$ (blue) and TreeBeST (green) inferred a large number of duplications on deep branches with many descendants. This trend was most extreme for TreeBeST, with 40% of all inferred duplications placed near the root of the tree (10 and 15 descendant species). Duplications inferred by SYNERGY (yellow) were slightly more evenly distributed across the branches with most duplications inferred at the WGD (six descendant species) whereas SPIMAP (orange) showed opposite trends with a large number of inferred species-specific duplications.

The majority of events that were consistently inferred by all methods (grey) were species-level duplications (one descendant species) which is in agreement with what has been proposed by Hahn (2007): the bias of reconciliation methods for excessive placement of duplications is strongest near the root while branches near the tips constitute "informative" branches on which duplications are accurately placed. The depth of inferred events in my study indicates that the large part of the additional duplications induced by Leaphy$_N$ and TreeBeST gene trees (see Table 4.1 for numbers) are likely to be due to incongruence between the reconstructed trees and the species tree, based on their placement on deep branches.

Neither SYNERGY nor SPIMAP were affected by this bias, although there was a large number of species-specific events inferred by SPIMAP but not SYNERGY. Interestingly, just under a third of duplications found by the three phylogenetic methods but not SYNERGY were species-specific and thus placed on informative branches, suggesting that SYNERGY might be "underinferring" duplications to some degree. Indeed, Rasmussen & Kellis (2010) found SYNERGY to be unable to detect gene conversion events in some cases, instead placing duplication events deeper within the tree. SPIMAP was shown to be more accurate than SYNERGY in recovering the real duplication age of gene conversion events (Rasmussen & Kellis, 2010) which might explain a portion of the large number of species-specific duplications it inferred.

To further examine the properties of the inferred duplications, I calculated the "Duplication consistency score" first introduced by Vilella *et al.* (2009). This score is calculated as the ratio of the overlap of descendant species present in the two subtrees below each duplication node to the total number of species below the duplication node (see Figure 4.8A for an example). The rationale of this score is based on the fact that frequently, misplaced duplications are due to relatively small incongruences between the gene tree and the species tree that force the placement of a duplication during reconciliation (Hahn, 2007; Vilella *et al.*, 2009). Such erroneously placed duplications will have a very small overlap in descendant species and a score near zero, while well-placed duplications will often be more balanced and score closer to one. Arguably, it is possible that all but a few paralogs have been lost following gene duplication such as in the examples outlined in Chapter 3.6.3. This would result in a small consistency score but can generally be expected to be rare. Furthermore, the inclusion of SYNERGY, which is not affected by the same gene tree reconstruction artefacts (see above) should provide an approximate "baseline" frequency of such events in the data.

The distributions of duplication consistency scores inferred by SYNERGY (yellow), Leaphy$_N$ (blue), TreeBeST (green) and SPIMAP (orange) are shown in Figure 4.8. While for most score categories the numbers of duplications falling into each category were very similar for all four methods considered, there was

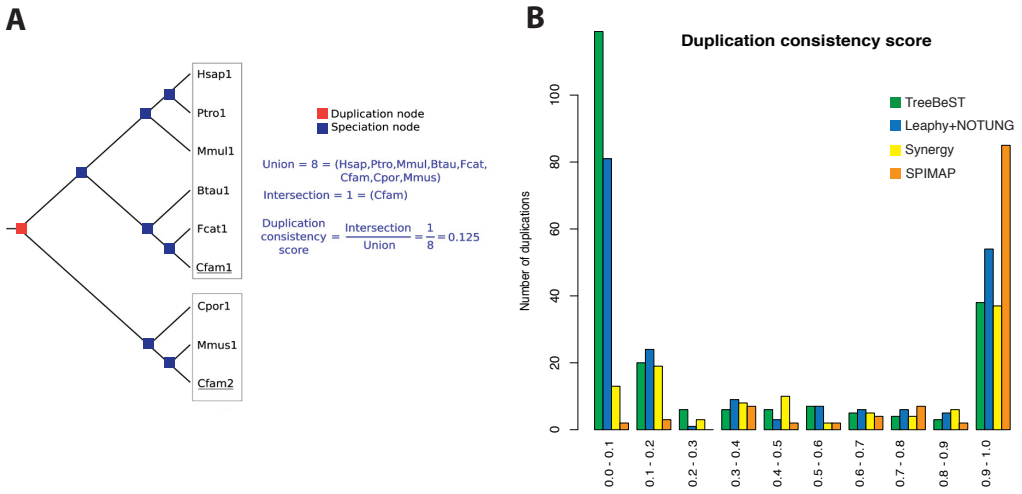


Figure 4.8: **A**: Calculation of the duplication consistency score and illustrated example. Figure modified from Vilella *et al.* (2009). **B**: Distributions of the duplication consistency score for all duplications inferred by SYNERGY (yellow), Leaphy$_N$ (blue), TreeBeST (green) and SPIMAP (orange). Scores near zero indicate low consistency whereas scores close to one indicate high consistency.

a very clear excess of low-consistency duplications proposed by both the reconciliation methods. In fact, the majority of additionally inferred duplications appeared to be accommodated by this category. Again, TreeBeST inferred more low-consistency duplications than Leaphy$_N$.

SPIMAP showed opposing trends with a very large number of high-consistency duplications. This was unsurprising given the large number of species-specific duplication events it inferred, as those are, by definition, of high consistency. Similarly, I also found an increase in abundance of high-scoring duplications in Leaphy$_N$ and to some extent TreeBeST, which reflected the additional species-level duplications proposed by the phylogeny-aware methods but not SYNERGY (Figure 4.7).

4.3.2 Orthology Assignments

Finally, phylogenetic and functional downstream analyses require the assignment of orthology relationships. By definition, one-to-one orthologs are separated only by speciation, not duplication events. In order to assess sensitivity and specificity of the orthology assignments induced by the proposed duplications of all four methods, I compared the pairwise orthology assignments to the syntenic orthologs annotated in CGOB (Fitzpatrick *et al.*, 2010). (Unfortunately, a parseable version of YGOB was not available for comparison as the raw data available via the website are unassembled homology assignments and it was not possible to determine orthology status of genes in the post-WGD genomes without execution of the YGOB code.) The CGOB gene browser contains manually curated orthology assignments based on the syntenic context of the individual genes as well as sequence similarity and phylogenetic evidence (Byrne & Wolfe, 2005) and is considered to be of high quality. Here, raw results were directly parseable for orthology assignments due to the one-to-one syntenic relationships of the genomes included. CGOB contains the species in the CTG clade as well as homology assignments to *S. cerevisiae*. To avoid introducing bias, I chose to ignore those assignments however, as *S. cerevisiae* only shares limited gene order conservation with the species present in CGOB (Byrne & Wolfe, 2005). For the gene tree reconstruction methods, a pair of genes was considered to be orthologous when the

nodes that separate them on the tree are speciation events only. The gene order browser datasets are arranged into "pillars" of orthologous genes and "tracks" of chromosomal environments. In the case of CGOB orthology assignment is trivial seeing that there is only one syntenic track for all species. The results of this comparison are summarised in Table 4.2.

All four methods performed well in comparison to CGOB, recovering between 83% and 92% of the orthology assignments that were testable. Both reconciliation-based approaches showed lower sensitivity than either SYNERGY or SPIMAP which was likely due to the biased placement of events on deeper branches resulting in unbalanced duplications (see above). Again, SPIMAP and SYNERGY performed similarly well, although SPIMAP showed both increased sensitivity and specificity in the CGOB comparison compared to SYNERGY. Encouragingly, the occurrence of positive orthology assignments that were not found in CGOB was very low for all methods ($\sim$ 0.1%) suggesting that false positive orthology assignments are unlikely to be of great importance irrespective of the method used for downstream analysis. These results mirror what has been found in a recently published comparison of SDI methods (Rasmussen & Kellis, 2010).

4.3.3 Quantitative Comparison

One of the aims of this study was to identify DBD families and groups of homologous TFs that have undergone large amounts of turnover compared to others. The previous comparisons considering the dataset as a whole, showed that there were large differences between the inferences made by different gene tree reconstruction methods. To study the impact of method choice on the statistics derived from inferred numbers of duplications and losses when focussing on individual families, I compared those when abstracted to the DBD family and orthogroup level. While in the global comparison SYNERGY and SPIMAP appeared to be the most correct methods, judging by the placement and consistency of the proposed duplication events, problems in the clustering step in the assembly of the FOR dataset and unavailability of the method details meant that SYNERGY inferences could not be used for at least one-third of my TF dataset. Based on the statistics discussed above, SPIMAP thus was the method of choice. Figures

Method	#Ortholog pairs	%CGOB	%not CGOB	Number of Shared Inferences			
				SYNERGY	Leaphy	TreeBeST	SPIMAP
SYNERGY	6705	89.6	0.10	—	4639(69.2%)	3870(57.7%)	6394(95.4%)
Leaphy	5182	83.0	0.14	4639(89.5%)	—	3494(69.2%)	4890(94.4%)
TreeBeST	4474	83.0	0.16	3870(86.5%)	3494(78.1%)	—	4150(92.8%)
SPIMAP	7650	91.7	0.04	6394(84.6%)	4890(64.7%)	4150(54.3%)	—

Table 4.2: Comparison of orthologous pairs inferred by different methods, when compared to the Candida Gene Order Browser (CGOB) and to each other. The numbers represent "positive assignment", i.e. whether sequence A and B are orthologous, but do not consider negative assignments, i.e. sequence A is not orthologous to sequence B

4.9 and 4.10 show the numbers of duplications and losses respectively inferred by each method when grouped by DBD or orthogroup. The comparisons shown here are based on the correlation between inferences derived from SPIMAP results when compared to each of the other three methods; the remaining pairwise comparisons are shown the Appendix, Figs. A.1 and A.2.

Despite a difference in the overall numbers of inferred duplications (cf. Table 4.1) I found a good correlation between SPIMAP results and those of other methods when considering the proposed number of events grouped by domain architecture (Figure 4.9A). When removing the most extreme values, the correlations became weaker but were still highly significant and strong when compared to SYNERGY and Leaphy$_N$ results, but considerably weaker for TreeBeST inferences. This was also reflected in the comparisons between TreeBeST and results derived from either Leaphy$_N$ or SYNERGY inferences (Fig. A.1), further underlining the notion that TreeBeST results were generally more noisy for the dataset analysed here. Furthermore, SPIMAP inferences correlated more strongly with those from the other three methods than any of the other pairwise comparisons (Fig. 4.9, Fig. A.1), suggesting a greater amount of shared signal (or noise). These conclusions held when the inferred numbers of duplications were considered by orthogroup (Figure 4.9B), although again correlations were less strong, probably due to larger amounts of stochastic error that had less strong effects when inferences were grouped by domain architectures seeing that DBD groupings often contain multiple OGs.

The correlations between methods became less strong when considering the numbers of inferred losses. This was unsurprising, given that differential placement of the same number of duplications will induce different numbers of losses. The effect of this became especially apparent when considering the number of inferred events by orthogroup (Figure 4.10B). Nevertheless, correlations were still significant especially when large domain architecture groups were included (Figure 4.10A). Here, SPIMAP showed stronger correlations with SYNERGY in contrary to the numbers of inferred duplications where SPIMAP showed stronger correlations with Leaphy$_N$, which reflected the Duplication Consistency Score and node level statistics where both SYNERGY and SPIMAP generally inferred

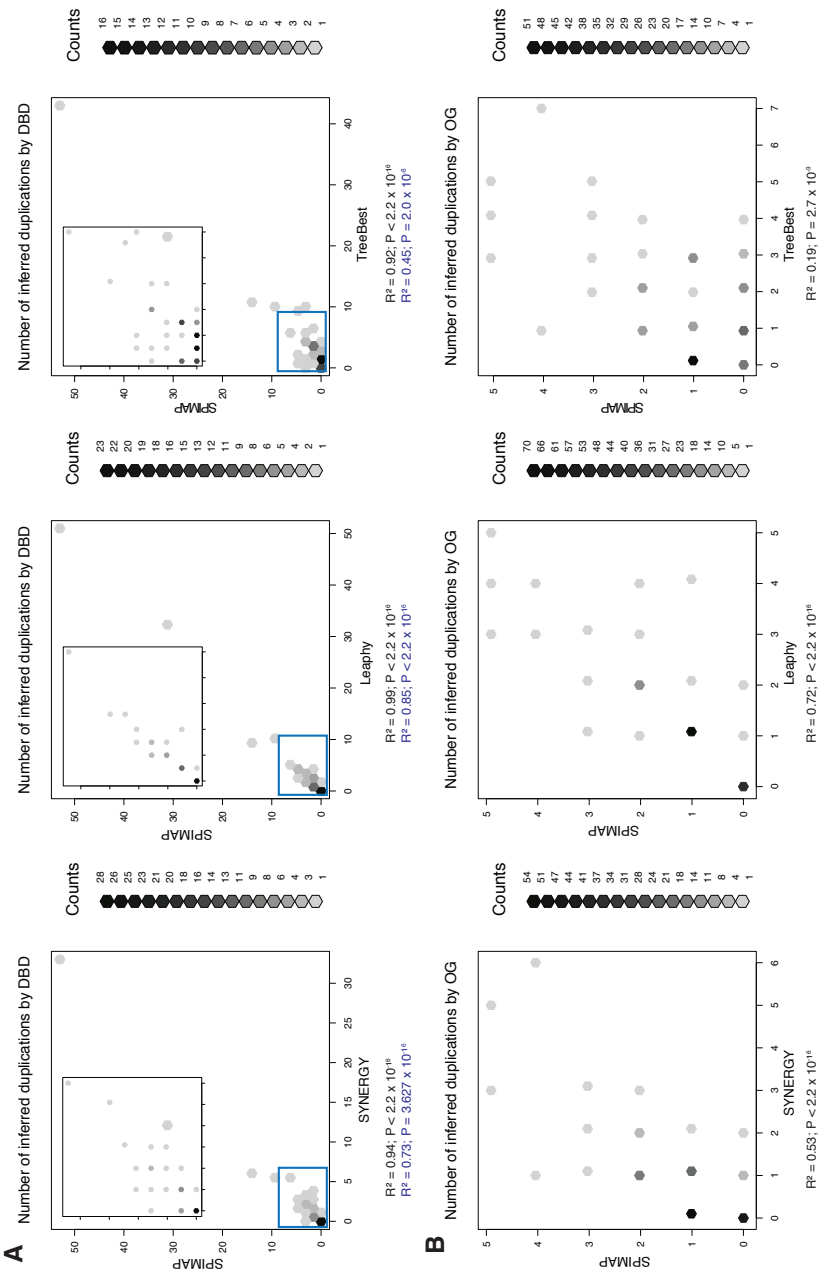


Figure 4.9: Three-way comparison between the number of inferred number of duplications when grouped by (A) DBD family and (B) orthogroup (OG). The sections boxed in blue delineate the close-ups shown within the respective plots. Hexagons are shaded based on the number of underlying data points. The correlation statistics shown beneath each plot were calculated on either the entire dataset (black) or the subset shown in the close-up (blue).

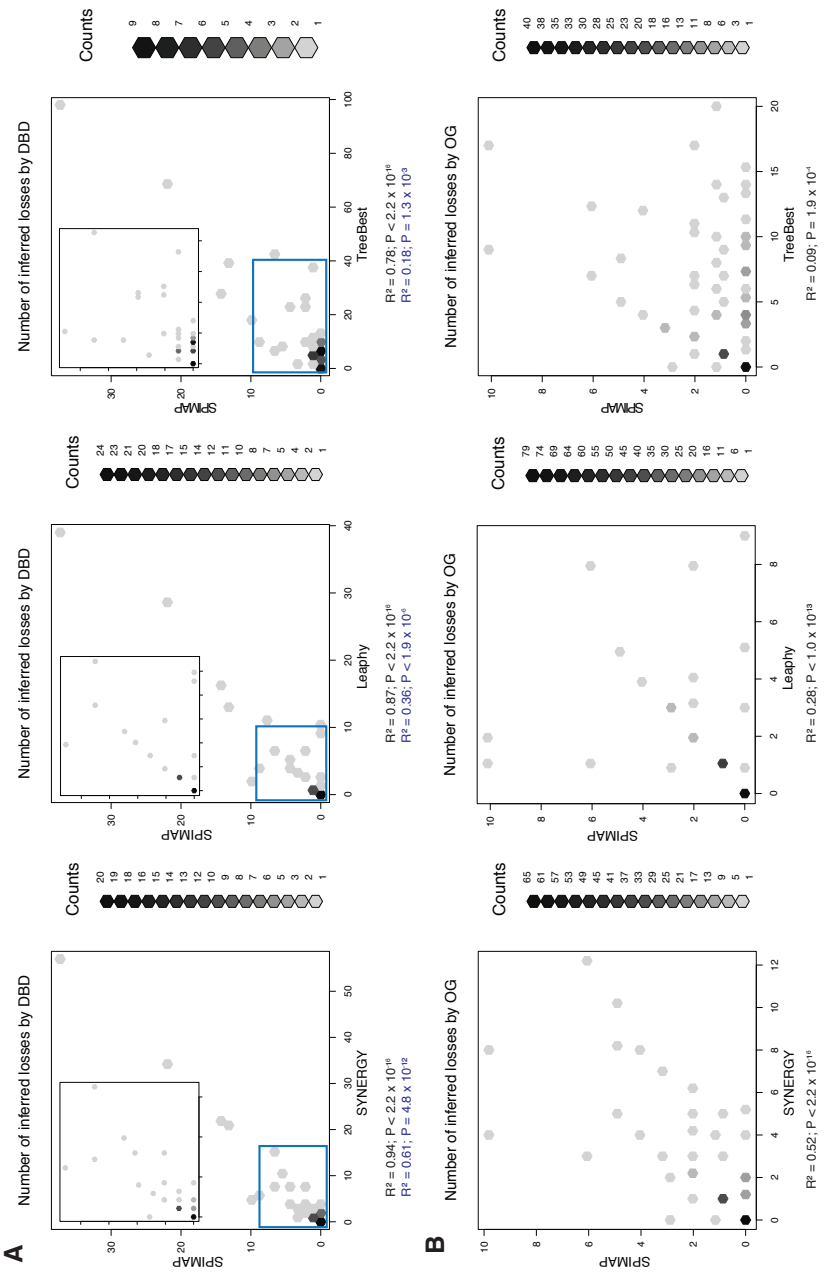


Figure 4.10: Three-way comparison between the number of inferred number of losses when grouped by (A) DBD family and (B) orthogroup (OG). The sections boxed in blue delineate the close-ups shown in the respective plots. Hexagons are shaded based on the number of underlying data points. The correlation statistics shown beneath each plot were calculated on either the entire dataset (black) or the subset shown in the close-up (blue).

duplications at more shallow node levels and of higher consistency thus inducing less loss events.

4.3.4 Conclusions

Inference of evolutionary histories of gene families involves two distinct analysis steps: firstly, the grouping of homologous sequences; and secondly, the reconstruction of gene trees for groups of given homologs. The SYNERGY algorithm includes both those tasks and while it has been shown to perform well compared to other methods for the former (Wapinski *et al.*, 2007a), there is somewhat contradictory evidence on performance in the latter task (e.g. Akerborg *et al.*, 2009; Rasmussen & Kellis, 2010). Because inferences made by SYNERGY are at the heart of the data collection pipeline used for this study (see Chapter 3.2.1) and the method promises increased accuracy in general due to incorporation of synteny information which is curated to a very high standard for some of the species included here (Byrne & Wolfe, 2005), SYNERGY was the initial method of choice. In order to assess performance in both steps, I investigated the consistency of the groupings of homologous genes, orthogroups, when compared to the available gene order browsers as well as the proposed gene trees and resulting inferences of duplications, losses and orthologs with respect to alternative widely-used gene tree reconstruction methods.

Careful examination of the realigned data revealed a substantial number of "misclustered" sequences, affecting approximately one-third of the collected orthogroups. Upon closer inspection I found that such artificial clustering frequently occurred between the *Saccharomycetaceae* and the CTG clade, possibly reflecting the large evolutionary distance between those species as well as the lack of high-quality synteny information. The most likely cause for the failure of the algorithm in those cases is the high BLAST E-value cutoff (0.1) used by SYNERGY for initial clustering of homologous sequences. This was possibly exacerbated by the nature of sequences themselves, often containing promiscuous but highly conserved domains accompanied by rapid sequence divergence outside those domains (see Chapter 5.4) which would have almost certainly been equally problematic for other clustering methods and probably presents one of the more challenging cases

to study due to those reasons. After reexamination, splitting and reclustering of the erroneously clustered orthogroups I obtained a high confidence dataset for further analysis.

Assessing the performance of SYNERGY as a gene tree reconstruction method was challenging for a number of reasons including the lack of full technical detail for the method and the fact that it performs two analysis steps simultaneously, making it difficult to hypothesise on potential shortcomings. In a comparison with their own probabilistic gene tree reconstruction method, Akerborg *et al.* (2009) found that SYNERGY performed well for small to medium-sized orthogroups but failed to recover gene trees for large orthogroups that were in the set of most probable trees proposed for those by probabilistic methods. It is difficult to discriminate, however, whether these results were due to similar misclusterings as mentioned above or the actual gene tree reconstruction process seeing that the authors used orthogroups calculated by SYNERGY as input for other software and especially large orthogroups were more likely to be affected by artificial clustering due to the resulting increase in evolutionary divergence between its members. Rasmussen & Kellis (2010) in turn found SYNERGY to perform very well, recovering over 99% of syntenic orthologs correctly with the only shortcomings found being the inability to detect gene conversion events when compared to methods using phylogenetic evidence.

In the comparison presented here SYNERGY performed well. It generally inferred fewer duplication events, showing high consistency and not heavily biased towards deeper branches of the tree, especially compared to the reconciliation methods Leaphy$_N$ and TreeBeST. SYNERGY's sensitivity for detection of orthologs as assessed by a comparison to CGOB, not included in the calculation of the FOR dataset, was close to 90%, suggesting that even in the absence of highly curated synteny information it can accurately infer orthology relationships, given that the homologs were clustered together in the first place.

The agreement on the exact placement of duplication events between different methods was very small, however, with at most 46% of inferred splits agreed by any two methods. This was especially true within the light of the orthogroups used for comparison, representing the portion of the dataset where no conflict in homolog clustering by SYNERGY was found and as such comprising the "easy

cases". Both of the reconciliation-based methods were found to exhibit the bias previously described, namely the forcing of duplications on deep branches of the tree due to errors in gene tree reconstruction (Hahn, 2007). This was reflected by the increased overall number of inferred events, low duplication consistency of the majority of those events, and the excess of duplications inferred on deeper branches of the tree. Leaphy$_N$ performed significantly better than the widely-used TreeBeST however, showing less extreme bias in all of those measures.

SPIMAP, a probabilistic and species tree-guided gene tree reconstruction approach, was later added to the comparison and performed well on all accounts. The overall numbers of inferred duplication events were very similar to those inferred by SYNERGY, showed no bias towards deep branches of the species tree and were of high consistency. There was an excess number of duplications inferred on the species-level, especially when compared to SYNERGY, although a proportion of those were also inferred by the reconciliation methods, indicating that those might represent cases of gene conversion where SYNERGY had been previously shown to fail to reconstruct accurate gene histories (Rasmussen & Kellis, 2010).

The failure of SYNERGY to appropriately cluster a considerable portion of the dataset meant that alternative methods had to be used for at least that part of the dataset. To examine how this would affect the outcome of planned downstream analysis, I investigated whether inferences made using different methods would influence conclusions drawn from the data such as enrichment for duplications and losses by orthogroup or domain architecture or orthology assignments. Although the placement of duplications frequently differed as described above, the overall numbers of duplications inferred were found to strongly correlate, especially when grouped by domain architecture. This meant that even when using a different method, inferring different numbers of duplication events in different locations, the relative numbers of duplications inferred for each DBD are similar between methods and would lead to equivalent results when comparing duplication rates between DBD families. The numbers of losses correlated less strongly but could still be considered to be informative. Although orthology inference varied in sensitivity, it was found to be highly accurate with ∼99% precision, minimising the risk of analysing non-orthologous sequences in downstream

analyses. Again, SPIMAP showed the strongest correlations of inferences of duplications and losses in pairwise comparisons between methods and performed best in terms of sensitivity and specificity of orthology assignments. I was thus confident to use SPIMAP for the following analyses.

4.4 Evolutionary Dynamics in TF Repertoires

4.4.1 Inference of Duplications and Losses

The full TF repertoire dataset comprised 391 OGs (the 240 stable OGs as well as 151 which were created as a result of splitting misclustered OGs; see above). Of those, 271 were non-trivial, meaning they contained more than three sequences, as for less than that SDI analysis is not meaningful. Gene trees for 271 non-trivial orthogroups were reconstructed using SPIMAP 1.0 (Rasmussen & Kellis, 2010) using the parameters determined previously (Chapter 4.3). SPIMAP reconstructs a gene tree that induces a reconciliation but does not output a reconciliation ("the mapping of events onto the species tree") itself. SPIMAP trees were reconciled with the species tree using NOTUNG (Chen *et al.*, 2000) to obtain a labelled species tree for each OG. Overall this resulted in the inference of 249 duplications and 303 losses. Figure 4.11 shows the distribution of inferred numbers of duplications and losses (red and blue, respectively) and the number of OGs rooted at a particular node level (yellow) across all DBD families. Recall, each OG contains one-to-one and one-to-many orthologous sequences that are inferred to be derived from a single sequence in the most recent common ancestor of the species included in an OG, the root. If the root species does not correspond to the most recent common ancestor of all species considered, I refer to the OG as an "orphan", rooted at the respective most recent common ancestor of the species containing representatives in the OG. (For a review of the orthogroup concept, see Fig. 3.2.)

The largest number of duplications, unsurprisingly, was inferred on the branch of the whole-genome duplication (Fig. 4.11). The lineages that diverged just after (*Saccharomyces castellii*, *Candida glabrata*, and the *Saccharomyces sensu stricto* species) also showed increased numbers of duplications, although these are likely

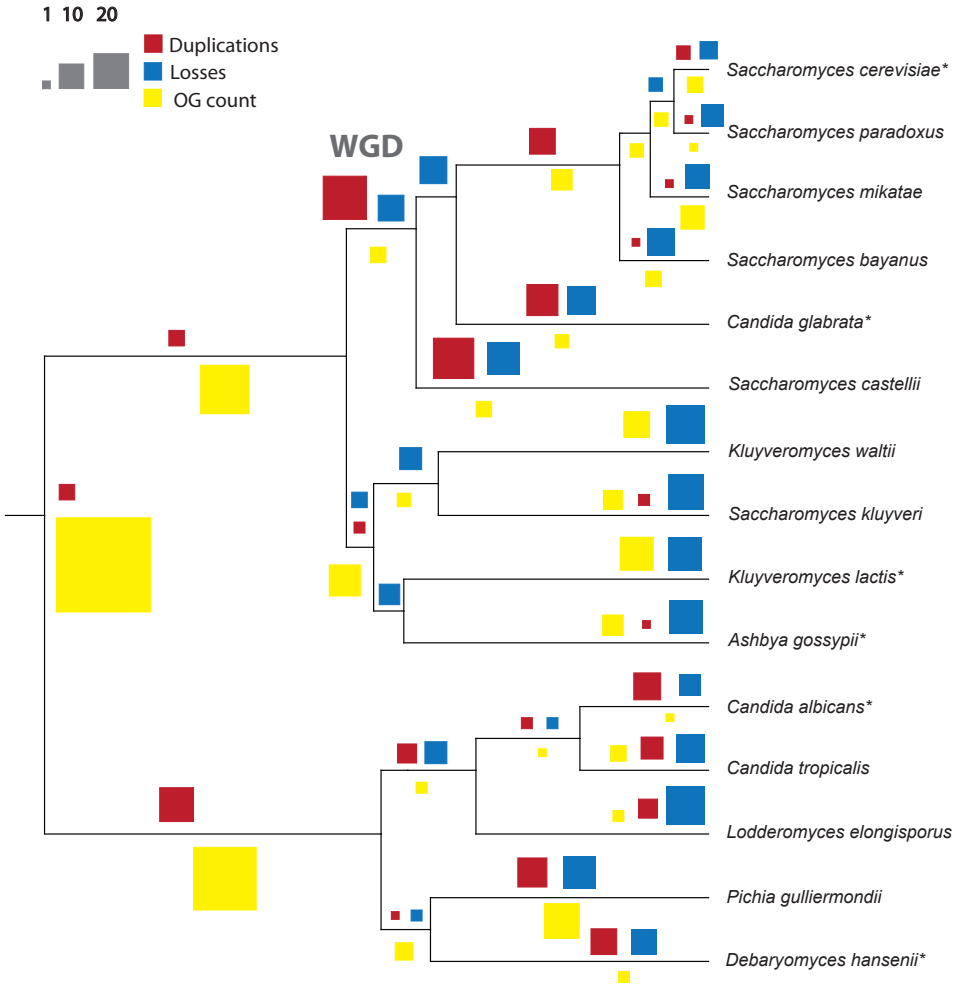


Figure 4.11: Number of duplications (red) and losses (blue) inferred for 271 orthogroups using SPIMAP. Yellow boxes indicate the number of orthogroups rooted at each of the respective branches. The whole genome duplication branch is marked "WGD". The area of the boxes is scaled according to the number of inferred events using an arbitrary unit (see in-figure legend). Asterisks denote the species for which the genome sequence was finished at the time when the dataset was collected.

to be due to gene tree reconstruction artefacts as indicated by the large number of accompanying losses inferred on the same branches. It was previously found that both *S. castellii* and *C. glabrata* have retained more WGD paralogs than the *sensu stricto* species and that those are enriched for transcriptional regulators (Chapter 3.6.3; Byrne & Wolfe, 2007), arguing for the true placement of these duplications on the WGD branch. Indeed, inspection of the gene trees of affected orthogroups revealed frequent cases where both *S. castellii* and/or *C. glabrata* duplicates were reconstructed as species-specific duplications on long branches where the generating duplications were inferred shortly after the WGD, supporting the idea that those might be WGD paralogs instead. Interestingly, this pattern was not only confined to cases where one of the duplicates was subsequently lost in the *sensu stricto* species but also occurred when both copies were retained in all descendant lineages, explaining the increased number of duplications inferred on the *sensu stricto* branch (Fig. 4.11). Examination of the large number of species-specific duplications that were detected in the comparative analyses discussed above (see Fig. 4.7) revealed that a large number of those were inferred on the *C. glabrata* and *S. castelii* branch, suggesting the possibility of high incidences of gene conversion in these lineages.

Only five duplications were inferred within the pre-WGD species (Fig. 4.11) indicating that this clade has mainly been dominated by gene loss in comparison to the other clades. Gene loss is somewhat more difficult to interpret within this context and will be discussed in more detail below. It is noteworthy, however, that all four pre-WGD species contain a large number of single-species orphan orthogroups containing single TFs that could not be attributed to any homologs by either SYNERGY or the YGOB.

The CTG clade in turn experienced relatively large amounts of gene gain, with duplications inferred on every branch. This was in line with the overall increase in numbers of TFs in those genomes as well as the overall increase in the number of protein coding genes they encode (see Chapter 2; Table 2.1). Furthermore, I found just under 20% of the orthogroups to be rooted at the base of the CTG clade and thus represent additional copies not present in other clades, further underlining this increase in size. Again, those could not be attributed to homologs in the *Saccharomycetacea* and it is possible that a number of CTG orphans are

indeed homologous to other non-root node orthogroups (e.g. those rooted at the *Saccharomycetaceae* or the pre- and post-WGD clades), especially seeing that remapping using the gene order browsers was restricted by the requirement of the presence of a *Saccharomyces cerevisiae* ortholog. This is unlikely to explain all of those orphan orthogroups however.

Inferred numbers of losses will be affected by the quality of the genome annotation of each species and accuracy of the placement of duplications, making interpretation of those more challenging. The majority of losses were inferred at tips of the tree, indicating that both issues might be affecting the estimates. When considering only estimates from finished genome sequences ($*$, Fig. 4.11), assuming that those annotations will be more complete, the numbers of losses were indeed smaller than in the surrounding species with lower quality annotations (with the exception of *C. glabrata*; the reasons for an inflated amount of losses here are explained above). The numbers of losses inferred on the deeper branches in turn are likely to be an underestimate in light of the large number of orphan orthogroups rooted at (e.g.) the base of the CTG clade or the *Saccharomycetacea*, some of which are likely to have been lost in other clades rather than gained at their root. It is not possible however to determine whether those orthogroups represent lineage-specific duplications or lineage-specific losses without further phylogenetic analysis to resolve the relationships between different orthogroups in the same DBD family. Initial analyses (not shown) suggested that this would be difficult to achieve, given the large amounts of sequence divergence and relatively short alignable regions in the species considered here, often resulting in poorly-supported family-wide phylogenetic trees. Further analyses were thus not attempted.

Despite these difficulties in interpretation, it was clear that the pre-WGD species have experienced increased amounts of TF loss compared to the other clades studied here, often in a species-specific manner. Interesting in this context is also the large number of single-species orphan orthogroups (yellow boxes at the tips of the tree in Figure 4.11) suggesting either a regime of independent losses (often retaining only a single copy in one of the species), ancient species-specific duplication, or rapid sequence divergence of ancient homologs in individual species to the extent where it is not possible to determine orthology

any longer. The fact that these orphans were frequently found within conserved syntenic regions according to YGOB argues for the earlier two possibilities to be more likely.

Overall, the three major clades studied here showed three very different patterns of gene family evolution. The post-WGD species experienced large-scale duplication of parts of their regulatory repertoire followed by a stable period with few duplications and a slightly larger number of losses. This suggested a burst of regulatory novelty followed by finer-scale specialisation and adaptation. The pre-WGD species were mainly dominated by TF loss which possibly indicates a process of fine-tuning of their existing repertoires in the adaptation to their individual niches. All species studied in the pre-WGD clade are distantly related to each other, with the speciations giving rise to the extant members being placed deeply in the tree (see Chapter 2, Figure 2.9). As a consequence, the regulatory networks in those species will have evolved independently for extended amounts of time, which in turn could be an explanation for the large amounts of species-specific losses and duplications and high levels of sequence divergence giving rise to the numerous orphans orthogroups observed in this clade. Again, it needs to be emphasised that the patterns observed here are difficult to distinguish from the ones that would arise from missing data, although the patterns found in finished genomes here are no different from those of lower quality. Finally, the CTG clade appears to have been under increased amounts of turnover, displaying both elevated rates of duplication and loss throughout. Seeing that a large number of species within this clade (see Chapter 1.5.2.2) are pathogenic to humans and other hosts, there is a possibility that this reflects active adaptive processes to fluctuations in environmental conditions such as a co-evolving immune system and it will be very interesting to explore the functional relevance of this increased plasticity in more detail (see below and Chapters 3.6.4 and 6.5).

The patterns observed here mirror genome-wide all-gene estimates of duplications and losses for these lineages (Wapinski *et al.*, 2007b), suggesting a tight coordination between increases in the numbers of transcriptional regulators and protein-coding content of the genome. This level of coordination argues for a facilitating role of increased numbers of regulatory proteins for the retention of gene duplicates. Whether this is through an increase in regulatory complexity, TF

dosage-dependent or both is unclear at this moment but would be an interesting question for further investigation.

4.4.2 Family-wise and Clade-wise Enrichment for Events

The contribution of different DBD families to TF repertoires in most organisms is highly asymmetrical with a small number of DBD families contributing the majority of TFs in a given genome (see Chapter 3.4). A WGD event can be the result of a hybridisation between different species or due to a spontaneous doubling of chromosomes, e.g. after a failed round of meiosis (reviewed in Dujon *et al.*, 2004) and we would not expect a strong bias towards the retention of particular families *per se*. SSD events in turn can be the result of a number of different mechanisms, some of which are likely to result in amplification of already large families. Duplicate genes can be the result of retrotransposition of mature mRNAs trough the machinery encoded by retrotransposons, non-homologous recombination, recombination between paralogs or errors during DNA replication (reviewed in Hurles, 2004). Especially in the latter two cases, large families will have greater tendencies to fluctuate in size due to the increased likelihood of the formation of unequal crossovers (e.g. Ahlroth *et al.*, 2001) or differential re-annealing after strand slippage during replication. To test whether any of the DBD families found in the *Saccharomycotina* experienced significantly increased rates of turnover relative to the other families, I calculated expected rates of duplication and loss (under the assumption of equal rates) and contrasted those with observed rates for each family across the entire tree as well as in a clade-specific manner. Significance testing was performed using permutation tests and results were confirmed using a probabilistic model of birth and death (BD) evolution as implemented in CAFE (Hahn, 2007).

4.4.2.1 Rate of Gain and Loss, Difference Statistic and Permutation Testing

The expected rate of duplication across the tree was defined as:

$$exp_T = \frac{dup_T}{anc_T} \tag{4.1}$$

where dup_T is the total number of duplications inferred across all families and anc_T is the number of ancestral orthogroups. Note that here all orthogroups were counted as ancestral, even if rooted at shallower nodes in the species tree and a duplication was inferred at the root node level of orphan orthogroups. This presents a simple shortcut to mitigate the effects of not knowing whether such orthogroups arose through duplication or losses, whilst acknowledging the increased copy number of TFs in that family. This slightly biases statistics towards duplications, as an alternative explanation of the observed orphan orthogroup would have been the loss of members in other clades. I considered this acceptable, however, as a single duplication is often the more parsimonious explanation, especially for orthogroups rooted near the tips of the species tree. In the following sections, I will distinguish between such assumed duplications and *bona fide* duplication events.

The expected number of duplications per family was defined as:

$$exp_{\text{fam}} = exp_T \times anc_{\text{fam}} \tag{4.2}$$

where exp_T the expected rate of duplication across the tree as defined above and anc_{fam} the number of ancestral orthogroups for that family.

Finally the difference statistic was calculated as follows:

$$S_{\text{dup}} = obs_{\text{fam}} - exp_{\text{fam}} \tag{4.3}$$

where obs_{fam} is the number of inferred duplications for a particular family and exp_{fam} the number of expected duplications for that family as defined above.

Similarly, I calculated S_{loss} for losses. A third statistic, describing the overall plasticity of a family, was defined as $|S_{\text{dup}}| + |S_{\text{loss}}|$.

In addition to looking for deviations from average rates of duplication and loss across the whole tree, clade-specific patterns of DBD family evolution were calculated as above but limited to a particular subtree for the post-WGD, pre-WGD, *Saccharomycetacea* and CTG clades by dropping orthogroup members from other clades not under consideration. The ancestral number of orthogroups here was taken to be the number of orthogroups that contained descendants in the clade analysed.

In order to determine the statistical significance of the differences between observed and expected numbers of events for each family, I implemented a permutation test by randomly re-labelling orthogroups with DBDs whilst keeping the overall number of orthogroups per DBD constant. P-values were calculated from 1000 samples and were adjusted to have a false discovery rate (FDR; Benjamini & Hochberg, 1995) of 0.05 to correct for multiple comparisons.

4.4.2.2 Results

Table 4.3 shows the results obtained from the analyses outlined above. None of the DBD families showed significantly accelerated rates of gene gain, loss or overall plasticity when considered across the entire tree. Analysis of clade-specific patterns however revealed several significant changes in gene family dynamics in each of the subtrees. The SRF-TF-like family of MADS box TFs was found to have experienced increased rates of turnover in the post-WGD species, showing significantly accelerated rates of duplication, loss and overall plasticity. Given that both the rate of gain and loss were found to be highly significant however, this result appeared dubious in light of the known biases of gene tree reconstruction where incorrectly reconstructed trees lead to inference of excessive numbers of duplications and losses. Indeed, inspection of the alignments and reconstructed trees of the two orthogroups annotated with an SRF-TF DBD suggested that this was the case. Although both orthogroups were duplicated during the WGD, representing true signal, excessive amounts of duplications and losses were inferred due to one of the biases discussed above where *S. castellii* and *C. glabrata* WGD paralogs tended to cluster together instead of with their syntenic orthologs. Manual inspection of other significantly accelerated DBD families (see below) confirmed that this was the only enrichment affected by gene tree reconstruction artefacts, however.

TF gain in the pre-WGD species was found to be dominated by the Zn_clus ($Zn(II)_2Cys_6$) as well as the Zn_clus:Fungal_trans domain-containing TFs. While very few *bona fide* duplications were inferred in this clade, all four species contained a large-number of orphan orthogroups, counted as assumed duplications, which were significantly enriched for those two families. This suggested that,

independent of whether these orphan orthogroups were the results of ancient duplications, losses or high levels of sequence divergence, these families have experienced different evolutionary pressures compared to the other DBD families in this clade.

When considering the *Saccharomycetacea* as a whole, I found accelerated rates of duplication of the C_2H_2 tandem zinc finger TFs. Again, this enrichment was most likely due to the large number of orthogroups that are specific to the *Saccharomycetacea*. Over 60% of orthogroups containing this domain architecture are only found in this clade (compared to 20% tree-wide and 20% CTG-specific orthogroups). Conceivably, this enrichment could be somewhat inflated given that the C_2H_2 tandem zinc finger is among the shortest domain motifs in this dataset and the sequences encoding it are thus more likely to have been misclustered by SYNERGY and subsequently fragmented (see Chapter 4.1). Nevertheless, even if all 20% of CTG-specific orthogroups were the result of artificial fragmentation and could be attributed to their orthologs in the *Saccharomycetacea*, this would still leave an excess of clade-specific C_2H_2:C_2H_2-containing orthogroups. Furthermore, I detected significant increases in overall plasticity of the SRF-TF family but, as before, this was most likely due to gene tree reconstruction artefacts.

Finally, investigation of patterns in the CTG clade revealed significantly higher rates of gene gain in the Zn_clus, Zn_clus:Fungal_trans and zf-C3HC4:Zn_clus:Fungal_trans families. These families were enriched for both *bona fide* inferred duplications and clade-specific orthogroups indicating clear propensity for expansion in these families within the CTG clade. This is also reflected in the overall composition of TF repertoires (Chapter 3, Fig. 3.6), showing that a large proportion of the TFs gained in the CTG species is accommodated within the Zn_clus and Zn_clus:Fungal_trans families.

The zf-C3HC4:Zn_clus:Fungal_trans domain architecture was exclusively found within the CTG clade (see Chapter 3, Fig. 3.6) and has been amplified on the branch leading to the CTG species as well as in *Candida albicans* and *Debaryomyces hansenii*.

Besides acceleration in the rates of gain and loss in different DBD families, I also tested for families showing significantly decreased rates of turnover (the lower tail of the probability distributions was calculated using permutation testing).

	DBD family	P-value(adj)
All		
Duplications	—	—
Losses	—	—
Plasticity	—	—
WGD		
Duplications	SRF-TF	< 0.001
Losses	SRF-TF	< 0.001
Plasticity	SRF-TF	< 0.001
Pre-WGD		
Duplications	Zn_clus	0.043
	Zn_clus:Fungal_trans	0.039
Losses	—	—
Plasticity	—	—
Saccharomycetacea		
Duplications	zf-C2H2:zf-C2H2	< 0.001
Losses	—	—
Plasticity	SRF-TF	< 0.001
CTG		
Duplications	Zn_clus	< 0.001
	Zn_clus:Fungal_trans	< 0.001
	zf-C3HC4:Zn_clus:Fungal_trans	0.043
Losses	—	—
Plasticity	—	—

Table 4.3: Accelerated rates of duplication, losses and overall plasticity in different DBD families

Due to the very large number of significant families in some clades, these results are not shown here but discussed below. Again, no family showed significantly decelerated rates of either duplications or losses when considered across the entire tree. When investigating clade-specific results, I found a large number of families showing significantly decreased rates of duplication in the WGD and pre-WGD clades (31 and 47, respectively) whereas only two families showed significantly decreased rates of duplication in the CTG clade. In light of the observed duplication rates in each of these clades (see above) this seemed counter-intuitive at first, i.e. one would expect a large number of families to show relatively decelerated rates of duplication in the CTG clade where the majority of inferred duplications was accommodated within a few families. The results can be explained by taking into consideration both the spread of duplications across different families and overall duplication rate: in the WGD clade, duplication rate is high and duplications are spread across many different DBD families, leading to high per-family expected rates when sampling across the TF repertoire during permutation testing and consequently a large number of significantly decelerated families. Similarly, in the pre-WGD species, the duplication rate is low and duplications are enriched in few DBD families and although per-family expected rates are low, only few DBD families showed rates surpassing this. Finally, the duplication rate in the CTG clade is very high and duplications are significantly enriched in few families. Duplications were inferred in a number of families however, albeit at lower numbers. I found no significantly decelerated rates rates of loss in either of the clades.

4.4.2.3 CAFE

In order to obtain an alternative validation of the above observations using a probabilistic model of gene family evolution, I used CAFE v2.1 (De Bie *et al.*, 2006) to identify families that were undergoing large changes in size. CAFE estimates the birth and death (BD) rate parameter (λ) across all families based on the number of observed gene copies in all the extant species. The BD model is subsequently used as a null hypothesis of random changes in gene family size to test for families that violate this model (Hahn *et al.*, 2005). Note that λ

represents a single parameter for both gene gain and loss, a condition which is essentially violated by the WGD. It is possible however to specify different values of λ for groups of branches ("λ-models"), a feature which should make it possible to overcome this effect to some extent.

CAFE was run using several different λ-models on the entire tree as well as on the WGD and non-WGD subtrees only. The results are summarised in Figure 4.12. Model testing was performed using likelihood ratio tests (LRTs) for nested models and the AIC criterion (Burnham & Anderson, 2004) otherwise (see Chapter 2.2.2 for definitions). Both LRT and AIC support a two-rate model (Fig. 4.12; yellow row ("Full") and tree) with different λ for WGD and non-WGD species. A three-rate model, specifying different values for λ for the WGD, pre-WGD and CTG clades was also tested and not rejected on the grounds of the AIC (ΔAIC <2), but the increase in likelihood was not significant according to the LRT. Despite the gain and loss dynamics in the pre-WGD and CTG clades being very different (see Fig. 4.11), this probably reflects the coupling of duplication and loss rate in a single parameter suggesting overall similar levels of plasticity despite the differences in directionality.

Besides the full model, I also estimated λ-models on the WGD and non-WGD subtrees in isolation to assess whether the WGD had an impact on parameter estimation and thus rendered the previous analysis invalid. Model testing confirmed the use of a single λ for each of those clades (Fig. 4.12; yellow rows ("no WGD" and "WGD only"). Although here in both cases the more complex model had a lower AIC value, the ΔAIC was not significant, supporting the use of the simpler model. LRTs also did not support multi-rate models. The λ values estimated for the single-rate models in the WGD and non-WGD species were very close to the values estimated for each subtree in the full model, supporting the validity of these analyses.

CAFE infers the numbers of gene copies for all families at each internal node across the species tree based on the number of observed copies in the extant species. Given the best-fit model of BD evolution and λ estimates it then calculates the probability of observing the inferred change in copy number given the null model for each branch, yielding an overall P-value for the test of the given family evolving according to the BD model as well as branch-wise probabilities for

Model	λ_1	λ_2	λ_3	ΔlnL	P(LRT)	ΔAIC
Full						
(((((1,1),1),1),1),1),1),((1,1),(1,1))),1),((((1,1),1),1),(1,1)))));	0.00238			11.03	—	19.21
(((((2,2)2,2)2,2)2,2)2,((1,1)1,(1,1)1)1)1,(((1,1),1),(1,1)1)1)1);	0.00193	0.00368			0.000	—
((((((2,2)2,2)2,2)2,2)2)1,((1,1),(1,1))1)1,(((1,1),1),(1,1))1)1);	0.00203	0.00351		4.30	—	7.75
((((((2,2)2,2)2,2)2,2)2,((2,2)2,(2,2)2)2,(((1,1),1),1),(1,1))1)1);	0.00197	0.00267		8.75	—	16.65
((((((2,2)2,2)2,2)2,2)2,((3,3)3,(3,3)3)3,(((1,1),1),(1,1))1)1);	0.00210	0.00369	0.00172	—	0.358	1.15
no VCD						
(((1,1),1),(1,1))1,(((1,1),1),(1,1))1);	0.00187			1.13	—	0.10
(((2,2)2,(2,2)2)2,(((1,1),1),1),(1,1))1);	0.00214	0.00156		0.08	0.147	—
(((2,2)2,(2,2)2)2,(((3,3)3,3)3,(1,1))1)1);	0.00225	0.00154	0.00204	—	0.323	1.85
VCD only						
((((1,1),1),1),1),1);	0.00317			1.48	—	0.97
(((((2,2)2,1),1),1),1),1);	0.00289	0.00484		—	0.085	—
(((((2,2)2,2)2,2)2,1),1),1);	0.00282	0.00348		1.12	0.391	2.23
(((((2,2)2,2)2,2)2,2)2,1);	0.00236	0.00343		0.80	0.243	1.61

Figure 4.12: Species tree showing the best-fit λ model for CAFE analysis and model testing results. The number profile in the Model column indicates which branches in the tree share which λ parameters. The best-fit λ-model for the full tree is shown on the species tree. Coloured branches indicate different λ's. The λ_1, λ_2 and λ_3 columns show the inferred values using each of the respective models. Note that these are not comparable to λ values estimated using SPIMAP as λ here is a combination of λ and μ. The remaining columns show the differences in log likelihoods and *AIC* for each model compared to the model with the highest likelihood and *AIC* value, respectively. P-values are derived from LRTs of nested λ-models.

Family	P-value	# Sig. Branches	Clades
zf-C2H2:zf-C2H2	0.001	3	WGD(1); pre-WGD(2)
Zn_clus:Fungal_trans	0.000	5	CTG(4); pre-WGD(1)
Zn_clus	0.000	3	CTG(2); pre-WGD(1)

Table 4.4: CAFE results: Families showing an overall P-value < 0.01. Significant branches indicates the number of significant branches found in each families and "Clades" indicates how many of these were found in each subtree. Significant branches were detected for both increase and decrease in numbers of TFs in each family, although these do not necessarily correspond to increases or decreases in rate as the test only surveys changes in the number of inferred family members at each node. A much higher number of TFs on one branch will lead to a higher number of inferred copies at the ancestral node and might indicate losses on the low-copy number branch, even if the true difference was generated by large expansion on the high-copy branch.

model conformance. Table 4.4 shows a summary of the families that were found to evolve differently from the BD null model with $P < 0.01$. Full visualisation of CAFE output for these families can be found in the Appendix, Figures A.3 - A.5.

Encouragingly, the CAFE results very closely mirrored the results obtained from my permutation tests reinforcing the notion that those were not generated by gene tree reconstruction artefacts or excessive fragmentation of misclustered orthogroups and indeed represent genuine signal. Virtually all families found significant in the clade-specific rate analysis were equally detected to have experienced significant changes in rates of turnover by the BD model in the corresponding clades. The one exception was the zf-C3HC4:Zn_clus:Fungal_trans family which, being specific to the CTG clade, did not yield a low enough P-value for deviation from the BD null model across the whole tree .

One observation that was not detected in the permutation tests was the very high retention rate of zf-C2H2:zf-C2H2 WGD paralogs in *Saccharomyces castellii*, which contains 29 TFs encoding this DBD compared to 23 in *Candida glabrata* and 18 or 19 in the remaining post-WGD species (see Appendix, Fig. A.3). This higher tendency for retention of WGD duplicates in *S. castellii* and *C. glabrata* has become obvious from overall examination of the composition of the TF reper-

toires in those species (Chapter 3; Fig.3.6) and does not appear to be family-specific. It would be interesting however to explore whether there are any biases towards retention of particular functional classes of TFs in those species in more detail.

4.5 Two Modes of Regulatory Network Growth

Evolutionary dynamics within the three clades studied here were found to be very different. While regulatory network growth in the post-genome duplication species was mainly achieved through the retention of WGD paralogous TFs, network growth in the CTG clade proceeded through series of small-scale duplications (SSD). These differences were also reflected when considering family-specific evolutionary dynamics. Whereas I found retention of WGD TFs to be non-family-specific, the SSD duplicates in the CTG clade were highly enriched for TFs belonging to the two largest DBD families, containing Zn_clus and Zn_clus:Fungal_trans architectures and contributing to an ongoing lineage-specific amplification of those. Indeed, recent duplicates in *C. albicans* and *D. hansenii* were often located close by on the same chromosome, implicating non-homologous recombination and slippage during DNA replication (see Chapter 4.4.2) as important mechanisms generating the lineage-specific amplification in the CTG clade.

This suggested the existence of two distinct modes of regulatory network growth that are potentially very different in their functional properties, topological features within the regulatory network and the resulting adaptive potential. TFs do not act in isolation but are part of higher order network structures where they cooperate with other TFs to regulate downstream target genes (see Chapter 1.2.3 for review). These networks have been shown to be hierarchically organised (e.g. Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009; Yu & Gerstein, 2006) and display a scale-free structure (e.g. Guelzim *et al.*, 2002), where a small number of TFs, denoted regulatory hubs, regulate a very large number of target genes while most TFs regulate a small number of target genes (see Chapter 1; Fig. 1.3). TFs in different hierarchical layers display distinct functional and topological properties, e.g. TFs in the higher levels of the hierarchy tend to be enriched for regulatory hubs (especially in the core layer which is highly interconnected) and be involved

in a larger number of biological processes than the TFs found at the bottom of the hierarchy, which are often stand-alone and specialist in their roles (Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009).

It is known that the properties of genes duplicated in small-scale duplications differ substantially from duplicates generated through WGD events (reviewed in Conant & Wolfe, 2008). While small-scale duplicates are often found in the less densely-populated regions of protein-protein interaction (PPI) networks (Li *et al.*, 2006) and generally exhibit smaller knockout fitness effects (He & Zhang, 2006), much the opposite is true for duplicates generated by WGDs (Maere *et al.*, 2005; Van de Peer *et al.*, 2009; Wapinski *et al.*, 2007b). This has been explained by the "dosage-balance hypothesis" (Edger & Pires, 2009; Papp *et al.*, 2003). SSDs of genes with large pleiotropic constraints that have many interaction partners or act in a concentration-dependent manner are likely to have detrimental effects due to the disturbance of the stochiometry of the modules or complexes they belong to. Genome-scale duplication in contrast allows for the simultaneous duplication of an entire module, thus relieving such pleiotropic constraints. Indeed, WGD paralogs that have been retained in the *Saccharomyces sensu stricto* species are often part of the same protein complex and perform central roles in the cell (Wapinski *et al.*, 2007b). Based on these findings one could expect similar tendencies for the retention of WGD duplicate TFs, i.e. increased retention of TFs at the core of the regulatory hierarchy and within the same pathways. Recent studies have started addressing some of those questions and have found that the post-WGD network in *S. cerevisiae* forms a highly-connected network layer with frequent cross-talk between duplicated and ancient pathways and that the subgraphs of WGD duplicate regulatory motifs are generally more complex than motifs created through SSD (Fusco *et al.*, 2010). Also, the ancestors of WGD duplicate TFs were found to be more influential, as determined by the extent of network fragmentation into non-connected subcomponents when computationally removed, than the ancestors of other TFs in inferred ancestral regulatory networks (Conant, 2010). These results hint that indeed, TFs retained after WGD were highly connected and part of core regulatory pathways and below I present a more detailed analysis based on my results.

In order to determine whether and how WGD duplicates differed from TFs duplicated trough SSD, I examined their placement within the *S. cerevisiae* regulatory network with respect to the number of regulatory interactions and position in the hierarchical organisation of the network. For these comparisons, I used the network compiled by Jothi *et al.* (2009), which is a combination of a series of biochemical and ChIP-chip experiments (see Jothi *et al.*, 2009 and references therein). This network contains 13,385 regulatory interactions between 158 TFs and 4369 target genes. Overall, 143 of the 158 TFs were shared between my dataset and the "Jothi2009" dataset. For corroboration of results I also used a regulatory network compiled from YEASTRACT (YT; Teixeira *et al.*, 2006) which is a literature-based, curated repository of regulatory associations. In my version of the YT network, I chose to only include regulatory interactions for which direct evidence, i.e. through binding of a TF, was available. This YT dataset contains 21,898 interactions for 184 TFs and 5833 target genes, of which 149 TFs were shared with my dataset. There are significant correlations between the number of regulatory interactions per TF in the Jothi2009 and YT networks (outdegree: $R^2 = 0.467$; $P < 2.2 \times 10^{-16}$; indegree: $R^2 = 0.666$; $P < 2.2 \times 10^{-16}$). Approximately 87% of the interactions in the Jothi2009 network are contained in the YT network. As such the Jothi2009 network is largely a subset of the YT network, a fact also reflected by the overall bigger number of interactions per TF in the YT network.

It is unclear which of these two datasets is a more biologically valid representation of the *S. cerevisiae* regulatory network. The Jothi2009 network is likely be biased towards the few experimental conditions surveyed in the original studies that were used to generate the dataset and might be missing a large number of condition-specific regulatory interactions. While those are more likely to be included in the YT data, the YT dataset likely suffers from study bias towards a small number of well-studied TFs. Indeed, the two TFs showing by far the highest outdegree in the YT network are two very well-characterised regulators (*STE12* and *RAP1*). For those reasons, I chose to use both networks for cross-validation of results whenever possible.

I used my results on the evolutionary histories of *S. cerevisiae* TFs to classify each as a WGD TF if it had arisen from a duplication event inferred on the WGD

branch and been retained in the *sensu stricto* species. Because of the inference biases around the *S. castellii* and *C. glabrata* nodes (see above), I furthermore included duplicates that arose on either the lineage leading to the *sensu stricto* species plus *C. glabrata* or the *sensu stricto* species only if their WGD status was supported by syntenic data in the YGOB (Byrne & Wolfe, 2005). Overall, 63 *S. cerevisiae* TFs were classified as WGD duplicates. Conversely, all other duplications since the root of the species tree were classified as non-WGD duplicates. A full list of WGD TFs is shown in Appendix, Table A.2.

4.5.1 WGD Paralogs Are Enriched for Highly Connected TFs

TFs with many regulatory interactions form important parts of the information processing structure of regulatory networks. Conceivably, the more regulatory interactions a TF has, the greater its potential for integration of signals (indegree; the number of TFs it is regulated by) or for their dissemination (outdegree; the number of target genes it regulates). Accordingly, one would expect duplication of such highly connected TFs to have large pleiotropic effects that might affect the accuracy and efficiency of signal processing. In order to examine whether TFs retained after WGD show similar tendencies to other WGD-retained genes, that is high pleiotropic effects with respect to the organisation of the regulatory network, I compared the distribution of indegrees for WDG TFs to the distribution of indegrees of non-WGD TFs as well as the distribution of outdegrees of WGD TFs to the distribution of outdegrees of non-WGD TFs using both the Jothi2009 and YT networks.

Figures 4.13 and 4.14 show the distributions of outdegrees and indegrees of non-WGD (grey) and WGD TFs (red) respectively. Comparison with the Jothi2009 network showed a clear trend for enrichment of high outdegree in the WGD TFs (Figure 4.13). This difference was statistically significant (one-tailed Wilcoxon test: $W = 3269$; $P < 0.001$). The trend was less clear in the YT network and did not yield a significant difference (one-tailed Wilcoxon test: $W = 2773$; $P = 0.231$), although there was a clear excess of WGD duplicates among the TFs regulating between 300 and 900 target genes. Interestingly, manual inspection of

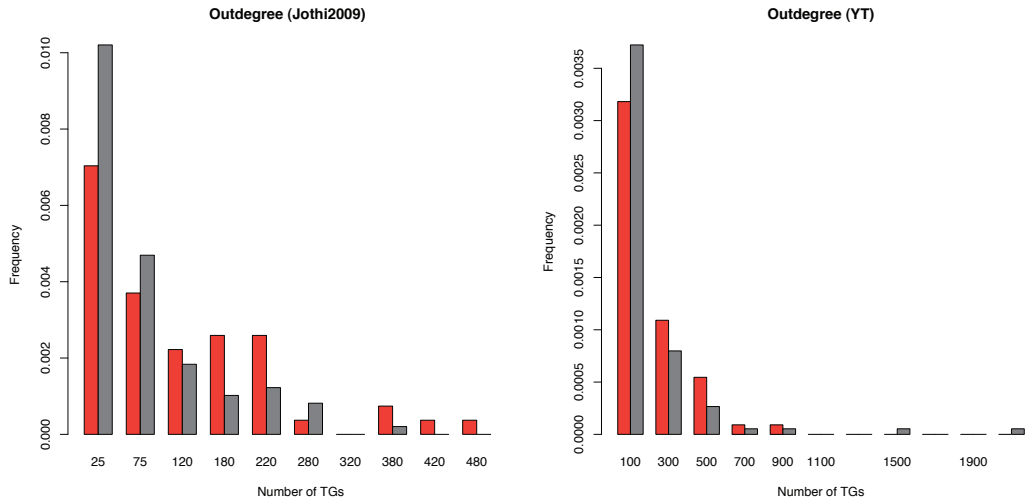


Figure 4.13: Distribution of the outdegrees of WGD TF duplicates (red) compared to non-WGD TFs (grey) in the Jothi2009 and YT regulatory networks.

the four TFs with the highest outdegree in the YT network revealed that three of those were indeed retained after WGD but only in either *S. castellii*, *C. glabrata* or both and then subsequently lost on the *sensu stricto* lineage and therefore not included in this analysis. It would be interesting to repeat this test when those results are included. The distributions of indegrees (Fig. 4.14) showed a clear enrichment for high indegree in WGD TFs compared to the distributions of indegree among non-WGD TFs and was statistically significant in both networks (one-tailed Wilcoxon test: W = 3485; P = 0.039 [Jothi2009] and W = 5530; P = 0.004 [YT]).

In order to investigate the discrepancies in enrichment results for high outdegree among WGD TFs between the Jothi2009 and YT networks, I examined the relationship between indegree and outdegree of TFs in each network (Fig. 4.15). As mentioned above, the YT network contains more regulatory interactions than the Jothi2009 network overall and includes the majority of the interactions present in the Jothi2009 dataset. Examination of corresponding outdegrees of WGD and non-WGD TFs highlighted the differences in outdegree distribution between the two networks. While at the lower end, the association between outdegrees was almost linear, there were large differences between the Jothi2009 and YT net-

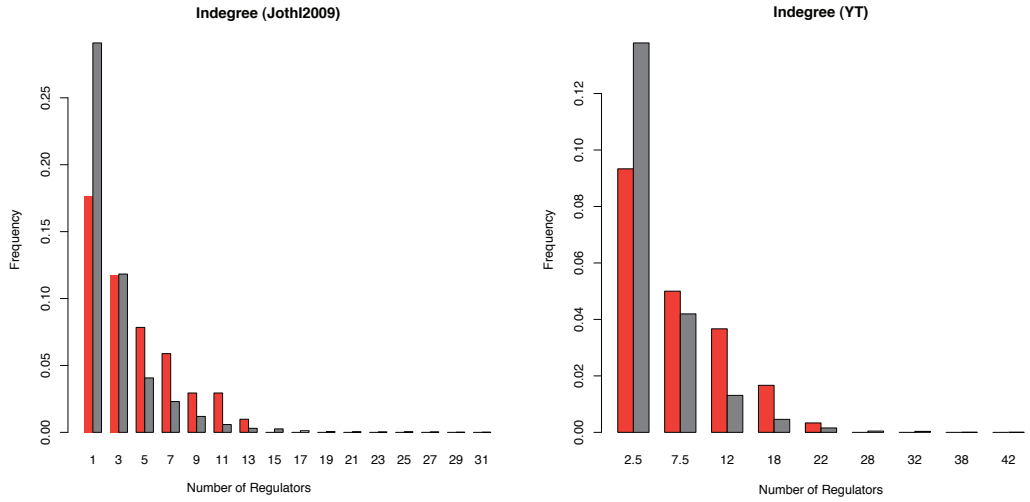


Figure 4.14: Distribution of the indegrees of WGD TF duplicates (red) compared to non-WGD TFs (grey) in the Jothi2009 and YT regulatory networks.

works in ranking order and magnitude of the high outdegree TFs. Those are likely responsible for the significant enrichments seen among WGD TFs in the Jothi2009 but not the YT network. In line with the stronger correlation between indegrees of all genes in the Jothi2009 and YT networks (see above), TF indegrees followed a more linear relationship (Fig. 4.15), Although the absolute numbers of regulators per TF differed, WGD TFs were distributed similarly compared to non-WGD TFs in both networks resulting in significant enrichments for large indegree in both the Jothi2009 and the YT network.

Independent of the choice of network used to investigate the properties of WGD TFs, I found clear trends for the retention of highly connected TFs following whole-genome duplication. Although these enrichments were not always significant depending on which network was considered, the degree distribution of these TFs clearly supported the notion that the WGD allowed for the retention of highly-connected network components, predicted to have large pleiotropic effects. To examine this further, I also considered the incidence of regulatory hubs, defined as the top 20% of highly connected TFs (Jothi *et al.*, 2009), in the Jothi2009 network. The results of this are shown in Table 4.5. Despite a significant enrichment for high-outdegree TFs however and a strong trend to-

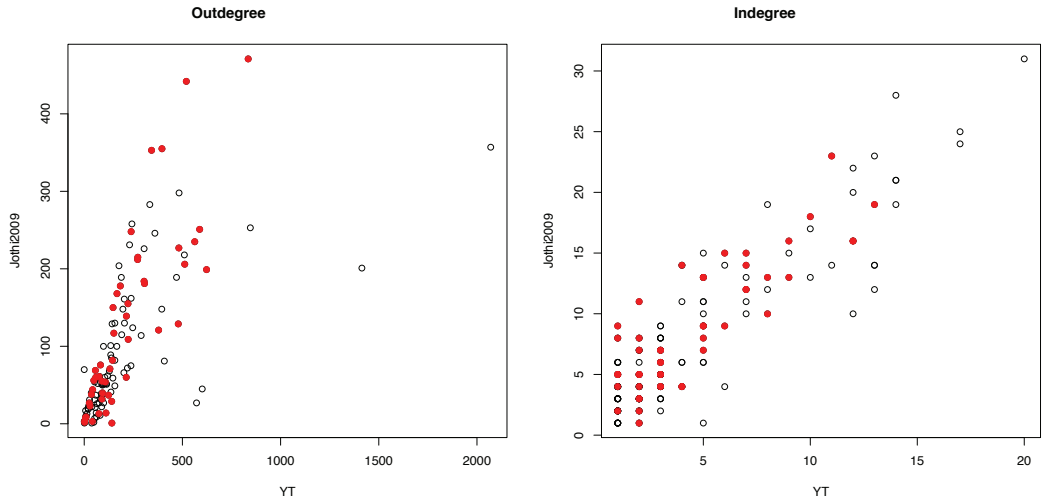


Figure 4.15: Distribution of outdegrees of WGD TF duplicates (red) compared to the overall distribution of outdegrees (black) in the Jothi2009 and YT regulatory networks (left plot). Distribution of indegrees of WGD TF duplicates (red) compared to the indegree distribution of all TFs (black) in the Jothi2009 and YT regulatory networks (right plot)

	Hubs	Non-hubs	N/A	Total
All	32	102	93	227
WGD	18	33	12	63
Expected	12.56	38.44	—	—
Expected (N/A)	8.88	28.31	25.81	—

Table 4.5: Number of regulatory hubs among WGD TFs. N/A refers to TFs in my dataset that were not included in the Jothi2009 regulatory network.

wards a bias for regulatory hubs, I found no statistically significant enrichment for regulatory hubs among the WGD TFs (χ^2 test: P = 0.056; df = 1).

Interestingly, when including the proportion of TFs that were not included in the regulatory network into the statistic, I found a strong underrepresentation of not-included TFs ("N/A"; Tab. 4.5; row 4) among the WGD TFs (χ^2 test: P < 0.001; df = 2) . In other words, WGD paralogs were more likely to have been included in experimental studies than non-WGD TFs. This can be interpreted as the WGD paralogs being of greater importance to the functioning and integrity of regulatory modules and hence having been subject to greater experimental characterisation. This idea is supported by the studies of ancestral reconstructed networks by Conant (2010), who found ancestral WGD TFs to be more influential (see above). Alternatively, this could be a side-effect of the balanced duplication of many, often well-defined DBD families (see Chapter 3, Fig. 3.6), many of which are more easily detected than the large zinc finger families and thus more likely to have been identified in earlier studies of TFs in *S. cerevisiae* (see Chapter 3.2.2.1).

The yeast regulatory network can be seen as being organised in a hierarchical manner, comprising a large feed-forward structure containing a top layer of TFs that regulate TFs in the core and bottom layers, a very highly connected core layer regulating downstream TFs in the bottom layer and a bottom layer that represents the end of the regulatory cascades, regulating non-TF target genes (e.g. Jothi *et al.*, 2009; see Chapter 1, Fig. 1.3C). Different layers have different properties. TFs in the top and core layers are generally associated with more

	Top	Core	Bottom	# N/A	Total
All	18	61	55	93	227
WGD	7	27	17	12	63
Expected	6.85	23.22	20.93	—	—
Expected (N/A)	5.00	16.93	15.26	25.81	—

Table 4.6: Distribution of WGD TFs among the hierarchical layers of the Jothi2009 network. N/A refers to TFs in my dataset that were not included in the Jothi2009 regulatory network.

biological processes than the bottom layer TFs as well as being more conserved (on the basis of presence or absence of homologs in 15 other fungal genomes). The core layer contains 27 of the 32 TFs classified as hubs and is involved in regulation of 87% of target genes (in contrast to 35% and 25% for the top and bottom layers, respectively) , making it the most important information processing structure in the regulatory network (Jothi *et al.*, 2009).

According to the hypothesis of increased retention of high-pleiotropy TFs after WGD, I would expect WGD TFs to be represented across all three network layers, including members of the top and core layers which are likely to have the largest pleiotropic effects. Table 4.6 shows the hierarchical classifications of TFs in my dataset as well as observed and expected numbers for the WGD TFs. Again, I found no significant differences in the distributions of WGD and non-WGD TFs across the hierarchical layers of the network (χ^2 test: P = 0.512; df = 2) but rather a balanced distribution.

In light of the high-pleiotropy hypothesis for WGD TFs that I had set out to test, it thus appeared that indeed there was unbiased retention of both high- and low-pleiotropy TFs following the WGD, arguing in favour of the gene balance hypothesis. Moreover, the significant enrichments for high indegree in both networks and high outdegree in the Jothi2009 network, and the (not significant but suggestive) trend for WGD TFs to be regulatory hubs certainly argue for the retention of important regulators following the WGD. One aspect of this might be that both hub status and position within the regulatory hierarchy are poor definitions of pleiotropy. Indeed, I found that e.g. hub status had no effect on

evolutionary rate constraints (Chapter 5; Fig. 5.5A) and only the top layer TFs had significantly more constrained rates compared to TFs in other layers (Chapter 5; Fig. 5.5A). This suggests that there might be other factors determining constraint and pleiotropic impact and will be discussed further in the conclusions of this Chapter and Chapter 7.2.2.

Alternatively, it is possible that what I observed for WGD TFs above is still very different from what is seen for SSDs. One can argue that a TF with few regulatory interactions, in the bottom layer of the regulatory network is no less likely to be retained after WGD than a highly-connected one. Conversely, retention of hubs and highly influential TFs after SSD might be strongly selected against in the absence of other, possibly interacting, duplicate TFs to balance the impact of the increased dosage. In order to explore this possibility, I examined the properties of SSD duplicated TFs which are discussed below.

4.5.2 The Properties of Small-scale Duplicated Transcription Factors

In continuation of the argument above and in contrast to the enrichment for highly connected genes among the WGD-retained TFs, one would expect duplicates generated through SSD to display the opposite trend, i.e. low connectivity, positioning in the bottom layer of the regulatory network and non-hub status. Assuming that those are subject to dosage-dependent pleiotropic constraints, their retention after SSD should be selected against and consequently rare. The very small number of recent SSDs (Table 4.7) on the lineage leading to *S. cerevisiae*, however, makes it impossible to perform statistical analyses analogous to the ones presented above. Examination of the properties of the three SSD-generated gene pairs provided some interesting insights.

All six TFs were found to have low indegree in both the Jothi2009 and YT networks. Similarly, the outdegree for three of the four TFs included in the Jothi2009 network were low (see Figs. 4.13 and 4.14 for overall distributions). This was in contrast to annotated interactions in the YT network where the two included TFs had 63 and 98 target genes. Only two TFs were found in the description of the hierarchical structure of the network, one in the top layer

	Jothi2009		YT			
SGD ID	**Out**	**In**	**Out**	**In**	**Position**	**Hub**
YER088C	N/A	2	N/A	7	N/A	N/A
YBL054W	1	0	N/A	9	N/A	N/A
YKL038W	9	1	63	3	Bottom	No
YBR033W	3	1	N/A	6	N/A	N/A
YLR098C	51	0	98	2	Top	No
YOR337W	N/A	3	N/A	4	N/A	N/A

Table 4.7: Degree distribution and position in the regulatory network of small-scale duplicated TFs. Paralogs are grouped together in the same section.

and one in the bottom layer, and none of those was found to be a hub. While these observations were somewhat inconclusive, especially in consideration of the YT network, a large part of these TFs was not characterised in either dataset or had a low indegree and outdegree arguing in favour of the low-pleiotropy hypothesis for SSD TFs. An especially interesting exception are the two paralogs in the last two rows of Table 4.7, YLR098C and YOR337W, where one member regulates a considerable number of target genes, whilst its duplicate has remained unstudied, raising the possibility of divergence after duplication or condition-specificity. Considering the short time-scale of binding site turnover (e.g. Schmidt *et al.*, 2010), this raises the question of how fast a newly duplicated transcription factor is assimilated into the existing network and can migrate into more densely connected parts of the network given that it arose as a duplicate of a sparsely connected TF.

In addition to the SSD duplicates that arose within the *Saccharomycetaceae*, I also investigated the position and connectivity of the *S. cerevisiae* homologs of TFs that had undergone lineage-specific amplification in the CTG clade. Because at this moment there is no complete characterisation of the regulatory network in any of those species, this required mapping to the *S. cerevisiae* regulatory network for comparison. In total, there were 126 orthogroups for which duplications (assumed and *bona fide*) were inferred, just under 80% of which were, however,

lineage-specific and were not mapped to a TF in *S. cerevisiae*. The remaining 27 orthogroups could be mapped to 31 *S. cerevisiae* homologs and an analysis of these is presented below. By definition these are all *bona fide* duplications due to the requirement of the presence of an ortholog in *S. cerevisiae*.

Figures 4.16 and 4.17 show the outdegree and indegree distributions of the TFs that were amplified in the CTG clade (red) with respect to the outdegree and indegree of non-amplified TFs in the *S. cerevisiae* regulatory network (grey). The distributions were not found to be biased towards highly connected TFs for either outdegree (Jothi2009: one-tailed Wilcoxon test: W = 1354; P = 0.547, YT: one-tailed Wilcoxon test: W = 1177.5, P = 0.820) or indegree (Jothi2009: one-tailed Wilcoxon test: W = 1830.5, P = 0.662, YT: one-tailed Wilcoxon test: W = 2559, P = 0.815) in either network. This was also evident from the relative connectivity distributions (Fig. 4.18) where CTG-amplified homologs often clustered in the lower connectivity regions. Nevertheless, there were numerous highly-connected TFs among the CTG-amplified homologs, arguing against a strict low-pleiotropy rule for retention of SSD duplicates. Again, interpretation is not straightforward given the large divergence time between *S. cerevisiae* and the extant members in the CTG clade. Several recent studies examining the conservation of binding events in the *Saccharomycotina* have shown how fast regulatory interactions are gained and lost and that TFs can migrate from having few regulatory interactions to becoming hubs within a few hundred million years (e.g. Lavoie *et al.*, 2010; Tuch *et al.*, 2008a).

The analysis of the CTG-amplified homologs in the context of the network architecture led to similar conclusions. When considering the frequency of hubs among the duplicates I found no significant difference from the background distribution (Table 4.8), although here the overall trends were reversed with respect to the WGD TFs, i.e. I found fewer hubs than expected among the CTG-amplified TFs as opposed to more than expected among the WGD TFs. Here, the proportion of TFs that was not included in the description of the regulatory network was almost identical to the expected number. This was a strong contrast to what I found in WGD TFs (χ^2 test: P = 0.514; df = 2). This lent further weight to the possibility that there might be biological reasons, such as importance in the

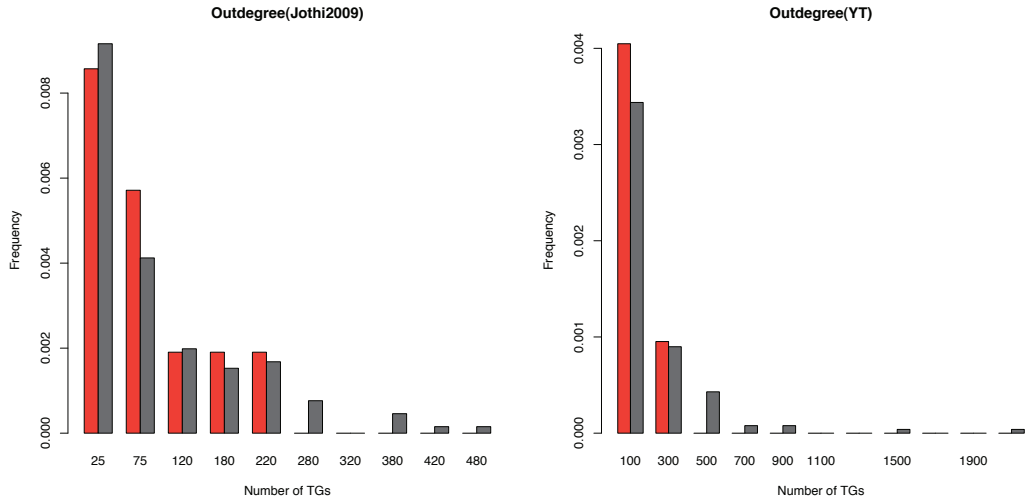


Figure 4.16: Distribution of the outdegrees of CTG TF duplicates (red) compared outdegree distribution of not amplified TFs (grey) in the Jothi2009 and YT regulatory networks.

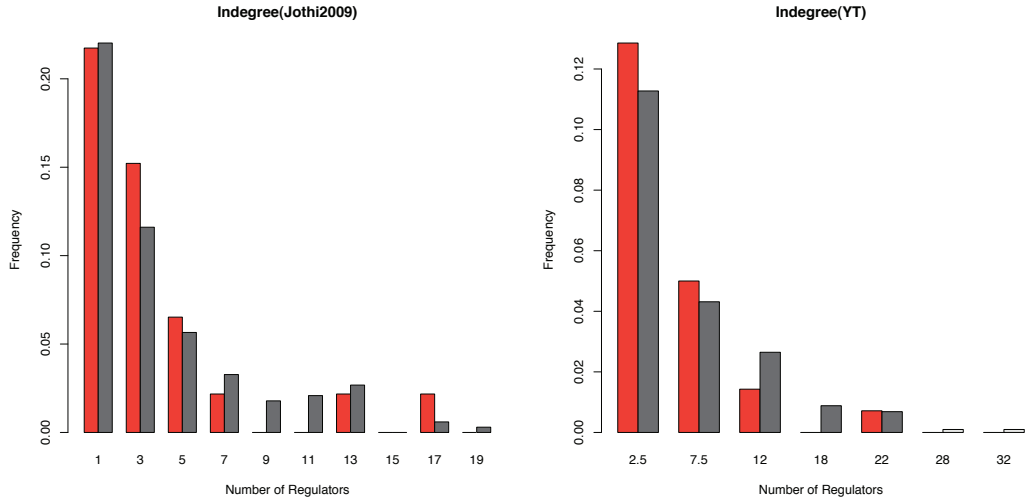


Figure 4.17: Distribution of the indegrees of CTG TF duplicates (red) compared to the distribution of indegrees of non-amplified TFs (grey) in the Jothi2009 and YT regulatory networks.

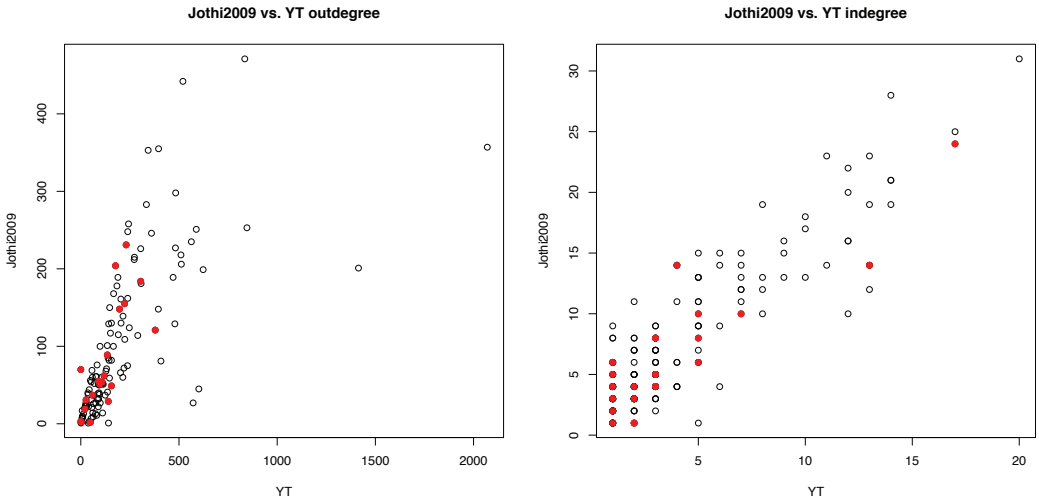


Figure 4.18: Distribution of outdegrees of CTG TF duplicates (red) compared to the overall distribution of outdegrees (black) in the Jothi2009 and YT regulatory networks (left plot). Distribution of indegrees of CTG TF duplicates (red) compared to the indegree distribution of all TFs (black) in the Jothi2009 and YT regulatory networks (right plot)

	Hubs	Non-hubs	N/A	Total
All	32	102	93	227
CTG	3	17	11	31
Expected	4.78	15.22	—	—
Expected (N/A)	4.37	13.93	12.70	—

Table 4.8: Distribution of regulatory hubs among TFs duplicated in the CTG clade. N/A refers to TFs in my dataset that were not included in the Jothi2009 regulatory network.

	Top	Core	Bottom	# N/A	Total
All	18	61	55	93	227
WGD	3	7	10	11	31
Expected	2.69	9.10	8.21	—	—
Expected (N/A)	2.46	8.33	7.51	12.70	—

Table 4.9: Distribution of CTG-amplified TFs among the hierarchical layers of the network. N/A refers to TFs in my dataset that were not included in the Jothi2009 regulatory network.

regulatory network or family-specificity that were captured by the strong under-representation of uncharacterised TFs among the WGD duplicates. Similarly, I found no significant deviations of the distribution of CTG-amplified TFs across the hierarchical layers of the network (Table 4.9; χ^2 test: P = 0.634; df = 2).

Taken together, the results obtained from analysis of SSD TFs were not strong enough to either reject or support the hypothesis of preferential duplication of low-pleiotropy genes in small-scale duplications. Nevertheless and in contrast to the WGD-retained TFs, which were significantly enriched for highly-connected TFs, I found fewer regulatory hubs than expected and more bottom layer TFs than expected among the SSD TFs although these enrichments were not significant. It is more than likely that some of the TFs that were amplified in the CTG clade have been rewired and experienced gain and loss of regulatory interactions since the divergence from the *Saccharomycetaceae*. It is known from

comparative studies of bacterial regulatory networks that both global and local structures can change drastically and completely different transcription factors can be "promoted" to become global regulators in different species (Babu *et al.*, 2006b). The extent to which this is true in eukaryotic regulatory networks, which are generally more complex and interconnected (Sellerio *et al.*, 2009), is currently unknown and it would be very interesting to examine degree distributions and placement of the CTG-specific duplicates within the network architecture of a more closely related species.

4.6 Functional Implications of Regulatory Network Growth

Analysis of the position of duplicated TFs within the regulatory network showed that WGD paralogs were enriched for highly-connected TFs and as such are likely to perform important roles in the regulatory network. To examine the biological relevance of the TFs that were retained after the WGD, I determined their functional role based on annotation available at the Saccharomyces Genome Database (SGD) as well as published literature. Annotations for the 63 WGD TFs can be found in Appendix, Table A.2. Functional classification revealed that a striking number of WGD TFs participate in the regulation of stress- and nutrient-signalling. In total, 35% of regulators are known to be directly involved in the response to a variety of extra- and intracellular stresses, including important master regulators such as MSN2/4 (general stress response), the YAP family (oxidative stress response, salt tolerance), AFT1 and AFT2 (iron homeostasis) and RLM1 (cell integrity). A further 10% of TFs regulate carbohydrate and nitrogen metabolism in response to nutrient availability and metabolic state (i.e. aerobic or anaerobic). Other highly represented groups included cell cycle regulators and a number of TFs mediating drug resistance.

The post-WGD species are Crabtree-positive, meaning that they preferentially ferment glucose even under aerobic conditions, and do so at a high rate (see Chapter 1.5.2.1). This is likely to be a derived metabolic feature: even

though most species that did not undergo a WGD event are facultative anaerobes, none of them displays an anaerobic metabolism as efficient as the post-WGD species (Merico *et al.*, 2007). This efficiency requires the tight coordination of metabolic processes in response to nutrient availability and extracellular conditions. In high-glucose environments *S. cerevisiae* grows exponentially and genes required for oxidative respiration, use of alternative carbon sources and response to stresses are repressed. Genes required in rapid growth conditions such as ribosomal proteins are highly expressed and cell cycle progression is fast (Brauer *et al.*, 2008; Gasch *et al.*, 2000). In contrast to the *Saccharomyces* species, high growth rates in non-WGD species are likely to be linked to aerobic metabolism, thus requiring the high-level, coordinated expression of different pathways in nutrient-rich environments. Indeed, recent studies of ribosomal protein regulation in the *Saccharomycotina* have shown incidences of large-scale rewiring that lead to the separation of the regulation of mitochondrial and cytosolic ribosomal proteins (e.g. Ihmels *et al.*, 2005) as well as a breakdown of coregulation between respiratory and cytosolic ribosomal proteins (e.g. Lavoie *et al.*, 2010) in the post-WGD species. The very high occurrence of stress- and nutrient-responsive regulators as well as cell cycle regulators among the WGD TFs I have collected thus provides evidence that the WGD event might have substantially influenced the rewiring of the regulatory network and the ability to adapt to new environments in *S. cerevisiae* and its close relatives.

Furthermore, I examined the functional annotation for TFs that were the result of amplification in the CTG clade. Striking here was that a small number of orthogroups were heavily amplified (see Appendix, Table A.3). One orthogroup containing TFs of unknown function experienced ten duplication events. Annotations for CTG TFs were much more sparse compared to the annotation available for *S. cerevisiae*, however: over 50% of the TFs were not experimentally characterised. Nevertheless, available annotation underlined the biological importance of the observed duplications with a number of TFs found to be involved in the regulation of drug resistance, copper and iron homeostasis and hyphal gene expression, all of which are known to be important for pathogenicity and virulence (Banerjee *et al.*, 2008; Miyazaki *et al.*, 2010; Sellam *et al.*, 2010). Another interesting observation was the heavy amplification of two orthogroups containing TFs

of unknown function that show homology to the *S. cerevisiae* lysine metabolism regulator *LYS14*. Overall, I inferred 14 duplications on various branches in these two orthogroups. A literature search revealed a recently published study that found a number of lysine metabolic genes to be up-regulated in high osmotic stress conditions in *C. albicans* (Bruno *et al.*, 2010), suggesting a possible role for these highly amplified TFs.

4.7 Conclusions

Analysis of patterns of gains and losses experienced by TF repertoires in the *Saccharomycotina* revealed clade-specific differences in evolutionary dynamics across the three major lineages studied (Fig. 4.11). Network growth in the species that underwent a WGD event ∼100 million years ago appeared to be mainly dominated by this large scale duplication, with very few SSD events to follow. The CTG clade in turn displayed constant rates of both gene gain and loss through frequent SSD events.

The differences in types of duplication events generating network growth in the WGD and CTG species were also reflected in family-specific rates of evolutionary turnover. While there was no particular family that experienced significantly increased rates of duplication, or rather retention in this case, I found significantly accelerated rates of duplication for the Zn_clus and Zn_clus:Fungal_trans TFs (the two largest families in *Saccharomycotina* TF repertoires) in the CTG species. This suggested that ongoing lineage-specific amplification through SSD of certain families was the predominant mechanism driving regulatory network growth in these species.

Furthermore, WGD TFs appeared to display different properties from SSD TFs with respect to their role and position in the regulatory network. In contrast to SSD duplicates, I found WGD duplicates to be enriched for highly-connected TFs. While this provided some evidence for the gene balance hypothesis, predicting retention of high-pleiotropy genes after the WGD, the balanced distribution of WGD TFs among regulatory hubs and the hierarchical layers of the network gave further support to the high-pleiotropy theory. The gene balance hypothesis is straightforwardly applied to PPI networks but leaves more room

for interpretation in regulatory networks. Disturbance in relative quantities of different TFs could lead to decreased accuracy of transcriptional initiation, over- or under-expression of target genes, competitive binding through an excess of free TF molecules or simply a dosage-dependent effect that would make duplication of a single TF without duplication of one or more of its target genes energetically unfavourable in absence of selection for increased dosage of the target gene. Other aspects that potentially contribute to the impact of duplication of a TF is whether it is expressed in a condition-specific manner and, if so, how many conditions it is expressed in. It is known, for example, that TFs can be regulatory hubs in some conditions but not in others (e.g. Luscombe *et al.*, 2004). This would be an interesting target for further study.

Nevertheless, there was a clear trend for retention of highly connected TFs after WGD. Enrichments were strongest for WGD TFs to have large indegrees and were robustly detected in both regulatory network used for testing. This suggested that WGD duplicates were themselves more highly regulated which may have aided quicker degradation of functional redundancy between the duplicate copies and thus increased selective pressure for retention of both copies. Indeed, a recent study of posttranslational modification among duplicate genes has shown that the number of phosphorylation sites in a protein is a strong determinant of gene retention (Amoutzias *et al.*, 2010), supporting this theory. Theoretically this should equally apply to SSD-generated paralogous TFs, although here I failed to detect similar trends in TFs that were amplified in the CTG clade. Again, conclusions drawn from the comparison of CTG homologs are vague at best, due to the inability to map more than 50% of homologs to *S. cerevisiae* TFs and the high plasticity of regulatory networks in general.

Independent of whether or not WGD and SSD duplications differ in their pleiotropic impact, it was evident that the WGD carried great adaptive potential through duplication of an excess of well-connected transcriptional regulators that provided ample raw material for diversification and fine-tuning of core regulatory pathways. Indeed, recent studies investigating the WGD subnetwork have found that WGD TFs were rapidly integrated into the regulatory network, do not show greater redundancy than ancient TFs and form a strongly interconnected structure (Conant, 2010; Fusco *et al.*, 2010). My functional annotation of WGD

duplicates showed a strong signature for retention of stress response, metabolic and cell cycle regulators, all of which are of great relevance to the differences in life style found between post-WGD species and the remaining *Saccharomycotina*. This was also reflected in analysis of evolutionary rates (Chapter 6.5) and will be discussed further in Chapters 6.5.2 and 7.2 and 7.3.

Besides these implications for regulatory evolution in the *Saccharomycotina*, the results obtained here provide important insights for the study of regulatory evolution in other eukaryotic organisms. The observations made in the *Saccharomycotina* suggest that lineage-specific amplification of certain families of TFs through SSD is the major evolutionary force driving the expansion of TF repertoires in most higher eukaryotes. These claims are not new; the importance of lineage-specific amplification has been appreciated in general, e.g. by Lespinet *et al.* (2002) and Iyer *et al.* (2008), as well as for gene regulatory networks and their evolution in particular, e.g. by Aravind *et al.* (2009) and Nowick & Stubbs (2010). A direct comparison between these different growth mechanisms shaping TF repertoires and their functional consequences for the evolution of regulatory networks from closely related organisms is however still lacking. In fact, the *S. cerevisiae* network, which is the only eukaryotic regulatory network that has been globally characterised to an appreciable extent, represents an evolutionary oddity due to the impact of the recent WGD. In order to start understanding questions such as how duplication of TFs contributes towards evolutionary novelty, how quickly new TFs are incorporated into the regulatory network and at what hierarchical level new TFs can arise, especially looking towards complex higher eukaryotes where recent whole-genome duplication events are very rare (e.g. Van de Peer *et al.*, 2009), another well-characterised regulatory network that has experienced more similar modes of network growth to those is needed.

Chapter 5

Evolutionary Rate in TF Repertoires

5.1 Introduction

Evolutionary analysis of turnover in transcription factor (TF) families revealed large-scale changes in TF repertoires through both gain and loss of transcriptional regulators (see Chapter 4). Besides changes in the numbers of TFs encoded in a genome, regulatory interactions can be altered through mutations in the protein-coding sequence of TFs, potentially leading to differences in DNA-binding affinity, protein-protein interaction (PPI) potential or post-translational regulation of TFs. In oder to determine the impact of protein-coding evolution in TF repertoires, I estimated evolutionary rates of orthologous regulators and related those to the regulatory network structure and degree dependence, especially when accounting for the effects of the whole-genome duplication (WGD) which contained many highly-connected TFs (see below). Furthermore I investigated relative rates of DNA-binding domains (DBDs) and non-DBD regions and their relative impact on associations seen between rate and network structure.

The rate of protein evolution in yeast has been shown to be dependent on a number of variables. The single most strongest determinant of evolutionary rate is absolute mRNA expression level and, related to that, codon bias (e.g. Drummond *et al.*, 2005; Xia *et al.*, 2009; reviewed in Pál *et al.*, 2006). Slowly evolving proteins tend to be enriched for high abundance, essential genes, frequent gene

duplication, a large number of interaction partners, low levels of intrinsic disorder and a larger number of regulators (Xia *et al.*, 2009). These associations are not straightforward however through interdependence of those factors and the small effect size of individual components (reviewed in Pál *et al.*, 2006). In PPI networks for example, the degree dependence of evolutionary rate first detected by Fraser *et al.* (2002) was later disputed (e.g. Jordan *et al.*, 2003) and shown to be dependent on whether a protein is a stable or transient interaction hub among other factors (Mintseris & Weng, 2005).

Similar studies of regulatory networks have failed to detect associations between the number of target genes a TF regulates (its outdegree) and evolutionary rates in bacteria (Lozada-Chávez *et al.*, 2006; Price *et al.*, 2007) and yeast (Evangelisti & Wagner, 2004; Jovelin & Phillips, 2009; Wang *et al.*, 2010). Instead, Wagner & Wright (2007) have found that redundancy, the number of alternative routes connecting regulators above a TF and a TFs target genes, was positively correlated with evolutionary rate. More recently, both Jovelin & Phillips (2009) and Wang *et al.* (2010) detected positive associations between the number of regulators TF is regulated by (its indegree) and evolutionary rates, in contrast to what has been found genome-wide, where slow-evolving proteins had a larger number of regulators (e.g. Evangelisti & Wagner, 2004; Xia *et al.*, 2009). Hierarchical organisation of regulatory networks has also been found to impact evolutionary rate (Bhardwaj *et al.*, 2010b), conservation of TFs between species (Jothi *et al.*, 2009) and phenotypic impact of regulator deletion (Bhardwaj *et al.*, 2010a) and as such appears to capture the evolutionary constraints of regulatory networks better than degree-dependence. The top layer TFs were found to be most conserved across species and showed the lowest evolutionary rates, whereas regulators in other layers appeared to be evolutionarily more flexible.

In my analyses of gene duplications and losses in the *Saccharomycotina*, I found that over a quarter of TFs in *S. cerevisiae* were duplicates that were retained after a relatively recent WGD event and that those were enriched for highly-connected regulators. It is known that following gene duplication, genes often experience accelerated rates of evolution through relaxation of selective constraints due to the initial redundancy of the duplicates (e.g. Byrne & Wolfe, 2007; Scannell & Wolfe, 2008). Conceivably, the effects of the WGD might contribute

to the lack of a strong correlation of evolutionary rates with outdegree or the positive correlation with indegree observed in the yeast regulatory network (see above) due to the relaxed constraints of highly connected duplicates. To test whether this was the case, I estimated evolutionary rates of TFs and dissected the observed correlations based on their duplication status (WGD and non-WGD) and hierarchical position in the regulatory network.

5.2 Estimating Evolutionary Rates

5.2.1 Relative Rate Scaling

Evolutionary rates for each orthogroup (OG) were estimated using a relative scaling approach as illustrated in Figure 5.1. ML branch lengths are defined as the inferred number of substitutions per site and as such represent a direct estimate of the amount of evolutionary change of a given gene. Unless speciation times are known, it is not possible to obtain an absolute estimate of evolutionary rate but the comparison of amounts of evolution experienced by different genes on the same set of branches allows us to derive a relative rate based on the fact that both subtrees will cover the same amount of real time. By scaling the branch lengths of the gene tree of each OG to those of a reference gene I was able to obtain a relative estimate of the evolutionary rate between different OGs. Here, the reference alignments were trimmed to the same set of species as those in the respective OGs, assuming a shared underlying phylogeny, thereby being able to estimate the best fit of the observed sequences to the common tree and derive a relative rate of evolution between OG and reference. Compared to other, more direct approaches for estimating evolutionary rates this diminished the need for each OG to contain the same set of species by dwelling on a relative rather than an absolute estimate of evolutionary rate for each OG which would require them to contain the same set of branches and thus cover the same amount of real time. This increased the number of comparable OGs. The relative rates estimated here are thus defined as the number of substitutions per site of the OG in question relative to the number of substitutions per site of a reference gene for the same underlying set of species.

Each OG was split into the number of one-to-one orthologous subgroups ("splits"), induced by the inferred duplications (see Chapter 4.4.1). This resulted in 633 splits, 371 of which contained more than two species. To maximise alignment quality, the amino acid sequences of each split were realigned using the domain-guided version of PRANK described in the previous chapter (Chapter 4.2) and backtranslated into codon alignments. Three reference genes, *TUB2*, *H2A* and *TBP* were chosen from the dataset of Wall *et al.*, 2003, who estimated genome-wide evolutionary rates for *Saccharomyces cerevisiae* protein-coding genes. All three were among the 10% of most slowly evolving genes and I selected them as suitable reference genes based on the reasoning that those will yield the maximum number of informative sites across the evolutionary range of species considered in my dataset. Reference genes were aligned equivalently. For each split alignment, the reference alignment was trimmed to the set of species contained in the split and both alignments were concatenated.

PAML 4.0 (Yang, 2007) was used to estimate a partitioned model (see Chapter 2.2.1 for an in-depth discussion) for each of the concatenated alignments, providing the species tree as the reference tree. A general time-reversible model (REV; Rodríguez *et al.*, 1990) with four categories of gamma-distributed rates was used as the base evolutionary model and separate nucleotide frequencies (π), rate parameters (s_{ij} for i and j in A,T,G,C) and gamma distribution shape parameter (α) were calculated for each reference gene and spilt partition to allow for between-gene heterogeneity. Branch lengths for the split partition were scaled to those of the reference partition in a linear fashion, yielding a scaling factor c, which served as the relative rate estimate (see "Mgene 4", Chapter 2, Table 2.1). PAML results were filtered for runs that failed to produce a reliable estimate of c, removing all instances where the standard error of the estimate was infinite or larger than the value of c itself.

5.2.2 Consistency of Estimated Relative Rates

To investigate whether c is a consistent estimator of the relative evolutionary rate, I repeated calculations using three different reference genes and checked for consistency. The three runs differed in the number of splits for which reliable

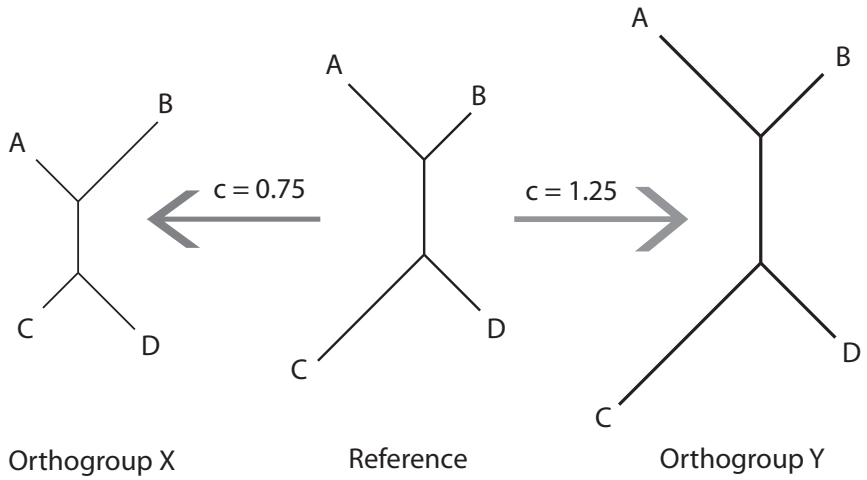


Figure 5.1: Relative rate estimation of individual orthogroups. Orthogroups X and Y are both scaled to a reference tree, obtaining a relative rate of evolution between orthogroup X and orthogroup Y.

estimates were obtained; *H2A* scaling performed best with 356 estimates, followed by *TBP* yielding 353 estimates and *TUB2* with 349. The number of splits for which good estimates were obtained using all three reference genes was 341. Figure 5.2A shows the correlation between c estimated for each split using *H2A* and *TBP*. Results comparing either of the scalings to *TUB2* were equivalent and are not shown here.

While the results correlated strongly ($P < 0.001, R^2 = 0.686$), I found three clear patterns of tight correlations, the main axis (Fig. 5.2A, red) and two additional components (Fig. 5.2A, a and b). These patterns were reproduced in any of the three two-way comparisons between different reference genes and always grouped together the same data points indicating that this might be due to properties of the data itself. Indeed, when investigating the splits that grouped into the different components I found that they were largely seperated by their phylogenetic depth. To further elucidate the factor causing these observed clade-dependent discrepancies in relative rates, I repeated the experiments as above but this time fixing the branch lengths of the input tree (Fig. 5.2B). Reference tree

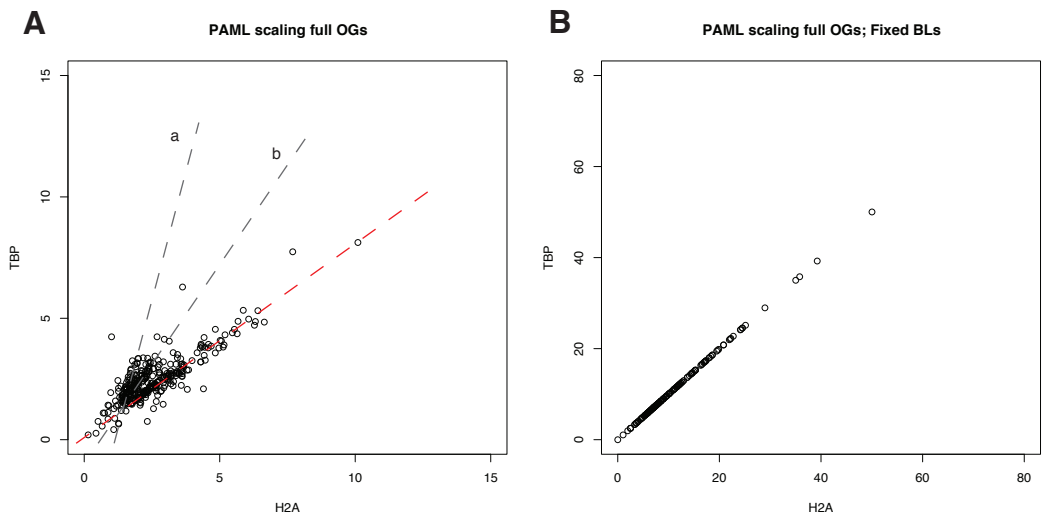


Figure 5.2: A: Correlations of relative rate c using two different reference genes, *H2A* and *TBP*. B: Equivalent experiment but with fixed branch lengths of the input tree.

branch lengths were calculated for each of the reference genes using the base evolutionary model REV+Γ. Here, I recovered a perfect correlation between results from different reference genes ($P < 0.001, R^2 = 1.0$) underlining the validity of the scaling approach. Fixing the branch lengths however decreased the number of runs that converged to a total of 305, with a further 49 and 102 splits failing to yield a reliable estimate for *H2A* and *TBP* respectively.

It seemed possible that these differences between estimates of c were caused by different tree shapes of the clade-specific subtrees dominating the scaling of branch lengths when not constrained by other clades in the alignment. To determine whether these effects were still dominant when considering each clade in isolation, I trimmed all splits to either the *sensu stricto* species, the WGD clade, the pre-WGD clade or the CTG clade and repeated the above experiments. Figure 5.3 shows the results for the respective clades. Here the patterns largely disappeared, also reflected in the improved correlations (*sensu stricto*: $P < 0.001, R^2 = 0.797$; WGD: $P < 0.001, R^2 = 0.865$; pre-WGD: $P < 0.001, R^2 = 0.960$; CTG: $P < 0.001, R^2 = 0.833$; CTG (no outlier): $P < 0.001, R^2 = 0.960$). While the slope of the correlations was close to one in most clades, it was closer to 1.5 in the

pre-WGD species indicating rate divergence between the two reference genes in this clade which could partially explain the patterns observed when c is compared across the entire tree.

I thus settled for using clade-wise relative rate estimates for my analyses, as these were less dominated by clade-specific rate effects while providing reliable estimates of c for most splits analysed. *H2A* was chosen as the final reference gene, as it provided the largest number of reliable estimates.

5.2.3 Evolutionary Rate Correlates

One of the hypotheses I set out to test was whether the previously reported lack of correlation between the evolutionary rate of a TF and its role in the regulatory network (e.g. Wang *et al.*, 2010) was due to the fact that analyses had not accounted for the effects of the WGD (see above). Evolutionary rate of protein-coding genes is known to be strongly influenced by their total expression level (Drummond *et al.*, 2005; Xia *et al.*, 2009). The Codon Adaptation Index (CAI; Sharp & Li, 1987) is a strong indicator of the expression level of a gene, based on the observation that highly expressed genes experience strong selection for optimal codon usage for efficiency and accuracy of translation. CAI for *S. cerevisiae* TFs was calculated using the implementation described in Xia (2007). Other factors potentially influencing evolutionary rates include the proportion of sites in unordered regions. TFs are known to contain extended unordered regions (e.g. Singh & Dash, 2007) which conceivably could have an impact on evolutionary rates, seeing that intrinsically disordered regions will be less structurally constrained and might therefore evolve faster. In order to see whether this was affecting the evolutionary rates calculated here, I used IUPred (Dosztányi *et al.*, 2005) to calculate the number and proportion of sites in intrinsically disordered regions for all *S. cerevisiae* TFs. As in the previous chapter, I used the regulatory network obtained from YEASTRACT (Teixeira *et al.*, 2006) as well as the one compiled by Jothi *et al.* (2009) (referred to as YT and Jothi2009 respectively) and their inferred hierarchical organisation to analyse evolutionary rates within the context of the network.

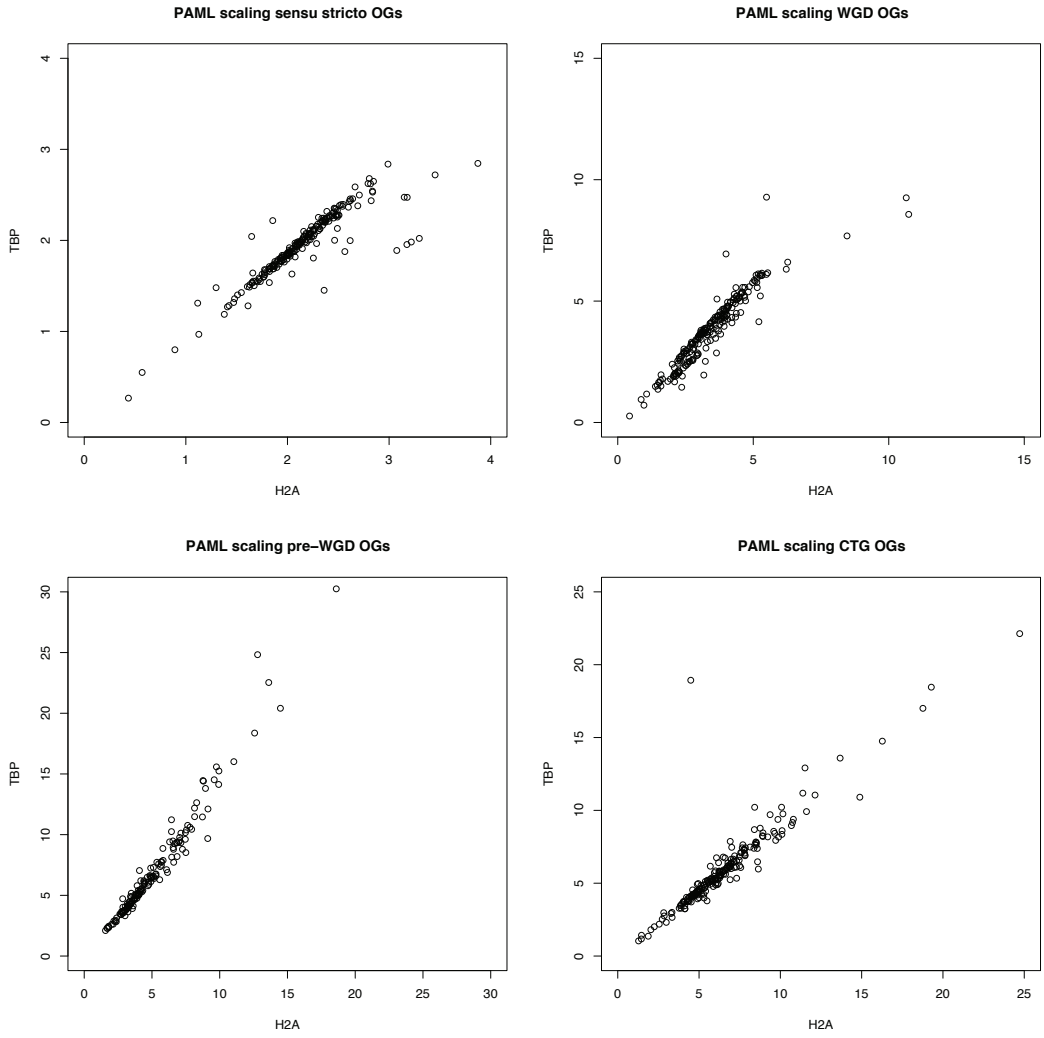


Figure 5.3: Clade-specific correlations of relative rate c using two different reference genes

5.3 Evolutionary Rate is Unrelated to TF Outdegree but Related to Indegree and Hierarchical Structure of the Network

To test hypotheses about the relationship between network architecture, TF connectivity and evolutionary rate, I focussed on the *sensu stricto* species only seeing that those are most closely related within the clade and should thus yield the most reliable rate estimates, as well as representing the group that the *S. cerevisiae* network architecture can be inferred on with reasonable levels of comfort (*Saccharomyces castellii* and *Candida glabrata*, although post-WGD, have divergent TF repertoires, certainly in terms of content [Chapter 3.6.3] and likely in terms of sequence [Chapter 4.4.1] and it is unclear how this affects network structure). Overall, I obtained reliable rate estimates for 220 of the 243 splits that contained representatives within the *sensu stricto* species on which the following analysis was based.

I first investigated the relationship between CAI and the relative rate estimates, the results of which are shown in Figure 5.4A. The majority of *S. cerevisiae* TFs had very similar CAI values, falling between 0.20 and 0.25. CAI values range from 0 to 1, with higher values indicating increased usage of the most common codons. As such the range observed here is relatively narrow and comparatively low. There were a few exceptions with high CAI values and these were indeed among the more slowly evolving TFs. Surprisingly however, I found significant positive correlation between CAI and relative rates, although this was very weak ($P = 0.001, R^2 = 0.043$).

Based on those results, it was clear that the CAI had no major influence on the evolutionary rate estimates in the *sensu stricto* species and other more important determinants of evolutionary rate exist. Similarly, I found a broad distribution of the percentages of residues in disordered regions among the TFs (Fig. 5.4B) and, perhaps surprisingly, no correlation of evolutionary rates with the percentage of residues in unordered regions ($P = 0.358, R^2 = -7.1 \times 10^{-}4$). When considering the absolute number of residues that are found in disordered regions (Fig. 5.4C) it emerged that both extremely fast and extremely slow evolving TFs appear

to be devoid of extensive disordered regions, while most other TFs do not show obvious patterns of rate-dependence ($P = 0.570, R^2 = -0.003$). This suggested that neither CAI nor instrinsic disorder exhibit a strong linear relationship with relative evolutionary rates of TFs and other functional constraints are likely to be in place.

In line with previous reports (e.g. Wang *et al.*, 2010), despite a slight negative trend, I found no significant correlation between a TF's rate of evolution and its outdegree in either the Jothi2009 or YT networks (Spearman's rank correlation; $r_s = -0.111; P = 0.185$ and $r_s = -0.004; P = 0.953$ respectively). Indegree, however, did show a significant positive correlation with evolutionary rate (Spearman's rank correlation; Jothi2009: $r_s = 0.241; P = 0.002$; YT: $r_s = 0.226; P = 0.001$). This also mirrored results obtained in previous studies (Jovelin & Phillips, 2009; Wang *et al.*, 2010).

When I examined the rates of regulatory hubs with respect to non-hubs, I found no significant difference (two-sided Wilcoxon test; W = 1435.5, P = 0.703) although the variance of relative rates of non-hubs appears to be bigger than that of hubs (Figure 5.5A). Finally, when considered within the hierarchical organisation of the network, I found that top-level TFs evolve significantly slower than TFs in the core and bottom layers (Kruskal-Wallis test; P = 0.021). Similar trends have been shown by Bhardwaj *et al.* (2010b) and Jothi *et al.* (2009), who examined ranked evolutionary rates of *S. cerevisiae* transcriptional regulators and the number of conserved orthologs in other fungi, respectively. Jothi *et al.* also determined the relative expression levels of TFs in different layers of the network and have shown that regulators in the top level are relatively more abundant. To see whether this difference in relative rates could be explained by codon bias as a proxy for expression level, I compared CAI values between the top and the core and bottom levels and found no significant difference (two-sided Wilcoxon test; $W = 1062; P = 0.909$), suggesting this to be a true evolutionary constraint imposed by the hierarchical organisation.

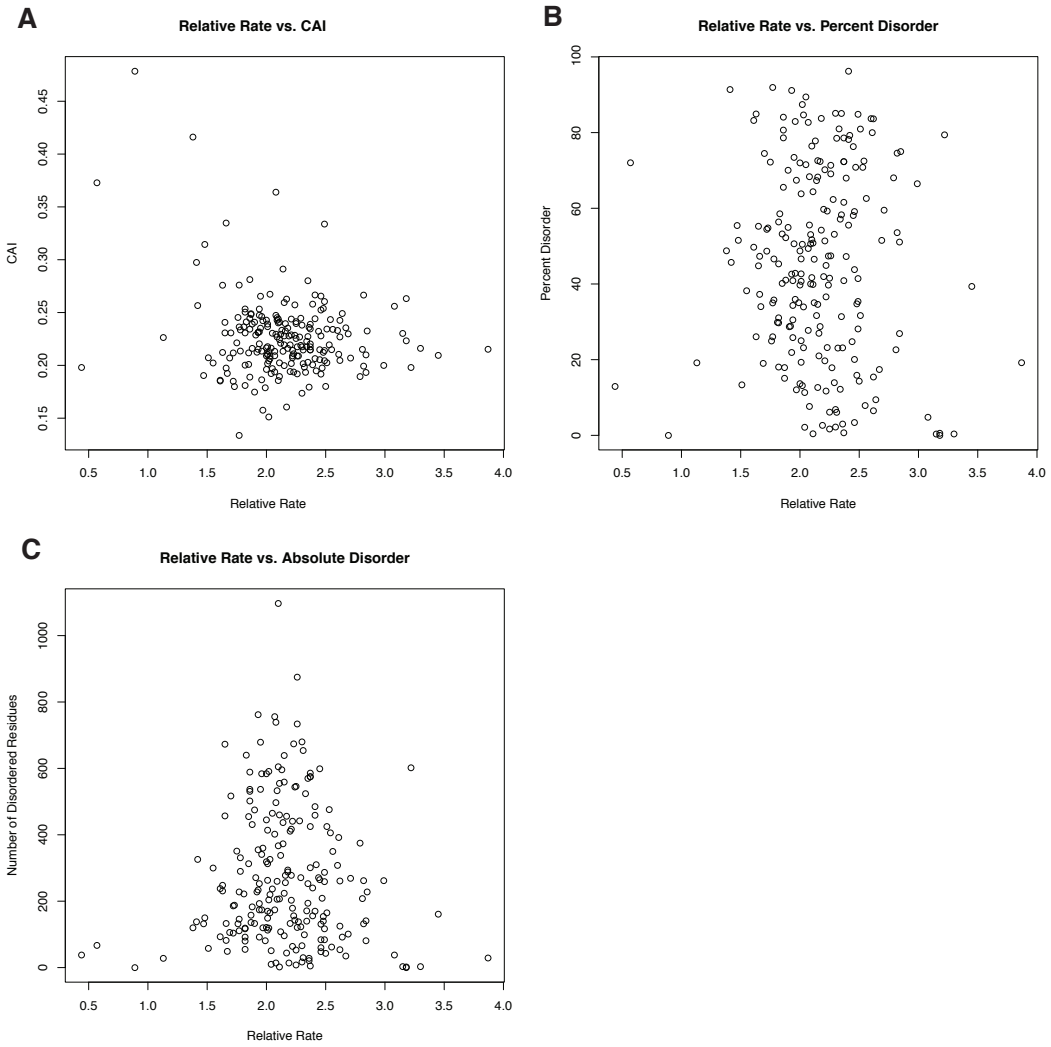


Figure 5.4: The influence of (A) CAI and (B) the precentage, and (C) the total number of residues in disordered regions on relative rate estimates among the *S. cerevisiae* transcription factors.

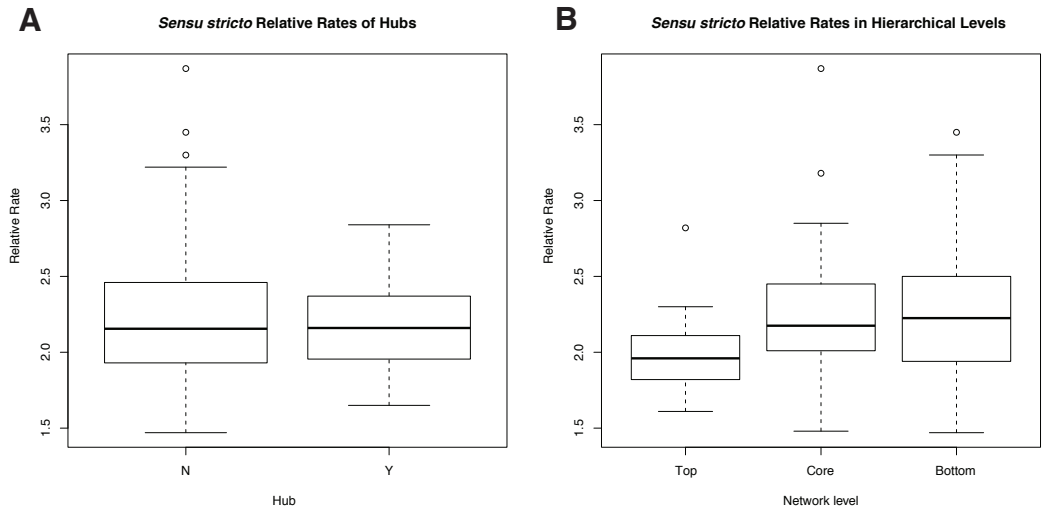


Figure 5.5: Relative rates of TFs that are (A) regulatory hubs and non-hubs and of for (B TFs residing in different hierarchical layers of the network)

5.3.1 Hierarchical Levels Show Distinct Trends for Degree-dependence of Evolutionary Rates

Hierarchical levels within the *S. cerevisiae* network show distinct functional and evolutionary properties. The top level contains master regulators that are by definition regulated by very few TFs and serve as integrators of intra- or extra-cellular signals, initiating a transcriptional response that is fed into the down-stream network (Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009). The core level is highly collaborative, regulating a large number of biological processes, whereas the bottom level is "stand-alone", corresponding to TFs that largely regulate specific sets of non-TF target genes. These differences are reflected in overall evolutionary rates (Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009 and this study; Fig. 5.5B), where top level TFs evolve significantly slower than core and bottom level TFs. Conceivably, degree dependence of evolutionary rates is likely to be different depending on whether a TF is at the initiating (top), processing (core) or distributing (bottom) level of the system. In oder to examine whether this was the case, I separated the yeast regulatory network into the three hierarchical levels described by Jothi *et al.* (2009) and analysed degree-dependence within

there levels.

Overall, indegree had been found to be positively correlated with evolutionary rate (Jovelin & Phillips, 2009; Wang *et al.*, 2010, see above). Whereas this held for the bottom layer of the regulatory network (Spearman's rank: $r_s = 0.274; P = 0.054$), no significant association was found in the core layer of the network (Spearman's rank: $r_s = 0.081; P = 0.545$). Indeed, when excluding TFs residing in the bottom layer, this positive association was no longer significant across the remainder of the network (Spearman's rank; Jothi2009: $r_s = 0.045; P = 0.291$; YT: $r_s = 0.132; P = 0.349$) indicating that this signal was mainly provided by bottom-layer TFs. Similarly, although not significant, I found a stronger negative relationship between outdegree and evolutionary rates in the top layer (Spearman's rank: $r_s = -0.283; P = 0.255$), compared to the core and bottom layers ($r_s = -0.080; P = 0.550$ and $r_s = -0.028; P = 0.842$, respectively). No difference in trends of outdegree-dependence between the hierarchical layers was found in the YT network, however (Table 5.1).

While the stronger trend for outdegree dependence in top-layer TFs was intuitive because one would expect a master regulator in the top layer with large number of downstream targets to evolve slower due to the (by definition) low levels of redundancy, the positive relationship between the number of TFs a bottom-level TF is regulated by and its evolutionary rate remained puzzling. It is known that TFs frequently contain a large proportion of intrinsically disordered regions and that this proportion is in positive correlation with the number of target genes a TF regulates (Singh & Dash, 2007 and this study, results not shown). While this conceivably affects evolutionary rate due to the relaxed structural constraint, potentially masking degree-dependence of evolutionary rates, I found no correlation between the proportion of sites in intrinsically disordered regions and the relative evolutionary rates in my dataset (see above). Moreover, indegree and the intrinsic disorder did not correlate at all (Jothi2009: $r_s = 0.016; P = 0.834$, YT: $r_s = 0.029; P = 0.644$), excluding relaxed structural constraints as an explanation for the positive association between indegree and evolutionary rate. This suggested the presence of an alternative evolutionary pressure, other than lack of structure or high expression level to be driving this positive relationship. Based on overall evolutionary rates, it appears that bottom layer TFs evolve

	All	Non-WGD	WGD
Network			
Indegree (Jothi2009)	**0.241 (P =0.002)**	**0.194 (P =0.033)**	0.315 (P =0.057)
Indegree (YT)	**0.226 (P =0.001)**	**0.204 (P =0.008)**	0.184 (P =0.226)
Outdegree (Jothi2009)	-0.111 (P =0.185)	-0.185 (P =0.062)	-0.036 (P =0.823)
Outdegree (YT)	-0.004 (P =0.953)	-0.043 (P =0.664)	0.021 (P =0.897)
Top			
Indegree (Jothi2009)	—	—	—
Indegree (YT)	0.114 (P =0.661)	0.292 (P =0.334)	—
Outdegree (Jothi2009)	-0.283 (P =0.255)	-0.312 (P = 0.277)	—
Outdegree (YT)	-0.094 (P =0.708)	-0.200 (P =0.492)	—
Core			
Indegree (Jothi2009)	0.081 (P =0.545)	-0.058 (P =0.739)	0.133 (P =0.545)
Indegree (YT)	0.121 (P =0.367)	0 (P =1.00)	0.101 (P =0.645)
Outdegree (Jothi2009)	-0.080 (P =0.550)	-0.175 (P =0.314)	-0.174 (P =0.427)
Outdegree (YT)	-0.125 (P= 0.360)	-0.020 (P =0.911)	-0.342 (P =0.128)
Bottom			
Indegree (Jothi2009)	**0.274 (P =0.054)**	0.237 (P =0.146)	0.532 (P =0.092)
Indegree (YT)	0.261 (P =0.067)	0.280 (P =0.084)	0.173 (P =0.610)
Outdegree (Jothi2009)	-0.028 (P =0.842)	-0.079 (P =0.630)	0.318 (P =0.339)
Outdegree (YT)	-0.074 (P =0.641)	-0.081 (P =0.655)	-0.017 (P =0.982)

Table 5.1: Correlations of relative evolutionary rates in indegree and outdegree in the Jothi2009 and YT networks when seperated into hierarchical layers as defined by (Jothi *et al.*, 2009). Values given are Spearman's rank correlation coefficient r^s as well as the respective P value for each of the correlations. Correlations for top-layer WGD TFs could not be calculated due to small sample size.

fastest in general (Fig. 5.5B) and that divergence of the regulatory network is manifested strongest here. Whether this might be due to positive selection on a specific biological process (remember top and core layer TFs regulate more biological processes) or generally relaxed functional constraints at this level, however, remains open.

5.3.2 Similarities and Differences between Rate Profiles of Whole-genome Duplication Transcription Factors and Other Regulators

I found WGD TFs to be enriched for regulators with many regulatory interactions, distributed across the different layers of the regulatory hierarchy (Chapter 4.5.1). It is well-known that after gene duplication one or both paralogs often experience accelerated rates of evolution, either through a relaxation in selective constraint due to redundancy or positive selection driving neofunctionalisation or subfunctionalisation of either copy (e.g. Scannell & Wolfe, 2008, see introduction to this chapter). This increase in evolutionary rates could have potentially blurred the impact of the regulatory network structure on evolutionary rates in light of the overrepresentation of highly connected regulators among the WGD TFs. To determine whether this was the case, I repeated the analyses outlined above but this time distinguishing between TFs that arose through WGD and the remaining portion. Furthermore, I examined whether TFs residing in different hierarchical layers of the network were subject to different evolutionary pressures, across the WGD and non-WGD subsets only. The results of this analysis are shown in Table 5.1 ("WGD" and "non-WGD" columns).

Overall, I found no significant difference in relative rates between WGD and non-WGD TFs (mean 2.25 and 2.14 respectively). When considering the Jothi2009 network, again indegree was found to be positively associated with evolutionary rates in both non-WGD and WGD TFs ($r_s = 0.194; P = 0.033$ and $r_s = 0.315; P = 0.057$, respectively). Although just above the significance threshold, this trend was markedly stronger in the WGD TFs, indicating that at least part of this might have been driven by post-duplication divergence. Separation of the signal into individual hierarchical layers again high-

lighted the differences in indegree-dependence in different layers. While I found no significant association between indegree and evolutionary rate in the core layer (WDG: $r_s = 0.133; P = 0.545$ and non-WGD: $r_s = -0.058; P = 0.739$), there appeared to be stronger association in the bottom layer. These associations were not statistically significant (WDG: $r_s = 0.532; P = 0.092$ and non-WGD $r_s = 0.237; P = 0.146$) although it is noteworthy that the trend displayed by WGD TFs is both stronger and closer to statistical significance. There remains the possibility of a small sample size effect (the total numbers of WGD TFs in each layer are relatively low (see Chapter 4, Table 4.5) so interpretation of this was unclear, also in light of the YT network not supporting this trend.

Outdegree again was not found to be significantly associated with evolutionary rates, although the negative trend in the non-WGD TFs was close to statistically significant ($r_s = -0.185; P = 0.062$) whereas it appeared that this constraint has been relaxed in the WGD TFs ($r_s = -0.036; P = 0.823$) in line with my predictions about the influence of relaxed selective constraints following gene duplication. Outgedree-dependence, although not significant, was consistently weakly negative across all network layers in the non-WGD TFs. This trend was reversed in the bottom-layer WGD TFs, showing a relatively strong, but non-significant positive association ($r_s = 0.318, P = 0.339$), but again this was not supported by the YT network and was suffering from small sample size making interpretation somewhat difficult.

Comparison of the overall evolutionary rates of hubs and in the different hierarchical levels of the regulatory network in turn did reveal important differences between WGD and non-WGD TFs (Figure 5.6). Neither the WGD nor the non-WGD regulatory hubs evolved significantly slower than non-hubs. I found the average rate of WGD hubs, however, to be higher than that of WGD non-hubs (Fig. 5.6B) whereas the opposite was true for non-WGD hubs (Fig. 5.6A). Differences in evolutionary rates between hierarchical levels were even stronger between non-WGD and WGD TFs (Fig. 5.6C and D). Here, WGD TFs reflected the the trends observed when considering all TFs, where the top-layer regulators evolve significantly slower than the core and bottom-layer TFs (Kruskal-Wallis test; $\chi^2 = 8.915; df = 2; P = 0.012$) whereas I found no significant difference in rate between the hierarchical layers among the non-WGD TFs (Kruskal-Wallis

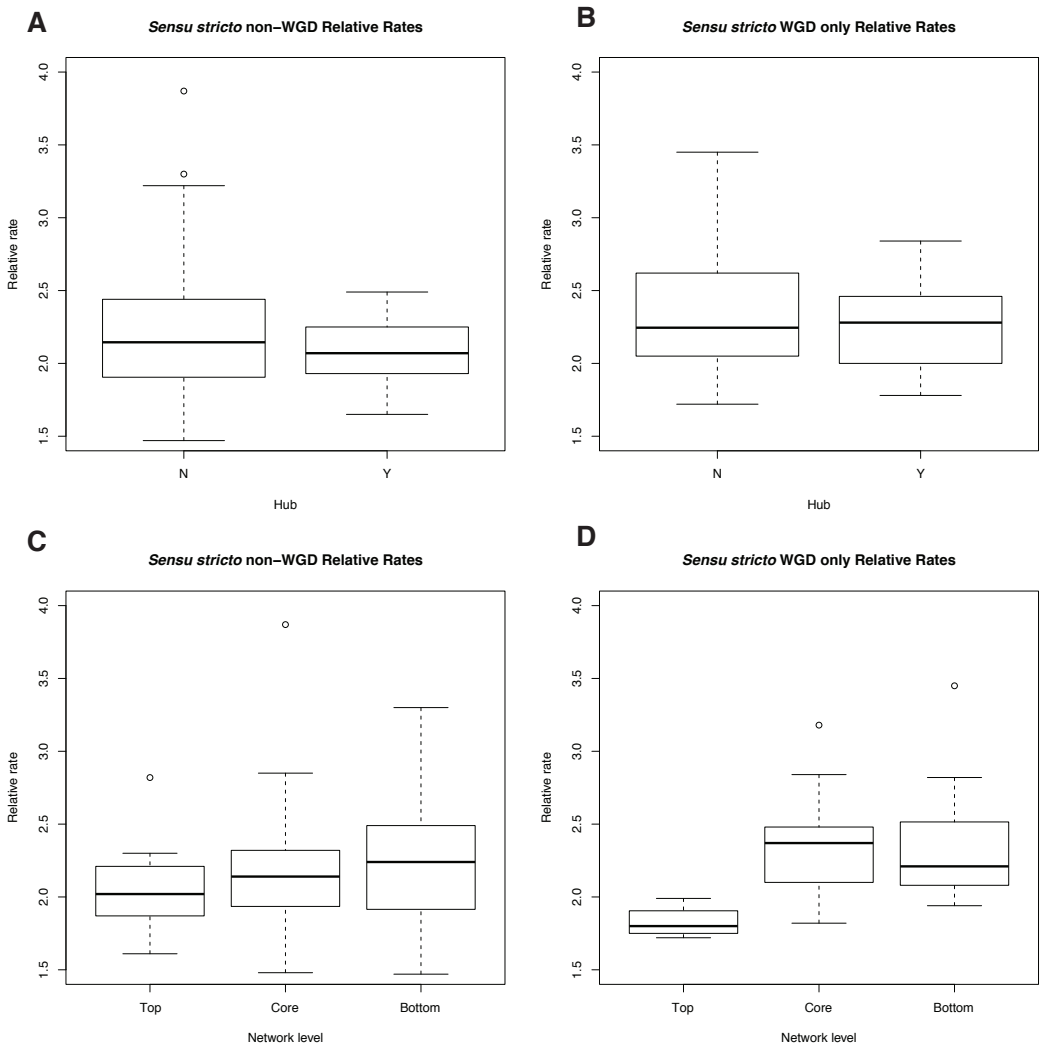


Figure 5.6: Relative rate estimates of WGD and non-WGD TFs in the context of the regulatory network. **A** and **B**: Rates of hub and non-regulators TFs in non-WGD and WGD TFs, respectively. **C** and **D**: Rate distribution across the hierarchical layers of the network for non-WGD and WGD TFs, respectively.

test; $\chi^2 = 2.4856, df = 2, P = 0.289$). This suggests two things; firstly, post-duplication divergence of regulators occurred mainly in TFs located further downstream in the regulatory cascades whereas there was stronger pressure to conserve the sequence and by extension, the function, of the top-layer TFs. Secondly, the signal for stronger conservation of top-layer TFs compared to other layers observed by me and others (Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009), might not hold when not taking into account the WGD and more data from other, non-WGD, species is needed to confirm whether this is indeed the case.

5.4 Different Evolutionary Rates of DNA-binding and Accessory Domains

Examination of the multiple sequence alignments of individual orthogroups and splits as well as numerous observations in individual TF families (e.g. *GAL4*, Zachariae *et al.*, 1993) indicated that DNA-binding domains (DBDs) tend to evolve slower than the non-DBD parts of TFs. In oder to see whether this was a general feature of TF evolution and whether there were any family-specific differences to this, I recalculated evolutionary rates using the method outlined above, this time separating alignments of each split into DBD and non-DBD regions. Thus for each split two scaling factors, a DBD and a non-DBD relative rate, were calculated. Furthermore, I wanted to reconsider the association between evolutionary rates of TFs in relation to their position in the network in light of the differences between DBD and non-DBD regions. I had found slower rates among TFs in the top layer and a positive association between indegree and evolutionary rate when considering the entire length of the protein. In order to determine whether this equally applied to DBD and non-DBD regions and to start being able to comment on where functional changes might be found, I considered DBD and non-DBD regions separately. As before, I will mainly concentrate on the *sensu stricto* species to ensure reliable interpretation of the network analyses although, with respect to the rate dichotomy between DBDs and other regions of the TFs, analysis of other clades gave equivalent results (discussed below).

Separating DBD and non-DBD regions reduced the number of alignment positions and as such the amount of phylogenetic signal available for estimation of the individual rate scaling factors. This was reflected in the number of splits that I managed to obtain a reliable estimate for (for filtering procedures, see above) which was reduced to a total of 182 splits compared to the 220 used in the rate analysis across all domains discussed above. Nevertheless, this still encompassed approximately 75% of splits with homologs in the *sensu stricto* species and as such should be a representative sample. To ensure consistency of the conclusions drawn from this subsample, I examined whether there were any biases with respect to the relative evolutionary rate, connectivity and membership in hierarchical levels between the 182 splits that yielded reliable between-domain ratios and the remaining 38 splits. I found no significant differences between relative rates, outdegree in both networks, hierarchical level and indegree in the YT network, although indegree distributions were significantly different in the Jothi2009 network (two-sided Wilcoxon rank-sum test; $W = 3531.5; P = 9.15 \times 10^{-4}$). Examination of the indegree distributions of each group gave indications that there was a larger proportion of cases with very large indegree among the splits that did not yield reliable between-domain rate ratios. This could essentially affect the positive association seen between indegree and evolutionary rate, as fast-evolving proteins will have been more difficult to align and as such not have provided sufficient phylogenetic signal for separate domain analyses and will be discussed in consideration of observed associations.

5.4.1 DNA-binding Domains Generally Evolve Slower than Non-DNA Binding Regions of Transcription Factors

Figure 5.7 shows the overall distribution of DBD (grey) and non-DBD (blue) relative rates in the *sensu stricto* species (Fig. 5.7A) as well as the relative rates between DBD and non-DBD regions on a per-split basis (Fig. 5.7B). I found DBD regions to evolve significantly slower overall (two-sided Wilcoxon rank-sum test; $W = 11488; P < 0.001$) confirming previous anecdotal observations. This was largely also true when examining results on a split-by-split basis, where I found DBD regions to evolve slower than non-DBD regions in approximately 72% of

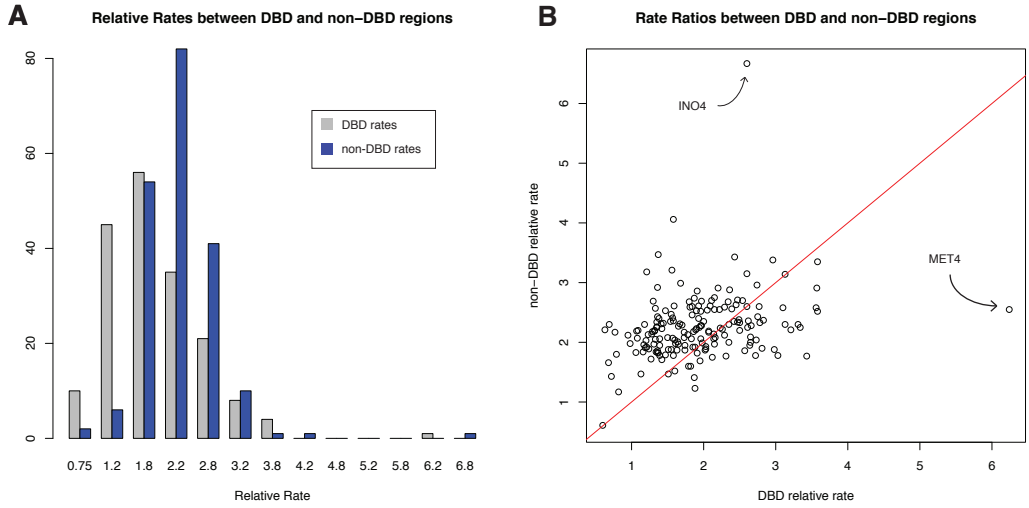


Figure 5.7: Relative rates in DNA-binding and non-DNA-binding regions of TFs in the *sensu stricto* clade. **A**: Overall distribution of DBD and non-DBD rates. **B**: Per-split ratio of DBD and non-DBD rates

splits (paired two-sided Wilcoxon rank-sum test; $V = 3784; P < 0.001$). This dichotomy in rates between DBD and non-DBD regions was even stronger when other clades encompassing larger phylogenetic distances were considered (Figure 5.8).

These results possibly reflect a relatively larger structural constraint imposed on the DBD compared to the remainder of the protein. To see whether this was dependent on the type of DBD, I also investigated whether I could see differences in such constraints between different families by considering DBD/non-DBD ratios for each family that had three or more estimates in each clade. While there were indeed significant differences between magnitude and directionality in the domain rate ratios between different families, I found little consistency with respect to which families contributed to those differences most across the different clades studied. The only noteworthy observation was the fact that the C_2H_2 and GATA zinc fingers were mostly found at the extreme lower end of the rate ratios, showing strong conservation of the DBD compared to non-DBD regions, suggesting that structural constraints on these relatively compact motifs are strong. As before I found no significant association between the percentage of positions in

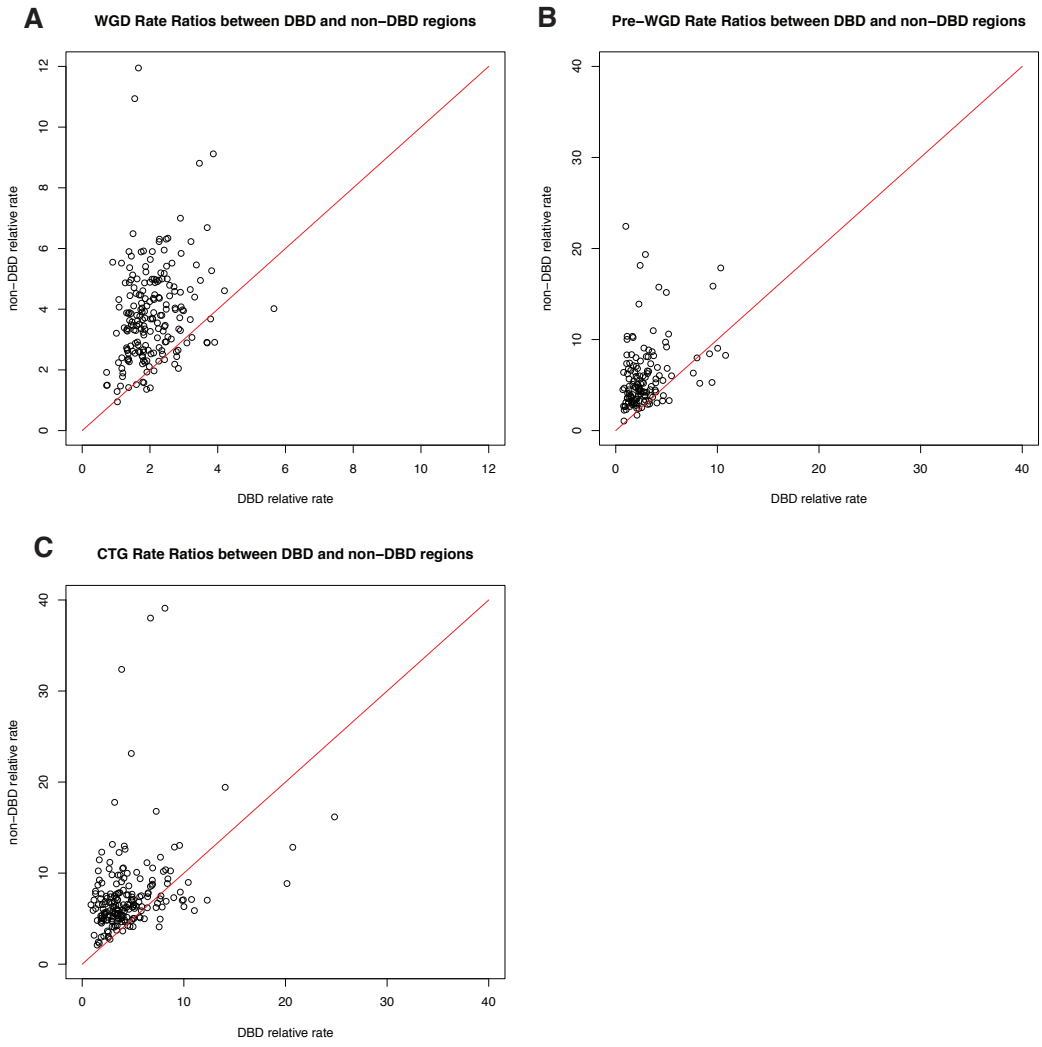


Figure 5.8: Relative rates in DNA-binding and non-DNA-binding regions of TFs in the (A) WGD, (B) pre-WGD and (C) CTG clades.

unordered regions in the *S. cerevisiae* TF and the domain rate ratio calculated for the respective split, indicating the lack of strong influence of intrinsic disorder on the dichotomous rate patterns. A more fine-grained analysis is necessary, however, to tease apart the contribution of other structurally constrained domains outside the DBD.

Consideration of individual atypical cases where I found the DBD to evolve much faster than the non-DBD revealed interesting insights with respect to possible functional changes. Figure 5.9 shows the alignment of the DBD of Met4 and its homologs, the split showing the most extreme DBD/non-DBD rate ratio in the *sensu stricto* species (see Fig. 5.7). Met4 contains a bZIP DBD, comprised of the basic region that directs DNA-binding and a leucine zipper further downstream that mediates interactions with other transcription factors. Met4, the main regulator of the sulfur metabolic network, lacks intrinsic DNA-binding ability but instead relies on co-factors for tethering to its target promoters (Lee *et al.*, 2010c). Based on the sequence alignment it appeared that DNA-binding ability was lost in the *Saccharomycetacea* because, in contrast to the homologs in the CTG clade, none of the homologs in those species encode a canonical basic region (Fig. 5.9). The requirement for specific co-factors, as well as the fact that Met4 is heavily post-transcriptionally regulated (Lee *et al.*, 2010c) might furthermore explain the decreased rate of evolution in the non-DBD regions of this TF. As such, Met4 provides an interesting example of the evolution of combinatorial gene regulation through the loss of DNA-binding ability and integration of information through different co-factors. I have found several lines of evidence suggesting changes in the regulation of sulfur metabolism, including rate shifts in the WGD duplicates Met31 and Met32, as well as Cbf1, the three main co-factors of Met4 (see Chapter 6, Table 6.3) and a more detailed discussion of this example will be provided in the final chapter of this thesis.

Besides the Met4 example, I examined the alignments of other splits showing a DBD vs. non-DBD rate of greater than 1.5. In all six cases examined, the increased DBD rate was due to a very short DBD region for which, despite highly conserved alignments, high relative rates were estimated likely due to the difficulty of accurate parameter estimation using such short stretches of DNA. In most cases, this was also coupled with a strongly-conserved non-DBD region,

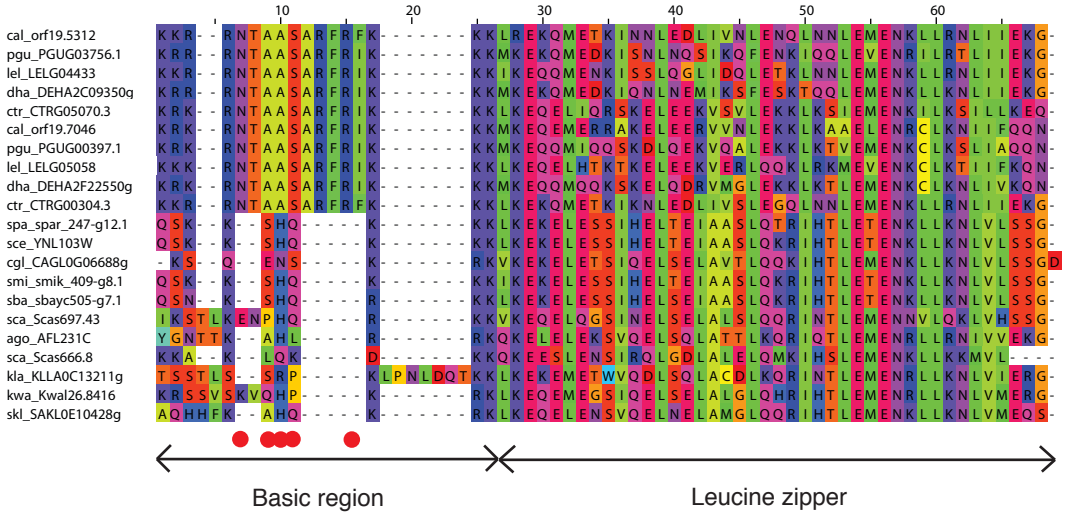


Figure 5.9: Multiple sequence alignment of the DNA-binding domain of Met4. Critical DNA-binding residues are marked using red dots.

hinting that relaxation in non-DBD regions rather than changes in evolutionary constraint in DBD regions might be responsible for the larger difference in relative constraint when larger evolutionary distances are considered (see Fig. 5.8).

5.4.2 Evolutionary Rates of Non-DNA-binding Regions Drive Associations Between Connectivity, Network Position and Rate

Evolutionary rates across all domains of TFs (see Section 5.2) confirmed two previously observed trends (Jovelin & Phillips, 2009; Wang *et al.*, 2010) with respect to a TF's position in the network and its evolutionary rate. Firstly, TFs in the top layer of the regulatory network evolved slower than ones residing in the core and bottom layers and secondly, there was a positive association between a TF's indegree and its evolutionary rate. I have previously shown that both those observations are influenced by the duplication status of a TF (i.e. whether it was recently duplicated) as well as its position within the hierarchical layers of the network. Furthermore, I have found evidence that DBDs generally evolve slower than non-DBDs and as such might be experiencing different evolutionary

	DBD	Non-DBD
Indegree (Jothi2009)	0.040 (P = 0.646)	0.241 (P = 0.003)
Indegree (YT)	-0.037 (P = 0.641)	0.226 (P = 0.002)
Outdegree (Jothi2009)	-0.090 (P = 0.328)	-0.169 (P = 0.053)
Outdegree (YT)	0.053 (P = 0.562)	0.032 (P = 0.715)

Table 5.2: Associations between TF connectivity and evolutionary rates of DBD and non-DBD regions of transcription factors. Values represent the correlation coefficient calculated using Spearman's rank correlation and the corresponding P-values

dynamics. To test whether there were any differences in how DBD and non-DBD regions contributed to the observed associations I repeated the analyses above with results as shown in Table 5.2.

This analysis revealed differences in the associations between evolutionary rates and between DBD and non-DBD parts of the TFs. While DBD regions showed no significant association with either indegree or outdegree, it appeared that the positive association of indegree and evolutionary rate was driven by the non-DBD regions of the TFs. Interestingly, in the Jothi2009 network, I also found a significant, but weak, negative association between rate and outdegree in non-DBD regions which probably reflects the tendency of regulatory hubs also to be interaction hubs (Wang *et al.*, 2010). Arguably, interaction hubs, especially if they are permanent, are likely to experience stronger constraints to conserve non-DBD regions as those are likely to mediate protein-protein interactions with different partners. This trend is not supported by the YT network however and its interpretation remains unclear. Similarly, the differences in rate between TFs in different hierarchical layers of the network was not significant considering DBDs only (Kruskal-Wallis test; df = 2; P = 0.779; Fig. 4.14A) but were significantly different in non-DBD regions (Kruskal-Wallis test; df = 2; P = 0.005), again driven by slower rates among TFs in the top level (Fig. 5.10B).

The non-DBD regions of TFs thus appear responsible for most of the associations that I had detected when analysing evolutionary rates within the context of the regulatory network. This was not very surprising, since DBDs are often structurally more constrained and changes in binding specificity are likely

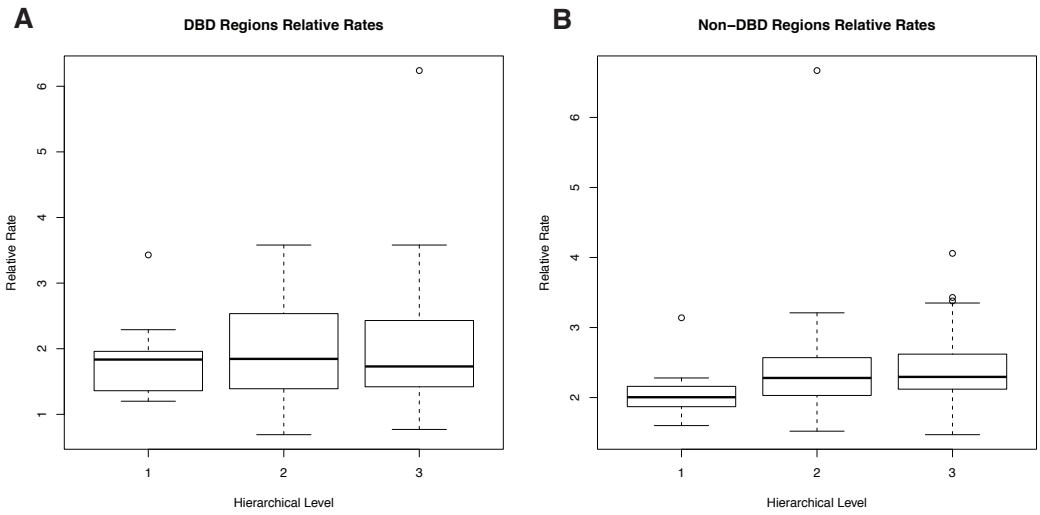


Figure 5.10: Evolutionary rates in hierarchical levels of the *S. cerevisiae* regulatory network for (A) DNA-binding domains and (B) non-DNA-binding domains.

to have larger pleiotropic effects. The fast divergence outside the DBD however also suggested ample raw material for adaptive changes through gain in interaction partners or post-translational regulatory sites such as phosphorylation or ubiquitination sites. Nevertheless, even if the overall rate of evolution across the DBD is low, single amino acid changes can alter DNA-binding specificity or even abolish DNA-binding all together (e.g. Ma & Ptashne, 1987) and in order to gain a quantification of changes contributing to differences in DNA-binding, detailed analysis on a family-by-family basis for each of those TFs is needed. Despite these limitations, the between-domain rate analysis highlighted interesting examples indicating changes in DNA-binding capability that are known to be of functional importance and provided hypotheses for further study which will be discussed in the last chapter of this thesis.

5.5 Conclusions

The nature of the evolutionary constraints acting on TFs have been the subject of several recent studies (e.g. Jovelin & Phillips, 2009; Wagner & Wright, 2007; Wang *et al.*, 2010). Conceivably, the more target genes a TF regulates (its outdegree)

and the more TFs it is regulated by (its indegree), the more important its role in the dissemination (outdegree) or integration (indegree) of signals and the larger a TFs pleiotropic constraint. In contrast to these assumptions however, none of the studies mentioned above have detected significant negative evolutionary rate constraints correlated with a TF's outdegree, and even more surprisingly, found positive associations between evolutionary rates and a TF's indegree (Jovelin & Phillips, 2009; Wang *et al.*, 2010). Wagner & Wright (2007) found a positive correlation between evolutionary rates and a TF's "dispensability" in terms of how many alternative pathways exist for the TF-gene pairs it connects. The mean number of pathways connecting any TF-gene pair in *S. cerevisiae* is 2.01 (Wagner & Wright, 2007) suggesting that redundancy is prevalent. This can be invoked as an argument for the lack of strong signal for outdegree-dependance of evolutionary rates, seeing that TF outdegree hubs are often found in the core layer of the regulatory network, which is highly interconnected and thus likely to include many redundant pathways (Jothi *et al.*, 2009). The positive relationship between indegree and rates however remained somewhat enigmatic even in the light of redundancy which would suggest that redundant intermediate regulators have higher indegrees which is not directly intuitive (but see below). Both Jovelin & Phillips (2009) and Wang *et al.* (2010) argued that positive selection might be responsible for this.

By distinguishing between the three functionally and architecturally different layers of the transcriptional regulatory network, I showed that the indegree-dependence observed here and elsewhere (see above) was strongest in the bottom layer of the network. This was true no matter which of the networks was considered. Indeed, the positive association between indegree and evolutionary rates was no longer statistically significant when the bottom layer TFs were excluded. Arguably the bottom layer of the regulatory network is also likely to contain a large number of intermediate redundant nodes seeing that TFs often participate in feed-forward loops and multiple input motifs that involve regulators from upstream layers (Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009). While again, this could be invoked to explain the overall higher evolutionary rates in bottom level TFs (Fig. 5.5B) it still did not account for the positive indegree-dependance.

When considering WGD and non-WGD TFs separately, I found that the positive association between indegree and evolutionary rates was stronger in the WGD TFs, indicating that post-duplication divergence and maybe positive selection might indeed have played a role in partly driving this trend as previously suggested. Again, my observation that hierarchical layers showed different levels of indegree-dependence held, although none of the trends was statistically significant in either the Jothi2009 or YT networks which is probably an effect of the small sample size among the WGD TFs when separated into layers. Another important difference between the relative rates of WGD and non-WGD TFs was the fact that the observed difference in evolutionary rates and conservation in different hierarchical layers was no longer significant when WGD TFs were removed from the dataset, suggesting that this signal was strongest just after duplication.

Overall, it appeared that the WGD had a considerable impact on the evolutionary rate correlates observed here and elsewhere (Bhardwaj *et al.*, 2010b; Jothi *et al.*, 2009; Jovelin & Phillips, 2009; Wang *et al.*, 2010). Given the overrepresentation of highly connected TFs among the WGD duplicates (see Chapter 4.5.1) and the fact that the WGD network was found to be tightly interlinked with the non-WGD network (Fusco *et al.*, 2010) it is conceivable that the overall evolutionary impact of the WGD on evolutionary rates extended further than the actual WGD duplicated TFs themselves, also affecting rates in TFs found upstream and downstream of WGD duplicates. Gene duplicates are by definition functionally redundant at first and redundancy has been shown to correlate positively with evolutionary rate (Wagner & Wright, 2007). Similarly, duplication of upstream TFs would lead to an initial doubling of regulatory inputs while redundancy would have been created upstream which in some cases could have lead to relaxed functional constraints on the TF in question thus driving a positive association of indegree and evolutionary rate. Another possible explanation would be the notion that TFs with high indegree are more highly-regulated themselves and thus perform specific roles. One would expect the selective constraint for a specific TF to be weaker than for a general, ubiquitous TF. In order to fully disentangle the determinants of evolutionary rates as well as the impact of the WGD one would need to consider the extended neighbourhood of a TF and the implied degree of redundancy. Other factors such as the amount of co-regulatory

and protein-protein interactions (although also of somewhat enigmatic character according to Wang *et al.*, 2010) and the dynamic properties of the regulatory network would also need to be considered. This analysis is not presented here due to time constraints but will be the subject of further study.

Separate analysis of DBD and non-DBD regions of TFs showed that DBD regions generally tended to evolve slower than non-DBD regions and that the rate associations described above were generally due to signal in the non-DBD regions. While without a structural model of each DBD and a resulting quantification of the impact of each mutation towards DNA-binding specificity it is difficult to assess the real relative amount of divergence in DBDs compared to their non-DBD parts, it was clear that non-DBD regions have high evolutionary plasticity and as such potential for the acquisition of new functions. Examination of TFs that showed opposite trends, e.g. very fast-evolving DBDs and slow-evolving non-DBDs, highlighted an example of DBD loss in the TF Met4 in the *Saccharomycetaceae* lineage, providing evidence for the evolution of combinatorial regulation of sulfur metabolism in those species which I have explored further in Chapter 7.4.

Chapter 6

Functional Signatures of Evolutionary Rates

6.1 Introduction

The evolutionary constraints acting on transcription factors (TFs) have largely been enigmatic so far. The number of target genes a TF regulates (its outdegree) and the number of TFs it is regulated by (its indegree) have proven to be a poor predictor of evolutionary rate (e.g. Evangelisti & Wagner, 2004; Jovelin & Phillips, 2009; Wang *et al.*, 2010). Some studies found redundancy, i.e. the number of alternative routes connecting its regulators and target genes (Wagner & Wright, 2007) to be positively associated with evolutionary rate and both Jovelin & Phillips (2009); Wang *et al.* (2010) have detected a positive correlation with evolutionary rate and its indegree. In my analyses of evolutionary rates, I have also detected this positive association with indegree and furthermore established that this association is dependent on the position of the TF within the hierarchical organisation of the regulatory network and its duplication status (see Chapter 5.3). Nevertheless, even when accounting for those, these correlates only explain a small percentage of the variation in evolutionary rate. In order to determine the impact of functional constraints on variation in evolutionary rates of TFs, I examined slow and fast-evolving TFs and differences in relative rates between orthologous TFs in each of the major clades studied here.

6.2 Clade-specific Variation in Evolutionary Rates

I used the relative evolutionary rate estimates calculated using the scaling approach discussed in Chapter 5.2.1 for clade-wise comparisons. Because the overall magnitudes of relative rates estimated and the number of splits that resulted in a reliable estimate for each clade were different (see Fig. 5.3), I ranked the different estimates of the scaling factor (c) within each clade from high to low and normalised the rank for each split by the total number of splits in the respective clades to obtain a comparable measure of the relative evolutionary rate of a TF within the remainder of the repertoire.

Relative rates between different clades were found to vary considerably. Figure 6.1 shows a clade-by-clade comparison between the rank of each split. As expected, I found most conservation between ranks in the *sensu stricto* and WGD species as a whole (top left plot), although even here I observed large amounts of variation between ranks for the same split. While other comparisons showed that there certainly was a biological component to this rate variation (the points falling on the diagonal are decreasing when considering more diverse clades), it was not clear how much variation one would expect due to small changes in rate variation across different genomic regions, alignment artefacts or other stochastic factors. I thus decided to consider only the extremes of the relative rate distributions (Figure 6.2). The fastest evolving 20% and slowest evolving 20% of splits in each clade were analysed in isolation, as well as in comparison to other clades. Splits were characterised as conserved if they fell into the top or bottom 20% of ranks in both clades (Fig. 6.2; blue) or divergent if they fell into the top 20% in one clade and the bottom 20% in another clade (Fig. 6.2; red). Figure 6.2 is an illustration of this classification based on a pairwise example. If more than two clades were compared, complete conservation of TF ranking in all of the clades considered was required for classification as conserved. Similarly, at least one ranking in the top 20% and one ranking in the bottom 20% of TFs was required for classification as divergent.

Besides manual inspection of results, I performed functional enrichment analyses on the downstream targets of TFs that fell into either slow or fast evolving

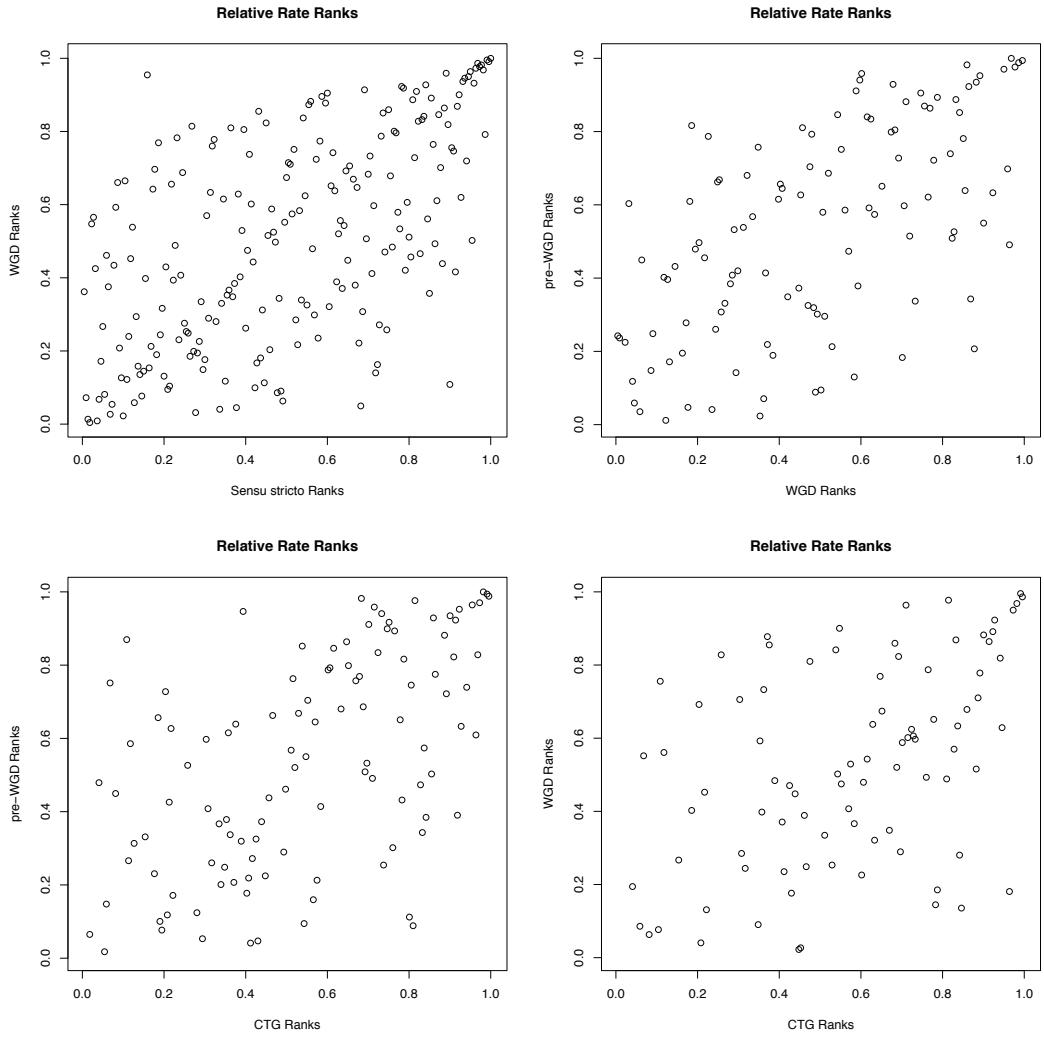


Figure 6.1: Clade-wise comparisons between the ranks of the relative rates for each split.

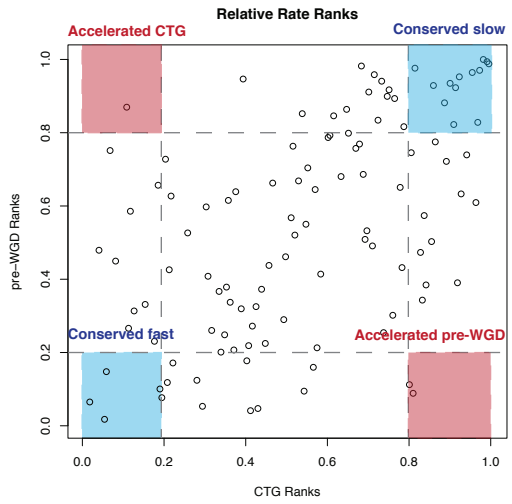


Figure 6.2: Categorisation of splits into rate categories. Rankings go from 0 (fastest evolving) to 1(slowest evolving). The top and bottom 20% of each distribution were classified as slow- or fast-evolving respectively. In comparisons across clades, splits were classified as conserved fast or conserved slow if they fell into the bottom or top 20% in all clades (blue) and as divergent if they fell into the top or bottom 20% in one clade but the opposite was true in others (red). When more than two clades are compared, TFs falling into the bottom or top 20% in *all* clades considered were classified as conserved fast or conserved slow.

categories to test for overrepresentation of particular functional classes of genes. GO term enrichment analysis was performed using Ontologizer 2.0 (Bauer *et al.*, 2008). P values were calculated using the Westfall-Young-Single-Step procedure and adjusted for multiple testing using the False Discovery Rate controlled at 5% (Benjamini & Hochberg, 1995).The GO annotations for all *S. cerevisiae* genes were obtained from the Saccharomyces Genome Database (SGD, 2010) and used as a background set. Study sets were created by combining all target genes of a given clade and category (slow/fast) using either the Jothi2009 or the YT data. Furthermore, I investigated overrepresentation of target genes in biochemical pathways using the SubpathwayMiner R package (Li *et al.*, 2009) which is based on information from the KEGG pathway databases (Kanehisa & Goto, 2000). As before, due to the uncertainty of how conserved the set of target genes is in other clades and due to the lack of reliable information about the regulatory network in the CTG species, I did not consider this clade in the enrichment analyses. Enrichments for the pre-WGD species were calculated based on *S. cerevisiae* target genes for the orthologous TFs, although these results have to be interpreted with caution.

6.3 Conservation of Evolutionary Rates Across Clades

Overall, I found little conservation of elevated evolutionary rates across all four clades. Only four orthogroups that had members in all four clades showed consistently fast rates of evolution. Those included *RME1*, a repressor of the main meiotic regulator *IME1* (reviewed in Vershon & Pierce, 2000). Indeed, meiosis is known to have changed drastically between the CTG species and the *Saccharomyces* clade. The CTG species lack a copy of *IME1* and other important genes required for meiotic recombination and chromosomal segregation (Butler *et al.*, 2009), suggesting that the elevated evolutionary rate of *RME1* could be driven by recent changes in the regulation of meiosis. Other consistently fast evolving TFs included *MGA1*, thought to be involved in the regulation of hyphal growth and the stress-response TFs *YAP3* and *XBP1*.

Conservation of very low evolutionary rates in contrast was more wide-spread with 10 splits showing consistently low rates in all clades. Four of those contained DNA-binding components of chromatin modification complexes. Furthermore, the slow-evolving splits included the general TF *NCB2*, the multiprotein binding factor *MBF1*, the general repressor *TUP1* as well as *HMO1* and *HAP1*. Intererestingly, *HMO1* has been shown to have diverged drastically in regulatory function, from a generalist TF in *C. albicans* where it was found to bind close to genes involved in mono- and polysaccharide metabolism, ergosterol metabolism and cell cycle regulation (Lavoie *et al.*, 2010). In *S. cerevisiae HMO1* mainly regulates ribosomal protein genes and binds to approximately a quarter of the number of targets as does its *C. albicans* ortholog and has thus become more specialist (Lavoie *et al.*, 2010). The same authors had detected a similar transition from a specialist to a generalist regulator for the TF *RAP1*, also involved in regulation of ribosomal proteins in *S. cerevisiae*, although here I observed a big shift in evolutionary rates, in line with the change in the numbers of promoters occupied (see next subsection). The rate conservation of *HMO1* across clades was thus surprising at first but might reflect the mechanisms of divergence and although *HMO1* binds fewer target genes in *S. cerevisiae*, it is part of a multi-TF co-regulatory complex and as such probably experiences strong evolutionary constraints.

In general, the conserved slow evolving TFs appeared to often be part of larger complexes or interacting with many different regulators and as such probably are subject to stronger evolutionary constraints. The overall lack of strong conservation of relative evolutionary rates between TFs across clades however also suggests that such constraints are either rare or can change within relatively short amounts of evolutionary time. Analysis of between-clade rankings of orthologs confirmed that these relative rate shifts occur frequently and I found good evidence for those being related to changes in functional constraints. These results will be discussed in detail in section 6.5 below.

6.4 Functional Signatures Among Slow and Fast Evolving Transcription Factors

In order to determine the adaptive significance of accelerated or decelerated rates of evolution, I examined the functional annotation of the downstream target genes regulated by the 20% fastest and 20% slowest evolving TFs in each clade. I performed a GO term enrichment analysis using the procedure outlined above as well as screening for enrichment of target genes among yeast biochemical pathways. Results from the GO term and pathway analyses are shown in Tables 6.1 (post-WGD) and 6.2 (pre-WGD) and 6.3 (post-WGD) and 6.4 (pre-WGD), respectively. Note that, especially in the pre-WGD species, these analyses are speculative at best and our understanding of the extent to which regulatory interactions are conserved between species is still very limited. The validity of the results obtained using target gene enrichments in the pre-WGD species are questionable and should be regarded as such.

GO term enrichment analysis of the target genes of fast and slow evolving TFs across the different clades revealed several interesting points. Generally, results obtained using the Jothi2009 and YT characterised target genes yielded similar results, although significant terms sometimes differed in granularity or process but where otherwise related, e.g. "oxygen and reactive oxygen species metabolic process" and "oxidoreductase activity" found enriched among the TGs of fast-evolving TFs in the Jothi2009 and YT TGs in the *sensu stricto* clade respectively. GO terms enriched in targets of fast and slow evolving TFs were similar between the *sensu stricto* and *WGD* clades (Table 6.1), as expected, with fast-evolving TFs involved in processes related to maintenance of an efficient anaerobic lifestyle, e.g. "oxygen and reactive oxygen species metabolic process", "oxidoreductase activity", "alcohol metabolic process" and "oxidation reduction", a distinguishing feature of the post-WGD species (Merico *et al.*, 2007, see Chapter 1.5.2). I furthermore detected significant enrichment for the terms "drug transport", "plasma membrane" and "fungal-type cell wall" which could be a signature for accelerated evolution in the regulation of the response to drugs and cell wall stress in those clades and was also reflected frequent rate shifts of TFs involved in these processes (see below) and the high representation of TFs involved in drug resistance

		Sensu stricto	WGD
Fast	Jothi2009	oxygen and reactive oxygen species metabolic process ($P_m = 0.58$)	plasma membrane ($P_m = 1.0$)
		drug transport ($P_m = 0.58$)	oxidation reduction ($P_m = 0.97$)
		—	transcription factor activity ($P_m = 0.87$)
		—	sexual reproduction ($P_m = 0.67$)
		—	vacuolar protein catabolic process ($P_m = 0.57$)
		—	ER-nucleus signaling pathway ($P_m = 0.56$)
	YT	plasma membrane ($P_m = 0.89$)	plasma membrane ($P_m = 1.0$)
		fungal-type cell wall ($P_m = 0.68$)	oxidation reduction ($P_m = 1.0$)
		alcohol metabolic process ($P_m = 0.59$)	cytosolic ribosome ($P_m = 1.0$)
		oxidoreductase activity ($P_m = 0.52$)	sequence-specific DNA binding ($P_m = 0.63$)
		—	fungal-type cell wall ($P_m = 0.62$)
		—	cellular amino acid and derivative metabolic process ($P_m = 0.52$)
Slow	Jothi2009	cellular amino acid biosynthetic process ($P_m = 0.99$)	cytosolic small ribosomal subunit ($P_m = 0.95$)
		cytokinetic cell separation ($P_m = 0.85$)	cellular amino acid biosynthetic process ($P_m = 0.91$)
		specific RNA polymerase II transcription factor activity ($P_m = 0.82$)	polysome ($P_m = 0.86$)
		cytosolic small ribosomal subunit ($P_m = 0.76$)	transcription repressor activity ($P_m = 0.54$)
		cytosolic large ribosomal subunit ($P_m = 0.76$)	telomere maintenance via recombination ($P_m = 0.51$)
		integral to plasma membrane ($P_m = 0.72$)	—
		iron ion binding ($P_m = 0.7$)	—
		fungal-type cell wall ($P_m = 0.53$)	—
		steroid biosynthetic process ($P_m = 0.52$)	—
		pheromone activity ($P_m = 0.50$)	—
	YT	cytosolic ribosome ($P_m = 1$)	plasma membrane ($P_m = 1.0$)
		cellular amino acid biosynthetic process ($P_m = 0.99$)	specific RNA polymerase II transcription factor activity ($P_m = 0.98$)
		fungal-type cell wall ($P_m = 0.84$)	cytosolic part ($P_m = 0.97$)
		organic acid transmembrane transporter activity ($P_m = 0.79$)	cellular amino acid biosynthetic process ($P_m = 0.91$)
		hexose catabolic process ($P_m = 0.57$)	iron ion binding ($P_m = 0.82$)
		intrinsic to plasma membrane ($P_m = 0.52$)	polysome ($P_m = 0.62$)
		—	cell death ($P_m = 0.53$)
		—	hexose catabolic process ($P_m = 0.53$)

Table 6.1: GO-term enrichment analysis for the fastest and slowest evolving splits in the post-WGD clades. P values are given as marginal posterior probabilities of a term being enriched in this given clade.

		pre-WGD
Fast	Jothi2009	transcription factor activity ($P_m = 1.0$) cellular aromatic compound metabolic process ($P_m = 0.99$) meiosis I ($P_m = 0.96$) plasma membrane enriched fraction ($P_m = 0.76$) mitotic recombination ($P_m = 0.73$) glutamine family amino acid metabolic process ($P_m = 0.56$)
	YT	response to temperature stimulus ($P_m = 0.98$) cysteine biosynthetic process ($P_m = 0.63$) purine base metabolic process ($P_m = 0.51$)
Slow	Jothi2009	NADP metabolic process ($P_m = 0.96$)
	YT	respiratory chain ($P_m = 0.95$) NADP metabolic process ($P_m = 0.6$) proton-transporting ATP synthase complex ($P_m = 0.5$)

Table 6.2: GO-term enrichment analysis for the fastest and slowest evolving splits in the pre-WGD clade. P values are given as marginal posterior probabilities of a term being enriched in this given clade.

and stress response that were retained after the WGD (Chapter 4.6). Interesting differences between the fast-evolving TFs in the *sensu stricto* and WGD clades were the enrichments for "transcription factor activity" and "sequence specific DNA-binding" among the WGD target genes only, which suggested accelerated rates of regulatory hubs just after the WGD. This in turn was another piece of evidence in support of large-scale changes in gene regulation after the WGD. Also uniquely enriched in WGD targets was the term "sexual reproduction" pointing towards changes in regulation of mating just after the WGD.

Target genes of slow-evolving TFs in the post-WGD clades were mainly enriched for genes involved in amino acid metabolism, telomere maintenance and ribosomal components, all essentially house keeping genes. Note that in addition to that, I found enrichment for terms that were also enriched among targets of the fast-evolving TFs, e.g. "plasma membrane", "fungal-type cell wall" and "cytosolic ribosome". Seeing that these terms are quite general, this suggested that while some components of the regulatory systems regulating those genes are evolving very slowly, other regulators are evolving very fast, probably reflecting functional divergence between different sets of conditions that these genes are

regulated under.

GO term enrichment in the pre-WGD species (Table 6.2), in contrast, looked very different. While here slowly evolving TFs were enriched for target genes participating in oxidative metabolism, e.g."respiratory chain" or "NADP metabolic process", the target genes of fast evolving TFs showed enrichment for genes involved in amino acid metabolism, meiosis and mitosis which could almost be interpreted as a reversal with respect to the post-WGD clades where I found TFs regulating housekeeping genes to evolve slowly and TFs involved in the regulation of oxidative metabolism to evolve fast. Interestingly, I also found a strong enrichment for "transcription factor activity" among the target genes of fast-evolving TFs indicating changes in the regulatory networks of those species. As mentioned before, the pre-WGD species are divergent from *S. cerevisiae* and it is unclear how conserved we expect the downstream targets of orthologous TFs to be across such long evolutionary distances, making interpretation of results more difficult. It is possible that this seeming reversal of classes of target genes under slow- and fast-evolving TFs is the result of extensive regulatory rewiring although this is probably unlikely. Alternatively, these differences could reflect variations in selective pressure on certain traits in those clades seeing that the life-style of pre- and post-WGD species is believed to be fairly different. Moreover, as shown by term enrichment among both slow and fast-evolving TFs among the post-WGD species, it appeared that increased selective pressure on the regulation of certain traits can potentially result in enrichment for target genes under the control of both slow-and fast-evolving TFs, suggesting that the regulatory system might gain new functions on one hand but experience very strong selective pressure to retain other parts of its functionality.

Pathway analysis (Tables 6.3 and 6.4) overall gave very similar results: target genes of fast-evolving TFs in the *sensu stricto* species were enriched for participation in metabolic pathways and biosynthesis of secondary metabolites (Table 6.3). Again, I found strong enrichment for genes involved in carbohydrate and oxidative metabolism among the target genes of TFs in the *sensu stricto* species and the WGD clade ($*$ in Table 6.3), reflecting some of the known changes in life-style in these species. As before, enrichments were found in target genes of both slow- and fast-evolving TFs, indicating fast divergence in subsets of the regulatory

		Sensu stricto	**WGD**
Fast	Jothi2009	Metabolic pathways ($P = 0.017$)	Oxidative phosphorylation ($P = 0.014$)*
		Biosynthesis of secondary metabolites (P = 0.017)	Sulfur metabolism (P = 0.015))
		—	Biosynthesis of secondary metabolites (P = 0.016)
		—	MAPK signaling pathway yeast (P = 0.017)
		—	Fructose and mannose metabolism (P = 0.017)*
		—	Starch and sucrose metabolism (P = 0.019)*
		—	Nitrogen metabolism (P = 0.019)
		—	Pyruvate metabolism (P = 0.031*)
		—	Meiosis yeast (P = 0.031)
	YT	Glycerophospholipid metabolism ($P = 0.021$)	Alanine, aspartate and glutamate metabolism ($P = 0.003$)
		Metabolic pathways ($P = 0.049$)	Methane metabolism ($P = 0.003$)
		Fructose and mannose metabolism ($P = 0.004$)*	Starch and sucrose metabolism ($P = 0.005$)*
		Nitrogen metabolism ($P = 0.005$)	Selenoamino acid metabolism ($P = 0.008$)
		Pyruvate metabolism ($P = 0.027$)*	Biosynthesis of secondary metabolites ($P = 0.010$)
		Starch and sucrose metabolism ($P = 0.029$)*	Glycerophospholipid metabolism ($P = 0.010$)
		Propanoate metabolism ($P = 0.032$)	—
		Tryptophan metabolism ($P = 0.032$)	—
		Amino sugar and nucleotide sugar metabolism ($P = 0.034$)	—
		Glycerolipid metabolism ($P = 0.039$)	—
		One carbon pool by folate ($P = 0.039$)	—
		Sulfur metabolism ($P = 0.039$)	—
		Galactose metabolism ($P = 0.041$)*	—
Slow	Jothi2009	Biosynthesis of secondary metabolites ($P = 0.000$)	Alanine, aspartate and glutamate metabolism ($P = 0.003$)
		Ribosome ($P = 0.000$)	Phenylalanine, tyrosine and tryptophan biosynthesis ($P = 0.003$)
		Alanine, aspartate and glutamate metabolism ($P = 0.001$)	Sulfur metabolism ($P = 0.004$)
		Arginine and proline metabolism ($P = 0.001$)	Glycine, serine and threonine metabolism ($P = 0.005$)
		Glycolysis / Gluconeogenesis ($P = 0.001$)*	Lysine biosynthesis ($P = 0.005$)
		Glycine, serine and threonine metabolism ($P = 0.003$)	Histidine metabolism ($P = 0.016$)
		Glyoxylate and dicarboxylate metabolism ($P = 0.009$)*	Cell cycle yeast ($P = 0.020$)
		Histidine metabolism ($P = 0.009$)	Citrate cycle (TCA cycle) ($P = 0.024$)*
		Pyruvate metabolism ($P = 0.013$)*	Valine, leucine and isoleucine degradation ($P = 0.027$)
		Citrate cycle (TCA cycle) ($P = 0.013$)*	Arginine and proline metabolism ($P = 0.042$)-
		Cysteine and methionine metabolism ($P = 0.017$)	—
		Lysine degradation ($P = 0.019$)	—
		Fatty acid metabolism ($P = 0.020$)	—
		Selenoamino acid metabolism ($P = 0.022$)	—
		Oxidative phosphorylation ($P = 0.025$)*	—
	YT	Ribosome ($P = 0.000$)	Ribosome ($P = 0.000$)
		Methane metabolism ($P = 0.005$)	Glycine, serine and threonine metabolism ($P = 0.005$)
		Glycine, serine and threonine metabolism ($P = 0.035$)	Lysine biosynthesis ($P = 0.028$)
		Starch and sucrose metabolism ($P = 0.046$)*	Pentose phosphate pathway ($P = 0.029$)
		Cell cycle yeast ($P = 0.046$)	Citrate cycle (TCA cycle) ($P = 0.039$)*
		Glycerolipid metabolism ($P = 0.047$)	Alanine, aspartate and glutamate metabolism ($P = 0.041$)
		Glyoxylate and dicarboxylate metabolism ($P = 0.047$)*	Galactose metabolism ($P = 0.041$)*

Table 6.3: Biochemical pathway enrichment analysis for the fastest and slowest evolving splits in the post-WGD clade. P values are corrected for multiple testing using FDR. Entries marked by ∗ are referred to in the main text.

		pre-WGD
Fast	Jothi2009	Thiamine metabolism (P = 0.017) Biosynthesis of secondary metabolites (P = 0.017) Galactose metabolism (P = 0.019) Sulfur metabolism (P = 0.019) Selenoamino acid metabolism (P = 0.025) Fructose and mannose metabolism (P = 0.048)
	YT	Fructose and mannose metabolism (P = 0.005) Galactose metabolism (P =0.005)
Slow	Jothi2009	—
	YT	—

Table 6.4: Biochemical pathway enrichment analysis for the fastest and slowest evolving splits in the pre-WGD clade. P values are corrected for multiple testing using FDR.

network controlling a set of functions while other parts of the regulatory network regulating the same functions were under strong selective constraint. Other enrichments among target genes of fast-evolving TFs in the WGD clade included sulfur metabolism for which I have found multiple lines of evidence indicating changes in the regulatory mechanisms controlling this, e.g. loss of DNA-binding ability in the master regulator *MET4* (Fig. 5.9), retention of *MET4* co-factors after the WGD and large difference in relative evolutionary rates of sulfur metabolic regulators in different clades (see below). The evolution of this system will be discussed in more detail in Chapter 7.4.

As before, enrichments among target genes of fast-evolving TFs in the pre-WGD clade (Table 6.4) were different to the post-WGD clade although here the differences were not as extensive. Again I found evidence for fast divergence in the regulation of amino acid metabolism. In addition to that however, I also detected enrichments for target genes involved in the utilisation of alternative carbon sources; galactose metabolism in particular. Galactose metabolism is known to have been rewired between *S. cerevisiae* and *C. albicans* (reviewed in Brown *et al.*, 2009; Askew *et al.*, 2009). Carbohydrate metabolism in *Kluyveromyces lactis*, a pre-WGD species, shares mechanisms of regulatory control of sugar utilisation with both *S. cerevisiae* and *C. albicans* (e.g. Zachariae *et al.*, 1993; reviewed

in Brown *et al.*, 2009), adding to the notion that divergence of carbohydrate metabolism has started before the divergence of the *Saccharomycetaceae*. Here, I found no significant enrichments for pathways among the targets of slow evolving TFs.

Overall, the analysis of functional enrichments among target genes of slow- and fast-evolving TFs provided interesting insights with respect to the evolutionary significance of relatively accelerated or decelerated rates of evolution among TFs and the link between TF divergence and known phenotypic changes. In the post-WGD species I found TFs evolving at the highest relative rates to be regulating genes involved in carbohydrate metabolism which is known to have been rewired extensively within the *Saccharomycotina* (e.g. Askew *et al.*, 2009; Brown *et al.*, 2009). I found another strong signature for divergence in the control of oxidative respiration, the importance of the effective regulation of which has been implicated in the evolution of an efficient fermentative life-style such as displayed by the post-WGD species (Merico *et al.*, 2007). It thus appears that the elevated evolutionary rates among those TFs might indeed be of functional and evolutionary importance. The lack of strong conservation of relative evolutionary rates between orthologous TFs in different clades, especially in fast-evolving TFs (see above), along with the differences in enrichments of target genes between clades, suggests that these accelerations are short lived. Only just over one third of the top 20% of fast-evolving TFs are shared between the 20% of TFs with the highest rates in the *sensu stricto* and WGD clades. I

Another interesting observation was the fact that on several occasions I found enrichment for certain processes or pathways among the target genes of both slow- and fast-evolving TFs in a given clade. This suggested that while parts of the regulatory network controlling a set of genes was evolving fast, other parts were under strong selective constraint. This might be due to effects of the whole-genome duplication, where it was found that paralogous copies of the same gene often experienced asymmetric rates of evolution, especially shortly after the WGD (Scannell & Wolfe, 2008). This is in line with my observations, where I found these "bi-partite" enrichments to some extent in the *sensu stricto* species but more strongly in the WGD species and not among the pre-WGD species, suggesting that this was a signal of post-duplication divergence, again underlining

the adaptive potential for gene duplication of TFs for the evolution of regulatory networks. Analysis of rate shifts confirmed this and will be discussed in the following section.

6.5 Between-clade and Between-paralog Rate Shifts Reflect Known Functional Divergence and Indicate Wide-Spread Evolutionary Divergence in Signalling Pathways

Broad functional analysis of downstream targets of extremely fast and slow evolving TFs suggested that changes in relative evolutionary rates between TFs may be indicative of changes in the gene regulatory programs they participate in and that those might be of adaptive significance (see above). In order to examine this more closely and determine whether I could find evidence for changes in selective pressure and by extension functional divergence between clades as well as paralogous TFs within clades, I grouped all ranked estimates of relative rates by their orthogroup (remember, relative rates were calculated for splits containing one-to-one orthologs and an orthogroup could result in several splits if gene duplication events were inferred within) and calculated the maximum difference in ranks between all possible pairs of splits and clades. I then selected OGs for which the maximum difference in ranks was larger than 0.6, according to the approach I had adopted for the identification of conserved or divergent rates of splits between clades (see Fig. 6.1). Table 6.5 shows all OGs for which I could detect such extreme rate shifts. Each OG was annotated with a "type" of rate shift, indicating whether it was found to be due to post-duplication divergence ("Dup"), between-clade divergence of one-to-one orthologs ("Clade") or both and the clade that showed the most divergent pattern compared to the rest of the tree.

Overall, I found 35 OGs that experienced rate shifts and more than half of those contained WGD-duplicated TFs. In total, 10 shifts were found to be purely due to post-duplication divergence, six showed shifts between duplicated paralogs but also between clades and the remaining 19 rate shifts were due to changes in

6.5 Between-clade and Between-paralog Rate Shifts Reflect Known Functional Divergence and Indicate Wide-Spread Evolutionary Divergence in Signalling Pathways

OG	Min. Rank	Max. Rank	Type	Clades	S. cerevisiae ID
129	0.9774	0.0452	Dup	CTG (+/-)	NDT80
21	0.9186	0.0091	Dup + Clade	WGD (+)	YAP1 / CAD1
139	0.9683	0.0591	Dup	*sensu stricto* (+)	MCM1 / ARG80
387	0.9276	0.0227	Dup	WGD (+)	SKN7/HMS2
150	0.9136	0.0409	Dup	*sensu stricto* (+)	MIG2/MIG3
24	0.9545	0.0947	Clade	*sensu stricto* + WGD (-)	HAC1
88	0.9955	0.1591	Dup	*sensu stricto* (+)	NHP06A /NHP06B
124	0.8462	0.0407	Dup + Clade	CTG (+/-)	DOT6/TOD6
12	0.9045	0.1086	Clade	CTG (+)	RAP1
102	0.9050	0.1124	Clade	Pre-WGD (+)	YOX1 / YHP1
148	0.9000	0.1086	Clade	*sensu stricto* (-)	NRG1 / NRG2
70	0.9638	0.1810	Clade	WGD (+)	CBF1
64	0.8273	0.0543	Dup + Clade	*sensu stricto* + pre-WGD (+/-)	GZF3 / DAL80
30	0.9864	0.2308	Dup	CTG (+)	HAP3 / HAP31
33	0.8818	0.1403	Clade	*sensu stricto* (-)	HAA1 / CUP2
266	0.7602	0.0318	Dup + Clade	*sensu stricto* + pre-WGD (+)	Unknown / RDS1
34	0.8455	0.1176	Clade	CTG (+)	MAC1
151	0.8100	0.0888	Clade	Pre-WGD (+)	MIG1
282	0.7227	0.0045	Clade	*sensu stricto* (-)	UME6
170	0.8136	0.1041	Dup	WGD (+/-)	NSF1 / RGM1
329	0.7964	0.0950	Dup + Clade	WGD + CTG (+)	HAL9 / TBS1
173	0.9095	0.2127	Clade	CTG (+)	MET31 / MET32
381	0.8773	0.1834	Clade	Pre-WGD (+)	MSN1
146	0.7515	0.0679	Clade	CTG (+)	RTS1
26	0.8639	0.1864	Clade	WGD (+)	BUR6
143	0.8778	0.2071	Clade	Pre-WGD (+)	TEC1
140	0.7692	0.0995	Clade	WGD (+)	STE12
138	0.7421	0.0773	Dup	*sensu stricto* (+/-)	RLM1 / SMP1
15	0.6968	0.0533	Dup + Clade	WGD (-)	CIN5 / YAP6
201	0.7828	0.1448	Clade	CTG (-)	RIM101
256	0.9502	0.3122	Dup	CTG (+)	Lysine biosynthesis TFs
344	0.7273	0.0909	Dup	*sensu stricto* (+/-)	PDR1 / PDR3
149	0.6818	0.0498	Clade	*sensu stricto* (-)	MSN2 / MSN4
347	0.8166	0.1855	Clade	*sensu stricto* + WGD (+)	CAT8
396	0.7376	0.1364	Clade	CTG (-)	TOS4

Table 6.5: List of orthogroups that have experienced a between-clade or between-paralog rate shift of greater than 0.6 in normalised ranks. Rows are coloured according to the dominant functional classes: blue, Stress response; orange: nutrient signalling; purple: Nutrient signalling and stress response. Signs in clade assignments indicate the relative directionality of rate changes (+, increase; -, decrease). Note that this does not imply absolute directionality compared to an ancestral state. (+/-) indicates highly asymmetric rates after gene duplication.

relative rates of one-to-one orthologs between different clades. Examination of the identity of individual TFs or TF pairs resulted in a wealth of striking associations between rate shifts and previously known rewirings, reinforcing the notion that such shifts were indeed indicative of changes in functional constraints. In addition to that I found ample evidence reinforcing the trends seen for the retention of the stress- and nutrient-response signalling regulators following the WGD (Chapter 4.6). In the following I will discuss a few of the better characterised examples as well as the broader evolutionary implications of my results. A more detailed account of individual examples will be given in the final chapter of this thesis.

6.5.1 Evolutionary Rate Signature of Rewiring of Ribosomal Protein Regulators

The availability of ribosomal proteins (RPs) is one of the strongest determinants of growth rate (reviewed in Lempiäinen & Shore, 2009), making the coordinated regulation of ribosomal protein expression and genes involved in energy homeostasis an important influence on growth rate in response to certain conditions. In *S. cerevisiae* their transcription is controlled in response to a number of intra- and extracellular signals (e.g. Gasch *et al.*, 2000) and this is known to be regulated by a complex of TFs acting in concert and in response to different stress and nutrient signals (e.g. Kasahara *et al.*, 2007; Rudra *et al.*, 2005). The TFs Rap1, Hmo1 and Fhl1 are generally associated with yeast RP promoters and in activating conditions recruit another TF, Ifh1, leading to high level expression of RPs(Martin *et al.*, 2004). In stressful conditions or under nutrient limitation, Ifh1 is released from the complex and replaced with the co-repressor Crf1, shutting down production of RPs (Martin *et al.*, 2004). Although the RP regulon is highly conserved across species, RP regulation in *C. alibicans* has been shown to be controlled by a different set of TFs and has been completely rewired between *C. albicans* and *S. cerevisiae* (Hogues *et al.*, 2008). Motif analyses furthermore suggested that RP regulation by the TFs Tbf1 and Cbf1, as found in *C. albicans*, represents the ancestral state and that rewiring occurred in the lineage leading to the *Saccharomycetacea* (Hogues *et al.*, 2008).

This rewiring has recently been characterised further, showing that occupancy of promoter regions by Rap1, Hmo1, Tbf1 and Cbf1 has changed drastically between *S. cerevisiae* and *C. albicans* (Lavoie *et al.*, 2010). The most prominent difference in promoter occupancy was found for Rap1, the main regulator of RP genes in *S. cerevisiae* which increased 10-fold compared to the number of bound regions in *C. albicans*. Furthermore Rap1, which was found to mainly bind in telomeric regions in *C. albicans*, has gained control over additional regulons in *S. cerevisiae* where it is known also to regulate the expression of glycolytic enzymes, silent mating type loci and telomeric genes, thereby integrating the control of expression of RPs with glucose metabolism, mating and telomere maintenance. Similarly Tbf1, the main regulator of RP genes in *C. albicans* was exclusively found at RP promoters in this species, whereas it was bound to a variety of genes in *S. cerevisiae*. Both these TFs have thus migrated from a specialist to a generalist role. Cbf1 was the only generalist RP TF that was detected to have considerable amounts of overlap between promoters bound in both species, showing a conserved role in the response to sulfur starvation and sulfur amino acid biogenesis and regulation of respiratory genes, although its association with RP genes was exclusively found in *C. albicans*.

In the rate analyses presented above I detected evidence for changes in functional constraints on several RP regulators, including extreme shifts in evolutionary rate between clades in two of the main RP regulators, Rap1 and Cbf1, as well as the paralog pair Dot6 and Tod6, two further TFs known to be involved in the regulation of ribosome biogenesis in response to resource limitation (Lippman & Broach, 2009). These rate shifts reflected the characterised functional changes in the TFs Rap1 and Cbf1. Rap1 was among the 20% of the fastest evolving TFs in the CTG clade but experienced a gain in evolutionary constraint to become one of the 10% of the most slowly evolving TFs in the *Saccharomycetacea*. This was in line with the acquisition of control over the RP and glycolytic enzyme regulons in this clade as well as the increased binding coverage in *S. cerevisiae* detected by Lavoie *et al.* (2010). Cbf1, I found to evolve fast in the *Saccharomycetacea* but be highly constrained in the CTG clade. Again, this rate constraint coincided with a role in the regulation of RPs and a greater number of genomic targets in the CTG clade, suggesting the presence of strong negative selection on

RP regulators. Cbf1 also plays a role in sulfur metabolism, another pathway for which I have found multiple lines of evidence throughout this study, suggesting its evolutionary divergence (see Section 7.4 for an in-depth discussion). The evolutionary relevance of the rate shift between Dot6 and Tod6 in the CTG clade is currently unclear and will not be discussed here, although it certainly is an interesting target for further study.

This demonstrates that rate shifts of orthologous TFs in different clades can indeed reflect changes in functional constraints and the relative importance of TFs in the regulatory network within the respective clades and as such provide strong hypotheses for further study. In fact, such rate shifts may be be an "underestimate" of the real extent of rewiring as shown by the RP regulator Hmo1, the TF with the second largest change in number of occupied promoters between *S. cerevisiae* and *C. albicans* (Lavoie *et al.*, 2010). As discussed above I found Hmo1 to be among the 10 TFs that were evolving consistently slowly across all clades. Interestingly here, although it has migrated from a generalist role in *C. albicans* to a specialist role in *S. cerevisiae* where it also underwent a four-fold reduction in genome occupancy (Lavoie *et al.*, 2010), I found no loss in selective constraint. This is probably due to the great importance of RP regulation.

6.5.2 The Evolution of Stress and Nutrient Signalling in the Post-WGD Species

Further examination of TFs that I had found to have undergone large shifts in relative evolutionary rates revealed several additional interesting examples that pointed towards evolutionary changes in a number of pathways. Among these is a rate acceleration of Arg80 in the *sensu stricto* clade, a paralog of the generalist TF Mcm1 that regulates arginine metabolic genes, as well as a Mcm1 co-factor Yox1, involved in cell cycle regulation, that was accelerated in the pre-WGD species. The Mcm1-cofactor regulons have been shown to have undergone rewiring within the ascomycetous yeasts (Tuch *et al.*, 2008a). Specifically Arg80, the paralog of Mcm1 appeared to have taken over Mcm1's role in the regulation of arginine metabolic genes leading to a loss of Mcm1 binding in the promoters of those genes, a subfunctionalisation which could explain the relaxed selective constraint

on Arg80 compared to its close homologs. Furthermore, I found relative rate changes in several regulators involved in meiosis and filamentous growth in the post-WGD species (Ume6, Tec1, Ste12). Rate changes in the CTG clade included a number of TFs that are potentially important for pathogenicity in these species such as the (also heavily-duplicated) homolog of Hal9 which works in osmotic stress response in *S. cerevisiae* (Ruiz & Ariño, 2007), one of the main regulators of drug resistance in *C. albicans*, Ndt80 (discussed in Chapter 3.6.4) or Rim101, which is involved in pH response (Davis *et al.*, 2000; Kullas *et al.*, 2007).

By far the most striking presence in my table of rate shifts, again, were TFs involved in stress response (Table 6.3; blue), nutrient signalling (orange), or both (purple). This reinforced results obtained from the analysis of WGD paralogs which first suggested extensive changes in signalling networks (discussed in Chapter 4.6). Upon exposure to a variety of stresses, *S. cerevisiae* mounts a largely non-specific expression response, termed the Environmental Stress Response (ESR) where hundreds of genes are induced or repressed (Gasch *et al.*, 2000). ESR-induced genes are enriched for genes arming the cell against nutrient scarcity by preparing for metabolism of alternative energy sources (e.g. carbohydrate metabolism, fatty acid metabolism, autophagy, membrane transport) and providing protection against extracellular (cell wall modification) and intracellular stresses both directly related (detoxification of reactive oxygen species, cellular redox reactions) and unrelated (protein folding and degradation, DNA damage repair) to aerobic metabolism. Repressed genes in turn include classes of genes required for fast growth such as ribosomal proteins, amino acid metabolism and nucleotide biosynthesis. More recent studies have shown that there is a strong correlation between groups of genes with expression correlated with growth rate and the ESR, where induced ESR genes are repressed under high growth rates and *vice versa*, providing a strong link between growth rate, energy metabolism and stress response (Brauer *et al.*, 2008).

The ability to efficiently ferment glucose and survive under strict anaerobic growth, such as observed in *S. cerevisiae*, has been postulated to require exactly such coordination between intra- and extracellular nutritional and stress signals to overcome the metabolical hurdles associated with anaerobic lifestyle. This requires for example the adjustment the redox balance in the absence of the electron

transfer chain and the provision of molecular oxygen for a number of biosynthetic pathways (reviewed in (Merico *et al.*, 2007)). *S. cerevisiae* and the other post-WGD species are unique in their ability to preferentially ferment glucose at high rates under aerobic conditions as well as strict anaerobic growth. Although some pre-WGD species were also found to be capable of strict anaerobic growth, they varied in their ability to do so and ethanol yield was generally lower (Merico *et al.*, 2007). Comparative analysis of ESR between different fungi showed that while both *S. cerevisiae* and the basal ascomycete *Shizosaccharomyces pombe* show great similarities in the ESR-induced and repressed genes, the regulatory network controlling this has been rewired (Gasch, 2007). Moreover, *C. albicans* shows no appreciable ESR under the same conditions although conserved *cis*-regulatory elements indicate that the ESR might be conserved to some extent but not activated in the same conditions (Gasch, 2007; Gasch *et al.*, 2004).

It is thus clear that the ESR and potentially stress- and nutrient-signalling in general have undergone major changes within the *Ascomycota*. This was strongly reflected in my evolutionary rate analysis, where I found almost half of the TFs that had undergone shifts to be among the most important players of a variety of stress- and nutrient-signalling pathways (Table 6.5). Rate shifts were often observed between WGD duplicates (e.g. Yap1 / Cad1, Mig2 / Mig3, Haa1 / Cup2 or Nrg1 / Nrg2) and together with the enrichment of signalling TFs found among the WGD duplicates this suggested extensive changes to stress- and nutrient-signalling after the WGD. Given the strong dependence of anaerobic metabolism on tight regulation of metabolic activity in response to changing conditions, this raised the possibility that the WGD might have provided the potential for the fine-tuning of those signalling mechanisms that lead to increased efficiency of anaerobic metabolism. Different signals could be integrated by the divergence of one of the paralogs in their regulation (transcriptional or post-transcriptional) whilst maintaining the same set of target genes or by transmitting the same signal to a different set of target genes through divergence in the DNA-binding specificity. The drastic differences in evolutionary rates between WGD paralogs in the WGD and *sensu stricto* clades, although certainly not unexpected (e.g. Scannell & Wolfe, 2008) provided further evidence that this might indeed be the case. Furthermore, rate changes between clades of one-to-one orthologous TFs

involved in a variety of stress-and nutrient signalling (e.g. Rim101, Mig1, Mac1, Cat8, Msn1 and Hac1) supported the idea that signalling network components between clades experience different selective pressures and might look globally very different. Closer examination of a number of examples found here suggested further mechanistic evidence on how combinatorial regulation and signal integration in stress- and-nutrient signalling pathways has evolved following the WGD and will be discussed in the following chapter.

6.6 Conclusions

Consideration of rate variation between orthologous TFs in different clades suggested extensive differences between the relative selective constraints on a TF when compared to the rest of the repertoire in the respective clade. Very few TFs were found to be evolving consistenly slowly, and even fewer consistently fast, in different clades. When examining the functional associations among the target genes that were regulated by slow and fast-evolving TFs in the post-WGD and pre-WGD clades, I found little overlap between the associations of the downstream targets of fast-evolving TFs in both those clades, but instead associations highlighted functional classes of genes that reflected known adaptations, at least in the post-WGD clade, such as involvement in oxidative metabolism, where regulation has been suggested to be important for the maintenance of the anaerobic lifestyle exhibited by post-WGD yeasts (Merico *et al.*, 2007). This argued for a functional component to these rate variations, suggesting an adaptive relevance of differences in selective constraints on regulatory pathways, given that the TF in question has retained its function. Furthermore, I found more functional enrichments among the fast-evolving WGD TFs compared to both the *sensu stricto* and pre-WGD TFs, suggesting that the WGD might have been followed by a period of relatively broad adaptive evolution across many parts of the regulatory network. Alternatively, this could have been a signal resulting from the *Saccharomyces castellii* and *Candida glabrata* regulatory networks, which I previously showed to have retained and lost very different sets of TF (Chapter 3.6.3) and therefore are also likely to experience different selective constraints.

By examining the orthologs that had undergone extreme shifts in evolutionary rates between clades and after duplication I detected several examples that were previously known to have undergone large-scale rewiring, acquired entirely different sets of target genes and changed in their hierarchical position within the network (e.g. Rap1, Lavoie *et al.*, 2010). This was clearly reflected in the magnitude and directionality of the rate shift, not only underlining the potential of this analysis to detect relevant functional changes in the regulatory networks studied, but also providing evidence for the presence of a functional signal in the evolutionary rates *per se* that I had failed to detect in the global rate analyses. It is unclear, whether in the case of Rap1 this change in rate was influenced by the gain of regulatory interactions or the fact that Rap1 was recruited to the ribosomal protein regulon, the regulation of which is of critical importance to the cell. This is further complicated by the fact that Hmo1, which also experienced a large turnover in the number of regulatory interactions did not experience such a rate shift. In the absence of experimental data, as here, these questions could potentially be addressed using motif analyses to try to quantify the extent of promoter binding of TFs in different clades and how this relates to their estimated relative rates. This will however not be discussed further here.

Similar to the large number of signalling TFs among the WGD duplicates, I found an impressive number of important regulators involved in stress- and nutrient-signalling that had undergone large shifts in relative evolutionary rates. These were not only confined to WGD-duplicated TFs but also included numerous one-to-one orthologs showing rate variation between clades, underlining previous hypotheses about the evolutionary impact of the WGD on stress- and nutrient-signalling pathways (Chapter 4.6) as well as their evolutionary plasticity in general. It is known that the regulation of stress response is among the most variable processes from the cell to the population level (reviewed in Thompson & Regev, 2009) so this was not an unexpected observation. The extent of it, however, was unexpected. The combined list of WGD duplicates and TFs that I found to have undergone rate shifts populates the top levels of transcriptional response of almost all signalling pathways in *S. cerevisiae*. This represents a great potential for fine-tuning and integration of novel signals and target genes into existing pathways (WGD duplicates) and is an indication that this has indeed

happened, reshaping parts of the regulatory network compared to other clades (rate shifts). A more detailed discussion of a selected number of examples showing potential mechanisms for the evolution of combinatorial regulation and signal integration through both mechanisms will be presented in the following chapter.

Chapter 7

Conclusions and Outlook

7.1 Phylogenomics Approaches for The Resolution of Species Trees

Phylogenomics approaches have become commonplace in the post-genomic era and their general use for resolving species phylogenies can only be expected to increase in light of the ever-growing availability of fully sequences genomes. These developments have been accompanied by the availability of phylogenomic analysis pipelines (e.g. Wu & Eisen, 2008) as well as new evolutionary models, aiming to incorporate increasing amounts of within-gene heterogeneity across sites and lineages (e.g. mixture models; Blanquart & Lartillot, 2008; Le *et al.*, 2008). Here, I have investigated the impact of between-gene heterogeneity, which conceivably will play an increasingly important role in phylogenomic datasets, with the possibility of the incorporation of thousands of genes (e.g. Marcet-Houben & Gabaldón, 2009) and has been shown to impact conclusions drawn in an analysis of the branching order at the root of the mammalian phylogeny (Nishihara *et al.*, 2007). While I found model choice to be an important determinant of the Maximum Likelihood (ML) tree recovered for the 18 species of ascomycetous yeasts in my study, partitioning by genes did not change the ML tree when appropriate evolutionary models were used. This indicated that in the case of the species studied here, within-gene heterogeneity, especially different evolutionary processes acting on different codon positions, resulted in stronger non-phylogenetic signal

than between-gene heterogeneity, possibly reflecting the small size of the yeast genomes. Besides providing a well-supported and trustworthy phylogeny of the 18 ascomycetous yeasts, the phylogenomics study highlighted several other interesting questions that will need to be addressed in the future when analysing genome-scale datasets.

Partitioning of the data by genes did in most cases result in significant increases in likelihood indicating that, despite not having an impact on the ML tree in my case, is important to consider in general due to the clear improvement in model fit. Whether or not partitioning by genes is the most appropriate way to partition however is unresolved and indeed recent studies have shown that "over-partitioning" can equally lead to the reconstruction of non-optimal trees (e.g. Li *et al.*, 2008; Ward *et al.*, 2010). The number of possible partitions increases rapidly with the size of the dataset, making it computationally prohibitive to compute likelihoods on all possible partitions of a genome-scale dataset (Li *et al.*, 2008). Most often genes have been grouped into partitions according to their evolutionary rate (e.g. Bevan *et al.*, 2007; Nishihara *et al.*, 2007) but other aspects such as local transition/transversion rates or nucleotide composition might also provide sources of between-gene heterogeneity that can result in model violation. In my dataset both the shape parameter of the Γ distribution used to model differences in rates across sites as well as the transition/transversion rate parameter showed large amounts of variability between genes (see Figure 2.5) and their relationship is more likely than not to be complex. It will thus be interesting to examine the influence of individual evolutionary process parameters on partitioning scheme and model fit and whether such conclusions hold across a wide range of phylogenomic problems or are specific to particular groups of species or ranges of divergence.

As important as using an appropriate model of evolution is the determination of which model most appropriately describes the dataset at hand. For ML analysis, this is classically determined using hierarchical Likelihood Ratio Tests (hLRTs) if models are nested or the AIC or BIC criterions which can also be used to test non-nested models (reviewed in Burnham & Anderson, 2004). Furthermore, there exist empirical approaches such as the Goldman-Cox (Goldman, 1993) and posterior predictive simulation (Bollback, 2002) tests, although in most

cases those are too computationally expensive to use. It is known that for example the BIC is relatively conservative and favours less parameter-rich models than the AIC (Burnham & Anderson, 2004; Weaklim, 1999) and that model testing approaches generally favour more complex models than empirical approaches (Ripplinger & Sullivan, 2010). This was also apparent in my amino acid analyses, where both LRT and AIC supported the use of partitioned models, but those were rejected by the BIC (see Chapter 2, Table 2.3). Although here this did not affect the ML topology, there are other cases where AIC and BIC disagreed on, for example, the number of partitions used which in turn affected the topology of the ML tree (Li *et al.*, 2008). As such model testing in phylogenomics datasets deserves a more thorough discussion, especially when partitioned models are used. For example here I chose to calculate the penalty terms for AIC and BIC on a per-partition basis based on the reasoning that additional parameters are introduced on a per-partition basis and are only relevant for the gain of information from individual partitions. This is not common practice for model testing in partitioned datasets however and it would be very interesting to empirically assess the effects of this on model choice and the inferred ML trees.

Finally, any tree inferred from phylogenetic analysis is a scientific hypothesis, providing the subject for further study. While I am confident that I have obtained a trustworthy species tree, there remain some controversies such as the branching order at the base of the clade that underwent a whole-genome duplication event (WGD) or the exact placement of *Pichia stipitis* for which bootstrap support was lower than for other nodes. These problems are likely to represent cases where ML phylogenetic analysis such as the ones performed here, even given a well-fitting model, does not provide sufficient resolution due to short divergence times between speciation events and the increased importance of population genetic processes such as independent lineage sorting in these cases (Degnan & Rosenberg, 2009). More recently, methods have been developed that consider individual gene evolutionary histories in the context of the coalescent to estimate the most likely speciation patterns. Those, especially the ones that do not rely on *a priory* gene tree inferences that are themselves likely to be subject of stochastic variation and model selection artefacts (see Chapter 2, Sections 2.3.1 and 2.3.2) such as the one proposed by Heled & Drummond (2010), will allow for the refinement

of the hypothesis I have presented here using alternative approaches. It will be interesting to investigate support for the branching order at the base of the WGD clade as well as the placement of *P. stipitis* using those methods.

7.2 Transcription Factor Repertoires in the *Saccharomycotina*

The role of transcription factors (TFs) in the evolution of regulatory pathways had until now remained largely unexplored, especially on a systematic level. Although some efforts have been made in characterising genome-wide repertoires of TFs (e.g. Charoensawan *et al.*, 2010b; Pérez-Rueda *et al.*, 2004; Riechmann *et al.*, 2000; Shelest, 2008), those have mainly focussed an fairly broad ranges of species and the identification of types of DNA-binding domains (DBDs) found in these repertoires. More recently, studies examining evolutionary rate constraints of TFs have emerged (e.g. Jovelin & Phillips, 2009; Wang *et al.*, 2010), but here the comparisons were based on very closely related species, lacking a broader context of TF repertoire evolution on an intermediate evolutionary range. None of the studies cited has attempted examination of the biological relevance of the trends they observed.

Here, I presented a comprehensive study of TF repertoires in 15 species of yeasts belonging to the *Saccharomycotina*. I have developed a pipeline to assemble TF repertoires from fungal proteomes and achieved high coverage using this approach. Compared to other, "off the shelf" datasets (e.g. available from the DBD database; Wilson *et al.*, 2008a), I achieved an increased coverage by as much as 30% (Chapter 3.2.2.1), highlighting the value of my *de novo* approach, despite the amount of manual curation that was required to obtain a high-confidence dataset.

I have characterised the families of TFs present in the genomes studied and their lineage- and species-specific patterns of presence and absence based on the type of DNA-binding domain (DBD) they encode. Furthermore, I have inferred

patterns of duplications and losses resulting in turnover of transcriptional regulators and analysed their evolutionary rates within and between different clades as well as in relation to their position in the regulatory network.

Overall, I have found that the composition of TF repertoires across the evolutionary range examined is highly asymmetric and globally similar with the same few DBD families contributing the majority of transcriptional regulators which reflects the results of existing studies (Chapter 3, Figs. 3.5 and 3.6). Genome-wide analysis of TF family copy number profiles allowed me to identify a number of lineage- and species-specific cases of gains and losses of both DBDs and domain architectures pointing towards biologically relevant pathways. Those included the multifunctional regulator *ABF1* which was gained in the *Saccharomycetaceae* (Chapter 3.6.1) and the carbohydrate metabolism regulator *SGT1* which was lost in the *sensu stricto* species (Chapter 3.6.2). The example of *ABF1* in particular, is one of the most highly-connected TFs in *S. cerevisiae* (Lee *et al.*, 2002), gave first indications of how fast newly gained regulatory proteins can become important, ubiquitous members of the regulatory network.

While the distributions of DBDs and domain architectures suggested no obvious expansion of any particular family after the WGD event, TF gain in the larger repertoires of the *Candida* species (the CTG clade; see Chapter 1, Fig. 1.9) appeared to be largely accommodated within the two largest families of TFs (Chapter 3, Fig. 3.6). This was confirmed by a detailed analysis of duplications and losses, showing that mechanisms of TF gain were different between the post-WGD species in the CTG clade (Chapter 4, Fig. 4.11). While regulatory network growth in the post-WGD species was mainly due to retention of WGD duplicates and non-family specific, TF gain in the *Candida* species was dominated by ongoing lineage-specific amplification of the $Zn(II)_2Cys_6$ and fungal-specific zinc finger DBDs, the two largest TF families. The species diverging after the split with the CTG clade but before the WGD appeared to be dominated by losses, although the analysis of such patterns is complicated by the quality of genome annotation and gene tree reconstruction artefacts (Chapter 4.3) and has not been explored further here.

Based on these mechanistic differences and the functional dichotomy between the duplicates retained after whole-genome duplication (WGD) and small-scale

duplication (SSD) observed for other classes of proteins, I examined the properties of WGD-retained TFs in the context of the regulatory network. When compared to their non-WGD counterparts, I found WGD TFs to have a significantly larger outdegree in one of the two networks tested and be significantly enriched for large indegree in both networks. Comparison of the proportions of WGD TFs among regulatory hubs and different hierarchical layers of the network was not significantly different from random expectation. This suggested the retention of a "cross-section" of the regulatory network that was somewhat biased for highly-connected TFs. This was in support of the dosage balance theory, predicting the retention of high-pleiotropy TFs following the WGD that would be expected to have deleterious, dosage-dependent effects in SSDs (see above; Chapter 4.5). It was not possible to empirically evaluate to what extent SSD duplicates share the same properties due to the small number of recent SSD TFs in the *Saccharomycetacea* and the unavailability of a fully-characterised regulatory network to compare recent SSDs in the CTG clade to. All six TFs that have arisen though recent SSD show small indegree and outdegree and do not contain regulatory hubs, leaving this an area of further investigation (see "Future Work" below).

Nevertheless, WGD TFs were clearly enriched for highly-connected TFs underlining the importance of the WGD for regulatory evolution in these species. This was also reflected in the functional signature of WGD TFs, over half of which are involved in stress- and nutrient-signalling and cell cycle regulation (Appendix, Table A.2). Coordination of all three of these processes is of critical importance to the metabolic patterns arising from the preferentially anaerobic life-style that distinguishes the post-WGD species from their relatives that did not undergo a WGD event Brauer *et al.* (2008); Gasch (2007); Merico *et al.* (2007). This over-representation of stress- and nutrient-responsive TFs was also reflected in my comparative analyses of evolutionary rates between orthologs in different clades, where I found extreme changes in evolutionary constraint between numerous TFs at the core of virtually all stress- and nutrient signalling pathways between the clades studied. Closer examination of individual examples (see below) suggested that the WGD event may have been catalytic for the integration of new signals and transmission of the same signal to different sets of target genes on a

very large scale, especially with respect to the metabolic caveats accompanying a "high-throughput" fermentative life style.

The difficulty to define the pleiotropic constraints of a TF was obvious not only from analysis of gene duplicates within the regulatory networks but also when considering evolutionary rate constraints. As others before (Jovelin & Phillips, 2009; Wang *et al.*, 2010), I have failed to find constraints on evolutionary rates depending on a TFs outdegree, arguably the most intuitive measure of pleiotropy. This was true even after accounting for the effects of recent gene duplication which could have potentially blurred such signal through relaxed selective constraints experienced by initially redundant functions (e.g. Byrne & Wolfe, 2007; Scannell & Wolfe, 2008). I also detected the previously found positive association between evolutionary rate and indegree and have furthermore shown that this association is confined to the bottom layer of the regulatory hierarchy. This most likely is the result of relaxed selective constraint in non-DBD parts of TFs and is partially driven by the WGD. The fact that this positive association is specific to bottom layer TFs might reflect functional diversification of specific biological processes, seeing that bottom layer TFs are often stand-alone and process-specific (Jothi *et al.*, 2009). Indeed, the functional signal in both the duplication and rate analyses was arguably the strongest among the correlates tested.

Returning to the original *cis* vs. *trans* debate (see above and Chapter 1.3) the results I have presented here as well as those published by others (e.g. Borneman *et al.*, 2007; Conant, 2010; Fusco *et al.*, 2010; Jovelin & Phillips, 2009; Lavoie *et al.*, 2010; Tuch *et al.*, 2008a; Wang *et al.*, 2010) strongly argue against the "evolutionary rigid" view of a TF advocated by the early evo-devo proponents. The lack of an appropriate way to describe the evolutionary constraints acting on TFs questions one of the strongest points of the arguments proposed, namely the high pleiotropic impact of changes in the coding sequences of TFs. Clearly, TFs are evolving fast and especially the non-DBD regions accumulate changes at a rate where sequences often become unalignable within a few hundred million years of evolution.

Overall this study has provided an in-depth, systematic view of evolutionary dynamics of TF repertoires at intermediate evolutionary distances. Careful data collection (Chapter 3.2) and choice of methods used for analysis (Chapter 4.3)

was crucial to ensure the high quality of the dataset itself and the inferences derived from it due to the promiscuous nature of domains commonly found in TFs and often limited phylogenetic signal. The dataset and analyses presented here have resulted in a larger number of hypotheses that can be addressed using experimental techniques and the insights to be gained are unlikely to be exhausted by what has been possible to achieve within the scope of this thesis.

7.2.1 Case Studies

One of the most distinguishing and best-studied metabolic features of the brewer's yeast *Saccharomyces cerevisiae* and its close relatives is it's ability to preferentially ferment glucose even under aerobic conditions. Although the overall yield of ATP per mole of glucose is lower in fermentation than through respiratory pathways, the rate of ATP production is higher, enabling the cells to grow faster (Rolland *et al.*, 2002). Furthermore, the ethanol produced during this reaction is exported from the cell and thought to confer a competitive advantage due to its toxicity to many other microorganisms. Yeast populations grow exponentially in glucose-rich conditions and, once all glucose in the medium is consumed, switch to aerobic metabolism of the ethanol produced during fermentation (Rolland *et al.*, 2002). This is called the diauxic shift.

Apart from the long-term metabolic cost, anaerobic fermentation comes with other limitations such as the dependence on the presence of molecular oxygen of some biosynthetic pathways, including sterol and fatty acid biosynthesis, heme, NAD or uracil (reviewed in Merico *et al.*, 2007). Also, the ability to adjust the redox balance and to avoid accumulation of genotoxic reactive oxygen species (ROS) is important for a successful oxygen-free lifestyle (Rigoulet *et al.*, 2004). Consequently, control of the cellular metabolism and response to nutrient availability requires tight control and the integration of intra- and extracellular nutritional and stress cues.

A recent comparative study of the energy metabolism and ability to grow under aerobic and anaerobic conditions and different nutritional requirements within the *Saccharomycotina* revealed extensive differences in metabolic efficiency under strict anaerobic conditions when minimal resources were provided among

this clade (Merico *et al.*, 2007). Most species were found to be able to grow on rich media when respiratory metabolism was impaired, suggesting that the ability to ferment is ancient and wide-spread. There was however a clear distinction in metabolic capability that appeared coincident with a whole genome duplication (WGD) event that is thought to have occurred about 100 million years ago. Post-WGD species were generally found to exhibit a preferred fermentative lifestyle, be able to grow in the absence of oxygen and generate respiratory-deficient petite mutants, lacking mitochondrial DNA (Merico *et al.*, 2007; see Chapter 1, Fig. 1.9). Pre-WGD species in turn varied in their ability to do any or either of those, suggesting a clear link between successful fermentative lifestyle and the WGD.

Indeed, others had previously implicated the WGD in the evolution of efficient fermentative metabolism. Hexose transporters and other glycolytic enzymes were found to be among the small percentage of genes that were retained after WGD, resulting in a potential overall increase of glycolytic flux (Conant & Wolfe, 2007). Other classes of genes that are known to have been preferentially retained include transcription factors (TFs), suggesting an important role for post-WGD regulatory adaptation (Conant & Wolfe, 2008). In my study of the evolutionary dynamics governing TF repertoires in the *Saccharomycotina* I found TFs retained after WGD to be highly enriched for well-connected regulatory. Moreover, besides general implications for the adaptive potential of this event for regulatory evolution, I found components of signalling pathways, including some of the key regulators of stress-response, MSN2/MSN4, SKN7, YAP1 and NRG1 to be highly represented among WGD paralogous TFs (Chapter 4.6). This was also reflected in an analysis of evolutionary rates which revealed frequent clade-specific accelerations as well as increase in negative selective pressure among TFs involved in nutrient- and stress-signalling, indicative of functional divergence and resulting changes in evolutionary pressure between clades (Chapter 6.5.2).

These observations suggested extensive change in nutrient- and stress-responsive signalling after the WGD. While often, information about the precise role of each of the WGD paralogous TFs with respect to each other is missing, I have investigated a number of the better studied examples and found striking evidence for the importance of the WGD in facilitating cross-talk between stress- and nutrient-

response signalling pathways which in turn is likely to have had a large impact on the ability to grow anaerobically. Two of those examples will be outlined below.

Furthermore, I found multiple lines of evidence suggesting changes in regulation of sulfur metabolism (Chapters 4.6, 5.4.1 and 6.5) and based on my observations will present a model for the evolution of combinatorial regulation of sulfur metabolism such as is seen in *S. cerevisiae* (Lee *et al.*, 2010c).

7.2.1.1 Two-way Control of Carbohydrate Metabolism Under Different Stress and Carbon Source Conditions

S. cerevisiae regulates carbohydrate metabolism and growth phases through a number of different pathways in response to carbon source availability and an array of other environmental cues (reviewed in Smets *et al.*, 2010). In the presence of glucose, genes required for the utilisation of alternative carbon sources, the Krebs cycle and oxidative phosphorylation and gluconeogenesis are repressed (Rolland *et al.*, 2002). This is mediated through the serine/threonine protein kinase Snf1 which is inactivated in the presence of glucose. Upon glucose depletion, Snf1 is activated and phosphorylates Mig1, the main repressor of this pathway, which is subsequently exported from the nucleus, relieving repression of its glucose-repressed target genes (Figure 7.1).

Besides the repression of genes redundant in glucose metabolism, *S. cerevisiae* also positively regulates transcription of genes required for glucose metabolism which is under control of Snf3/Rgt2 sensor (Rolland *et al.*, 2002). In glucose-rich conditions, Snf3/Rgt2 is activated, leading to the degradation of Mth1 and Std1 which are required for the expression of Rgt1, the main repressor of glucose induced genes (Fig. 7.1; Li & Johnston, 1997). Mig2, a homolog of Mig1, is repressed by Rgt1 in low-glucose conditions and upon Snf3/Rgt2-dependent glucose induction becomes active and acts as an additional repressor of glucose-repressed genes. Mig1 and Mig2 have been shown to be fully- or partially redundant for the repression of different subsets of genes, with Mig1 being the main regulator (Westholm *et al.*, 2008). Furthermore, the authors found functional differences between the genes that were repressed fully or partially redundant as well as Mig1-dependent. The latter class includes genes for uptake of alternative carbon

sources, while partially redundant genes were enriched for genes involved in their metabolism. Interestingly, the fully redundant class contained numerous phosphate metabolism genes suggesting a link between phosphate and carbohydrate metabolism.

Mig3 in turn, the WGD paralog, and as such "closer sibling" of Mig2, was not found to be involved in the regulation of glucose-repressed genes. Instead the only gene that was detected to be significantly upregulated in a *mig3* null mutant was *SIR2*, a histone deacetylase involved in maintaining genome integrity (Westholm *et al.*, 2008). Unlike its WGD paralog and in common with Mig1 however, Mig3 is directly regulated by the Snf1 kinase which in addition to glucose-dependent signalling has been implicated in the response to genotoxic stress (Dubacq *et al.*, 2004).

It thus appears, that here the WGD has facilitated the evolution of a combinatorial circuit by a) the migration of a TF from one control system, glucose repression, to another control system, glucose induction and b) the control of different target genes under the same signal. In the earlier scenario, this potentially allowed for a fine-tuning of the timing of expression for different sets of genes as supported by the fact that Mig1 and Mig2 are only partially redundant and genes in different "redundancy classes" have different functions (Westholm *et al.*, 2008). It is possible for example that the glucose repression pathway acts faster than the glucose induction pathway (it contains fewer reaction steps Smets *et al.*, 2010) and it is conceivable that the first set of genes to be repressed in the presence of glucose should be transporters of other carbon sources to avoid accumulation of unused compounds. Similarly, metabolising enzymes will be needed for longer to use up the remaining, already imported molecules. This is true equally for the reverse scenario, where upon carbon source shift from glucose to alternative sources, the first proteins required are transporters. A further aspect of this is the fact that Snf1 signalling is not only responsive to glucose but also to genotoxic stress (Dubacq *et al.*, 2004), thereby separating the glucose repression regulon into genes that are more or less important for stress response depending on whether they are regulated by Mig1 only, partially or fully redundant.

In the case of Mig3, less is known about its exact function. While it has not been implicated in the regulation of glucose repressed genes (Westholm *et al.*,

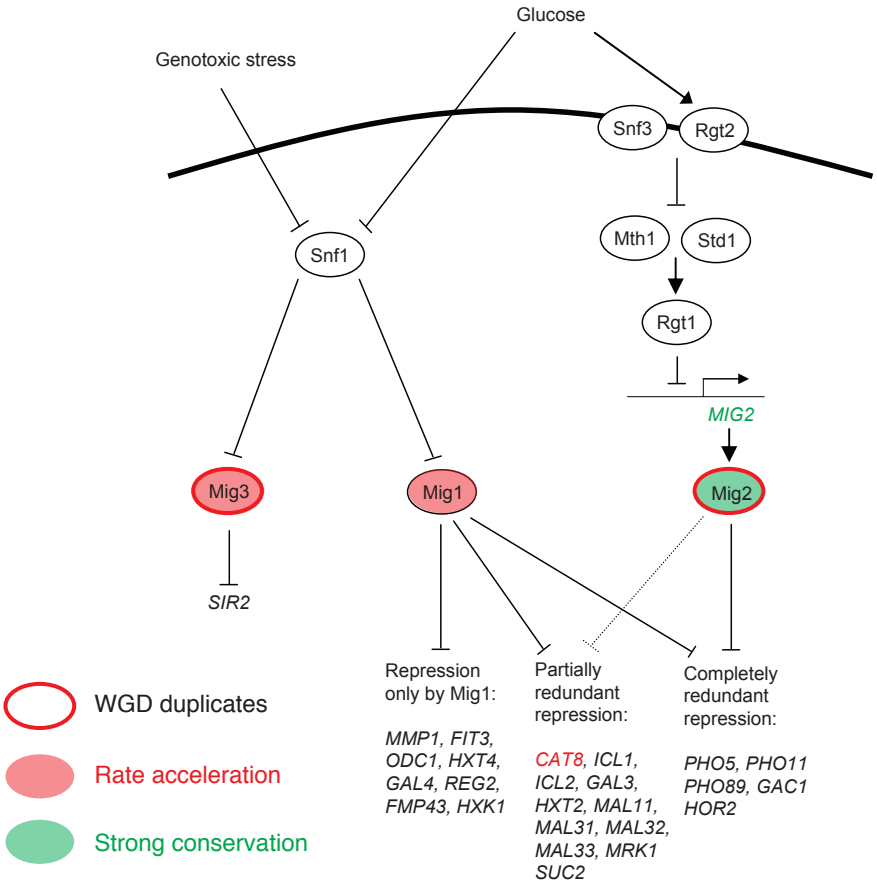


Figure 7.1: Regulation of glucose metabolism in *S. cerevisiae*. Figure adapted from Westholm *et al.*, 2008.

2008), it remained under control of Snf1 signalling (Dubacq *et al.*, 2004) and as such probably serves in integrating information about the nutritional status and genotoxic stress level to influence yet other metabolic processes. Sir2, its only known target gene, is known the have roles in silencing of the HMR and HML loci (silent mating type casesstes; Rine & Herskowitz, 1987) as well as telomeric and rRNA silencing (reviewed in Rusche *et al.*, 2003). Again, this provides interesting hypotheses about possible functional implications, for example a possible influence of nutrient availability or genotoxic stress on mating-type switching

Further evidence for changes in functional divergence among Mig1, Mig2 and Mig3 is based on an analysis of evolutionary rates. I found all three TFs to show great clade-specific rate variation. While the ancestor of Mig2 and Mig3 is among the 10% of most rapidly evolving TFs in the pre-WGD species, I found Mig2 among the 10% of slowest evolving TFs in the *sensu stricto* species indicating a gain of strong functional constraint. This is in contrast to Mig3 which I found to be still among the most rapidly evolving TFs. Similarly, Mig1 is among the most slowly evolving TFs in the CTG clade, whilst having experienced great acceleration in evolutionary rates in both the post-WGD but especially the pre-WGD species. Indeed, glucose repression by Mig1 in *Candida albicans* is thought to be under control of Rgt1 but not Snf1, analogous to to Mig2 in *S. cerevisiae* (reviewed in Sabina & Brown, 2009). *C. albicans* also contains a Mig2 homolog which I found to evolve neither greatly accelerated nor under extensive selective constraint in the CTG clade and it's functional role is still unknown (Sabina & Brown, 2009). CaMig1 and CaMig2 are highly conserved in their DNA-binding domain suggesting greater redundancy between the two TFs in the CTG species.

7.2.1.2 Evidence for Evolution of a Post-WGD Feedback Mechanism to Sense Glycolytic State

Copper and other metal compounds are of critical importance to an organisms viability by providing cofactors for many fundamental metabolic processes (Rutherford & Bird, 2004). Copper ions are found in different redox states in the cell and are important catalytic cofactors for a variety of redox chemical processes, including the detoxification of free radicals and respiratory metabolism (Peña

et al., 1999; Rutherford & Bird, 2004). An excess of copper in the cell however, has severe cytotoxic effects through the generation of ROS, inducing severe cell damage, or the displacement from other cofactors from key cellular signalling proteins (Peña *et al.*, 1999). This coupled essentiality and toxicity hence requires tight regulation of copper levels within the cell.

S. cerevisiae uses two main pathways to control copper homeostasis (Figure 7.2). Copper-uptake is mediated by the TF Mac1, which activates high affinity copper transporters and other genes in response to copper deficiency. Mac1 binds do copper-responsive elements as a homodimer and is directly regulated by Cu(I) ions which bind to a polycysteine cluster at it's N-terminal end. This precludes both homodimerisation and DNA binding, thus rendering Mac1 inactive when sufficient concentrations of copper are present in the cell (reviewed in Rutherford & Bird, 2004).

Cup2 (also known as Ace1) is active in high concentrations of copper and regulates proteins involved in copper sequestration and protection from toxicity. Similarly to Mac1, It is directly regulated by copper concentration through binding of copper ions in a N-terminal domain "polycopper cluster". In contrast to Mac1 however, Cup2 requires bound copper ions for its DNA-binding and thus regulatory activity (reviewed in Rutherford & Bird, 2004). Interestingly, the WGD paralog of Cup2, Haa1, although well-conserved in both DNA-binding domain and transactivation domain is not regulated by copper levels (Keller *et al.*, 2001). Instead it was recently found to be active in response to weak acid stress (Rutherford & Bird, 2004) which is in line with the observation by Keller *et al.* who detected upregulation of genes involved in cell wall maintenance.

Again, the WGD paralogs have greatly diverged in function and are now under control of different signalling mechanisms. Whether or not Haa1 and Cup2 regulate shared target genes remains to be determined, although the strong conservation within the DNA-binding domain first detected by (Keller *et al.*, 2001) but also apparent in my *Saccharomycotina*-wide multiple sequence alignments argues for similarity between their binding specificities. The response of Haa1 to weak acid stress is specific to the less lipophylic acids including acetic acid and succinic acid and is indeed strongest for acetic acid (Fernandes *et al.*, 2005).

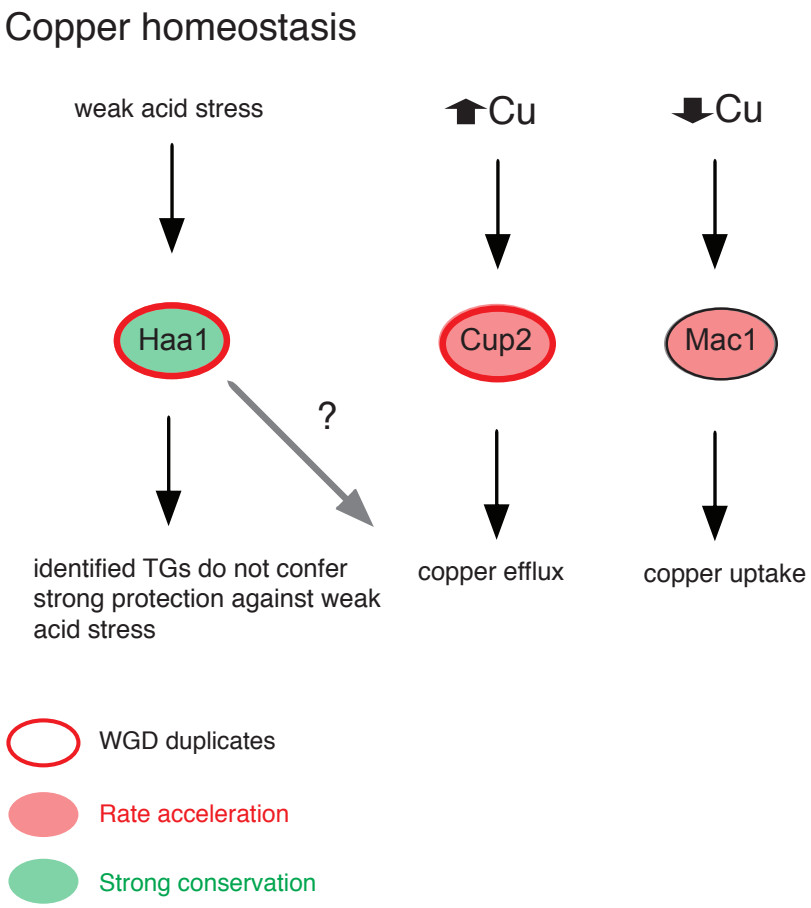


Figure 7.2: Regulation of copper homestasis in *S. cerevisiae*

Interestingly, however, while Haa1 was induced in weak acid stress, the transcriptional response induced only conferred limited protection against weak acid stress suggesting an alternative functional role.

If Cup2 and Haa1 indeed regulate shared target genes this WGD-generated duplication might present a mechanism for the evolution of cross-talk between two stress responses, high copper levels and weak acid stress, conferring cross-protection against both stresses. Moreover, the high affinity of Haa1-mediated stress response for acetic acid opens other interesting perspectives. Both acetic acid and succinate are byproducts of both glyceropyruvic and alcoholic fermentation that accumulate mostly at the beginning of fermentation, most likely due to overflow of intermediate products of an incomplete TCA cycle (reviewed in Vilela-Moura *et al.*, 2010). As such the response of Haa1 could function in signalling entry into fermentative metabolism. Depending on the nature of the hypothetical set of shared target genes between Haa1 and Cup2, this could represent a signal amplification seeing that copper ions are no longer needed for the purpose of respiratory metabolism which might also serve as a sensor for entry into fermentation. So Haa1 and Cup2 could either transmit this signal to a set of target genes unrelated to weak acid response and copper homeostasis but is important for fermentative metabolism, or Haa1 could aid induction of copper sequestering proteins to avoid copper accumulation and creating of ROS, or indeed both, upon entry into a fermentative cycle.

As in the previous example, I found extensive clade-specific differences in evolutionary rates for both Haa1 and Cup2 as well as Mac1 indicating changes in functional constraint. Both Haa1 and Cup2, and their pre-duplication homolog were found to evolve fast in the pre-WGD and CTG clades as well as immediately after the whole-genome duplication while experiencing a large decrease in rate in the *sensu stricto* species arguing for a gain in functionality. Similarly, Mac1 was found to be evolving very fast in the CTG clade only.

7.2.1.3 Evolution of Combinatorial Regulation of Sulfur Metabolism in *S. cerevisiae*

The sulfur metabolic network is a multibranched regulatory network, controlling sulfur assimilation in the absence of organic sources of sulfur, the synthesis of sulfur-containing amino acids, glutathione (an important antioxidant) and other metabolic compounds important for the biosynthesis of polyamines, vitamins and modified nucleotides (reviewed in Lee *et al.*, 2010c). As such, various aspects of sulfur metabolism contribute to a number of different cellular metabolic processes. Sulfur metabolism in *S. cerevisiae* is controlled by a heteromeric complex at the heart of which is Met4, the master regulator of sulfur metabolism in *S. cerevisiae* (Lee *et al.*, 2010c). Met4 lacks intrinsic DNA-binding ability but instead relies on a number of cofactors for recruitment to target promoters (Blaiseau & Thomas, 1998). The cofactors, Cbf1, Met28, Met31 and Met32 in turn lack the intrinsic ability to activate transcription of the sulfur metabolic network but require Met4's transactivating ability for induction of target genes (Blaiseau & Thomas, 1998; Blaiseau *et al.*, 1997; Kuras *et al.*, 1996).

Met4 associates with Cbf1, either of Met31 and Met32, or both Cbf1 and Met31/32, to specifically induce different sets of downstream target genes in a condition-specific manner (Fig. 7.3; Lee *et al.*, 2010c). Depending on the promoter structure (e.g. Cbf1 only, Cbf1-Met31/32 or Met31/32 only) and the strength of the Met31/32 binding site, different functional categories of sulfur metabolic genes are expressed and the timing and activation strength differs according to promoter architecture and hence the affinity and constitution of the associated Met4-containing complex. Met28 acts as a stabilising factor in both complexes. While Cbf1 is constitutively bound to promoters, the activity of Met4, Met28, Met31 and Met32 is regulated at various levels, transcriptionally and posttranscriptionally thus allowing for combinatorial control of different subsets of sulfur metabolism under a variety of signals (Lee *et al.*, 2010c). Formation, concentration and association with DNA target sites of each of these different complexes will not only depend on the presence or absence of each of the factors involved, but likely also on the concentration of their conformational states rendering them more or less likely to interact with each other, DNA and the transcriptional

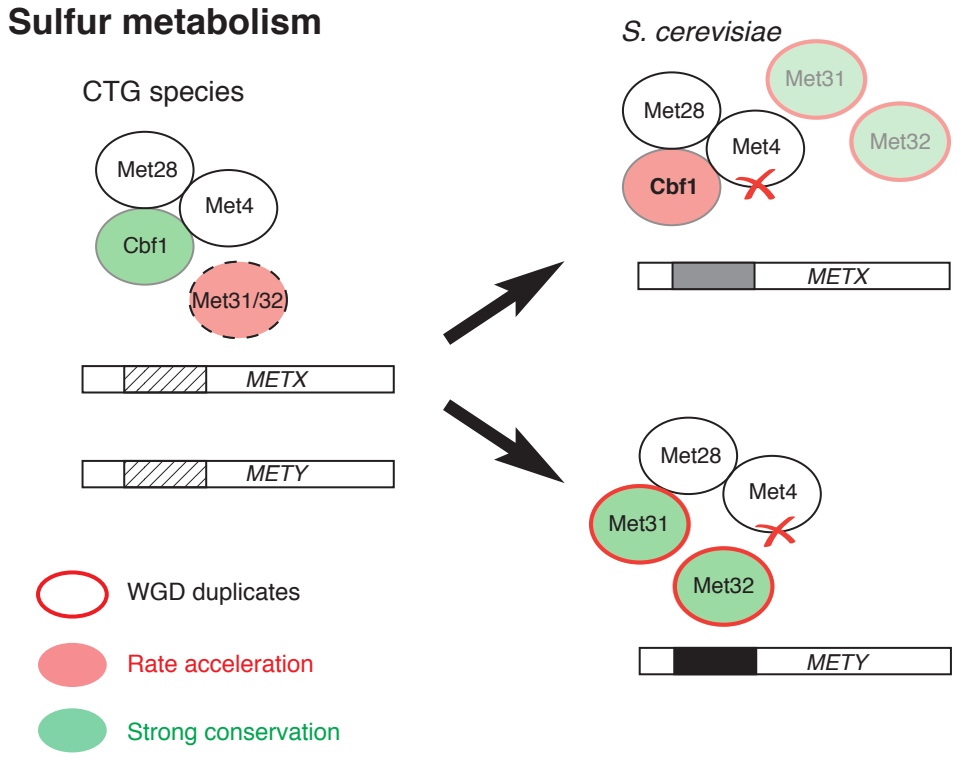


Figure 7.3: The evolution of combinatorial regulatory mechanisms of sulfur metabolism through loss of a DBD in Met4 and gain of new interaction partners.

machinery (see Buchler & Louis, 2008 and Chapter 1.2 for further discussion). Combinatorial control is thereby able to dynamically tune regulatory output in response to small environmental changes.

My evolutionary analysis of TF repertoires has provided several lines of evidence suggesting changes in regulatory mechanisms controlling sulfur metabolism within the *Saccharomycotina*. The most important of those changes is possibly the loss of DNA-binding ability of Met4 in the *Saccharomycetaceae* (see Chapter 5, Fig. 5.9). While both Met4 homologs in the CTG clade have a canonical basic region, required for DNA binding in bZIP TFs, only the leucine zipper dimerisation region has been maintained in the *Saccharomycetaceae*, thus render-

ing Met4-induced regulation entirely dependent on its cofactors. The association of Met4, Cbf1 and Met28 with promoter regions of sulfur metabolic genes is conserved within the *Saccharomycotina* (Lavoie *et al.*, 2010), arguing for the ancestral presence of a combinatorial mechanism that has become "hard-wired" through the loss of DNA-binding ability of Met4. It is unknown whether the single-copy homolog of Met31 and Met32, which I found to be WGD duplicates, has a conserved role in sulfur metabolism in the CTG and pre-WGD species or represent a novel combinatorial interactions that have arisen within the *Saccharomycetaceae*. Nevertheless, there is good evidence that this duplication provides another example where WGD retention resulted in the potential fine-tuning of an existing regulatory circuit. Although the distribution of Met31 and Met32 at promoter regions does not appear to be very different on a large scale, Met32 is bound in higher quantities than Met31, despite similar levels of protein abundance, suggesting the possibility of functional differences between the two (Lee *et al.*, 2010c).

Furthermore, I detected changes in the relative rates of Cbf1 and Met31 and Met32, compared to their orthologs in the CTG clade. I found Cbf1, which is also involved in the regulation of ribosomal proteins in *C. albicans* (Lavoie *et al.*, 2010; Tuch *et al.*, 2008a; see Chapter 6.5.1), to be among the 10% of most slowly evolving TFs in the CTG clade. This was in contrast to the post-WGD species where Cbf1 was among the 20% of fastest evolving TFs. Similarly, relative rates of Met31 and Met32 in the post-WGD clade were extremely slow while there appeared to be less constraint on the single ortholog in the CTG clade. Arguably, the relaxation of selective constraint on Cbf1 in the post-WGD species will have been greatly influenced by the loss of control over the ribosomal protein regulon. Nevertheless, it is also possible that the "handover" of a subset of sulfur metabolic genes to Met31/32 might have contributed to this relaxation. This handover could also explain the shift towards strong evolutionary rate constraints on Met31 and Met32 in the post-WGD species. Interesting here is that both copies appear to be heavily constrained, further arguing for their non-redundant functions.

7.2.1.4 Conclusions

The examples provided here show how a mixture of the three different scenarios introduced in Chapter 1.4.2 can contribute to regulatory novelty. I found evidence for changes in the regulatory mechanisms that these TFs are involved in, based accelerated and decelerated rates of sequences evolution. In all three cases, gene duplication was likely to have played an important part by allowing for fine-tuning of existing combinatiorial mechanisms (Section 7.4) or sub-or neofunctionalisation through integration of the same signal to different target genes (Section 7.2) or, putatively, different signals to affect the expression of shared target genes (Section 7.3). Most importantly, the examples presented here are three out of a very long list of duplicated TFs (Appendix A.2 and A.3) and ones that have experienced extreme rate changes between different clades (Chapter 6, Table 6.5). Systematic evaluation of the remaining examples should not only provide many more case studies such as the ones above but also feed into a bigger picture of the evolution of signalling pathways in the *Saccharomyces* species and the connections between them.

7.2.2 Future Work

In search for a more accurate description of the evolutionary constraints experienced by TFs, the most obvious extension to the work presented here is the refinement of network-based hypotheses. Regulatory networks are not static but change in a condition-specific manner (e.g. Luscombe *et al.*, 2004). TFs that are hubs in one condition are not necessarily as highly-connected in other conditions and arguably pleiotropic constraints will be dependent on i) how many conditions a TF is expressed in and ii) how many target genes it regulates in each of those conditions. Furthermore, although a TF might have a large number of regulatory interactions, a large proportion of those could be "redundant", i.e. achievable via other pathways, and as such not be of great importance. Conversely, a regulator with a relatively small number of interactions might have very large pleiotropic constraints if it is the only TF regulating these target genes. The importance of an individual regulatory interaction could be scored by it's redundancy factor,

i.e. the number of alternative regulatory pathways connecting a TF's regulators and the target gene in question and as such a redundancy-corrected outdegree could be obtained for comparisons. Again, those could be calculated on either a static representation of the network or in a dynamic fashion.

Although here I found no statistically significant differences between the distributions of WGD TFs among regulatory hubs or different hierarchical layers of the network, they were significantly enriched for regulators with large numbers of regulatory interactions and the trends were suggestive of enrichments among hubs in the core layer of the regulatory network. In order to explore differences between TFs duplicated during WGD and SSD, systematic profiling of recently duplicated TFs in the CTG clade could provide interesting insights into the properties of SSDs. Determination of their target genes through ChIP analyses would not only help to determine their functional importance but also give insights into their relative positioning in the regulatory network. Arguably, connectivity information about a few TFs is less useful for these purposes without a fully characterised regulatory network but should provide a start to formulate further hypotheses. In any case, it would be of great use to have a second fully characterised eukaryotic regulatory network to be able to understand network evolution and test hypotheses regarding different mechanisms. Finally, losses in the regulatory network are certainly no less interesting to study than duplications and a careful reexamination of my dataset combined with data mining approaches to delineate a set of "high-confidence" losses should also provide interesting insights into their impact on regulatory evolution.

Besides these larger evolutionary properties of regulatory network evolution, my study has highlighted a number of interesting targets to study in more detail in inter-species comparisons. Those include stress- and nutrient-response signalling and cell cycle regulation and could be addressed in a manner similar to the comparative studies examining the evolution of ribosomal protein regulation and the Mcm1-cofactor regulons as done by Lavoie *et al.* (2010) and Tuch *et al.* (2008a), respectively. This will not only help to better understand how regulatory evolution has influenced adaptation to the metabolic difficulties brought about by a fermentative life-style in the *Saccharomyces* species but should also provide insights into how the pathogenic species in the CTG clade face the challenge

of the human host and maybe deliver new targets for treatment and control of fungal infections.

Appendix A

Appendix

A.1 DBD pipeline

A.2 Speciation-Duplication Inference

A.2.1 Comparison of SDI methods

A.2.2 CAFE output

Figures A.3 to A.5 show the raw output of CAFE for gene families that were found to evolve significantly different from the estimated birth and death model (see Chapter 4.4.2.3). λ values and branch lengths for each branch were removed for the purpose of visualisation (those are equal between the different trees). Each node in the species tree is annotated with the inferred number of copies and the P-values for observing a transition in copy number from the inferred copy number to the ones inferred at each of the child nodes.

A.2.3 TF duplications

Species	DBD pipeline	DBD-DB	Shared	DBD pipeline only	DBD-DB only	Total
Ashbya gossypii	186	135	124	62 (31.47%)	11 (5.58%)	197
Candida glabrata	214	163	148	66 (28.82%)	15 (6.55%)	229
Kluyveromyces lactis	201	155	140	61(28.24%)	15 (6.94%)	216
Candida tropicalis	234	183	165	69 (27.38%)	18 (7.14%)	252
Pichia guilliermondii	254	201	181	73 (26.64%)	20 (7.3%)	274
Lodderomyces elongisporus	217	157	140	77 (32.91%)	17 (7.26%)	234

Table A.1: Comparison of the collected TF repertoires (DBD pipeline) with the automatically generated TF repertoires in DBD-DB (Wilson *et al.*, 2008a).

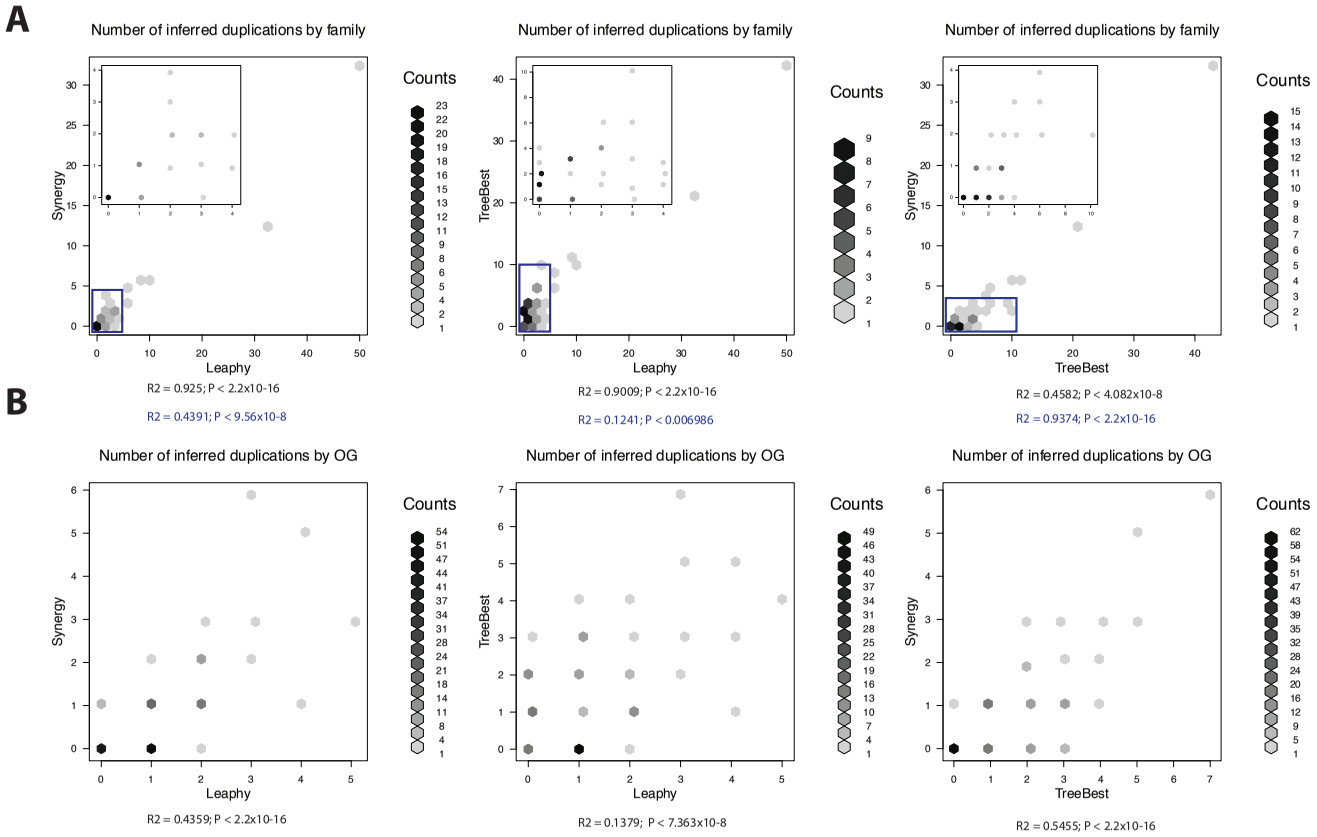


Figure A.1: Three-way comparison between the number of inferred number of duplications inferred by Leaphy$_N$, TreeBeST and SYNERGY when grouped by (A) DBD family and (B) orthogroup (OG). The sections boxed in blue delineate the close-ups shown within the respective plots. Hexagons are shaded based on the number of underlying data points. The correlation statistics shown beneath each plot were calculated on either the entire dataset (black) or the subset shown in the close-up (blue).

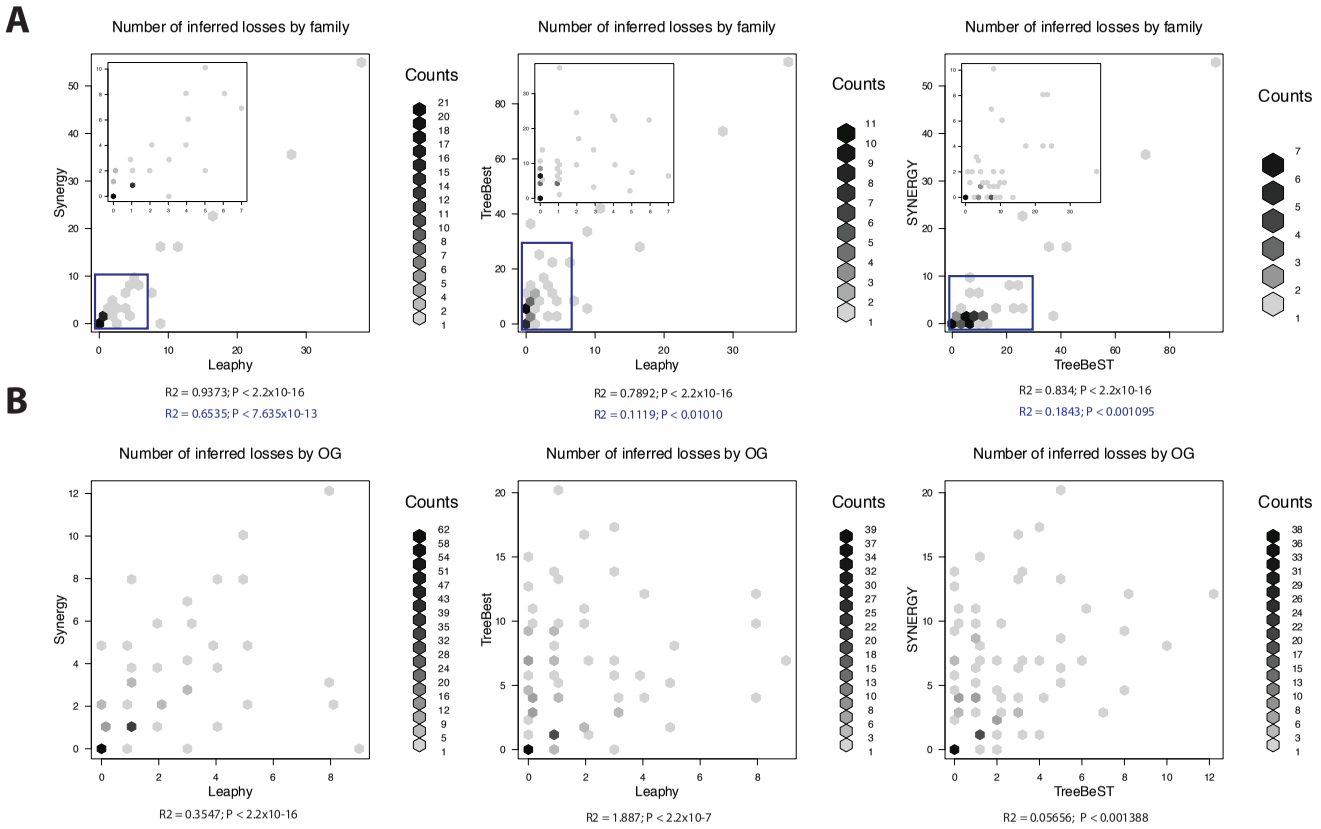


Figure A.2: Three-way comparison between the number of inferred number of losses inferred by Leaphy$_N$, TreeBeST and SYNERGY when grouped by (A) DBD family and (B) orthogroup (OG). The sections boxed in blue delineate the close-ups shown within the respective plots. Hexagons are shaded based on the number of underlying data points. The correlation statistics shown beneath each plot were calculated on either the entire dataset (black) or the subset shown in the close-up (blue).

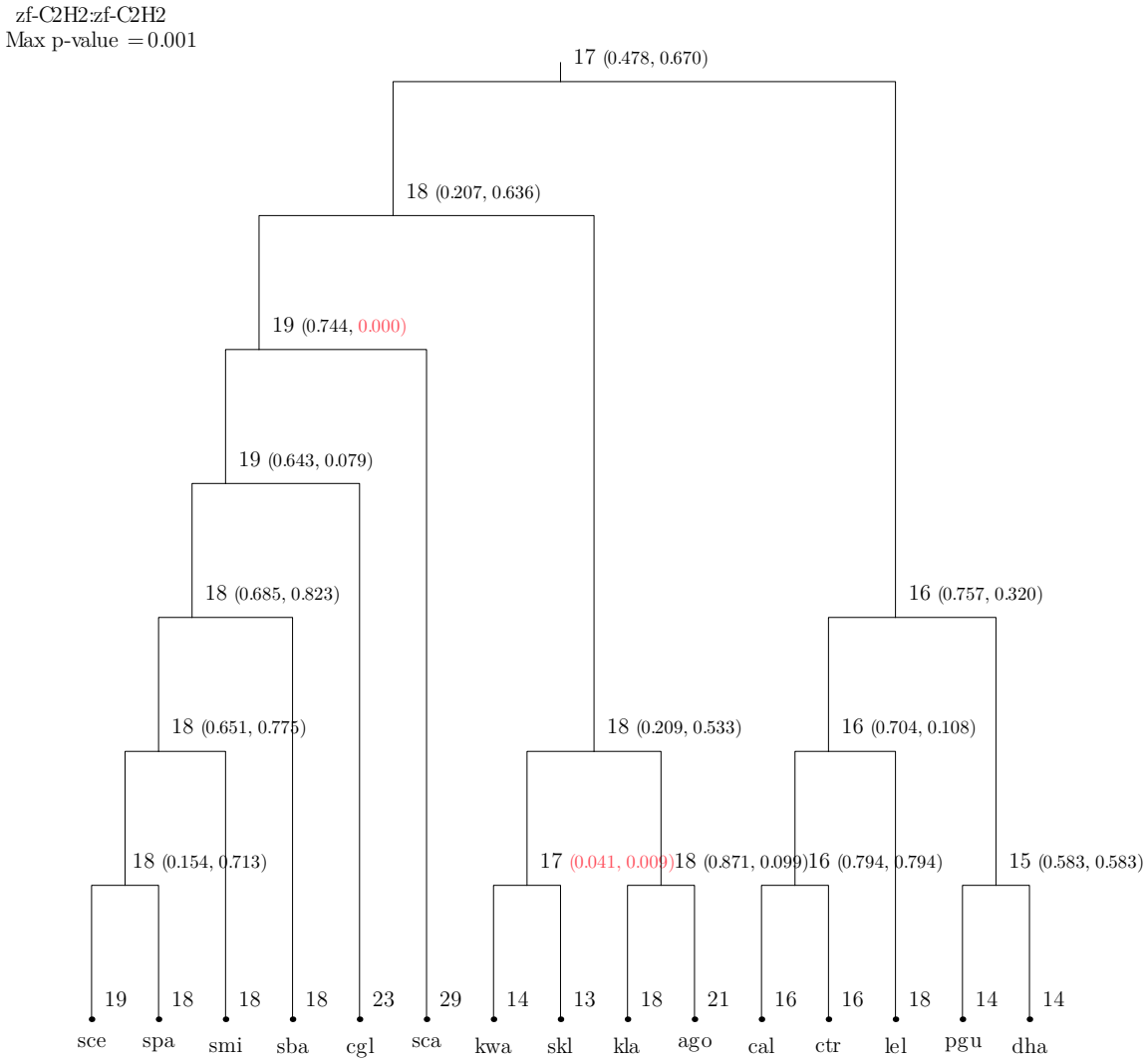


Figure A.3: CAFE output for C_2H_2:C_2H_2 TFs.

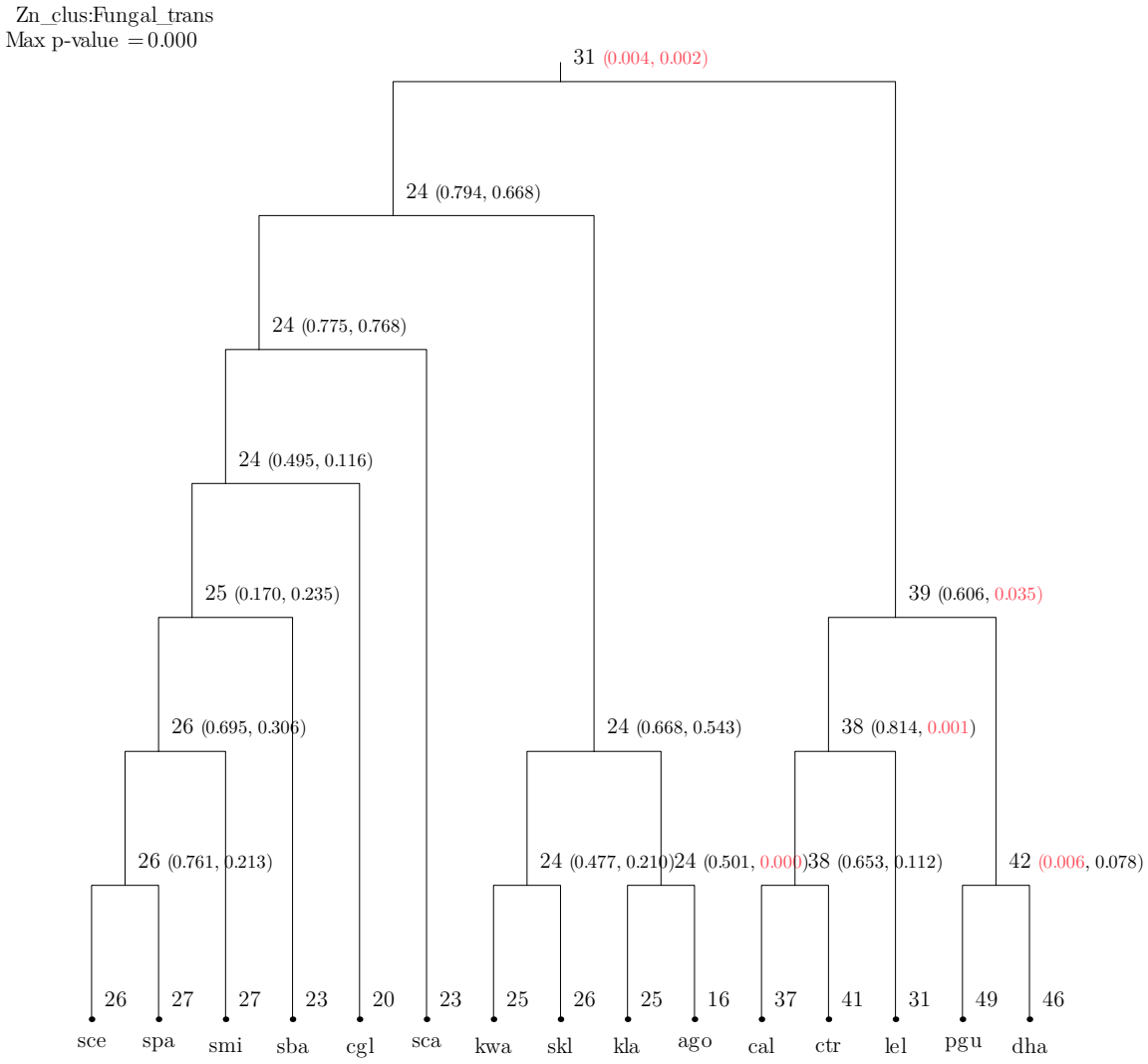


Figure A.4: CAFE output for Zn_clus:Fungal_trans TFs.

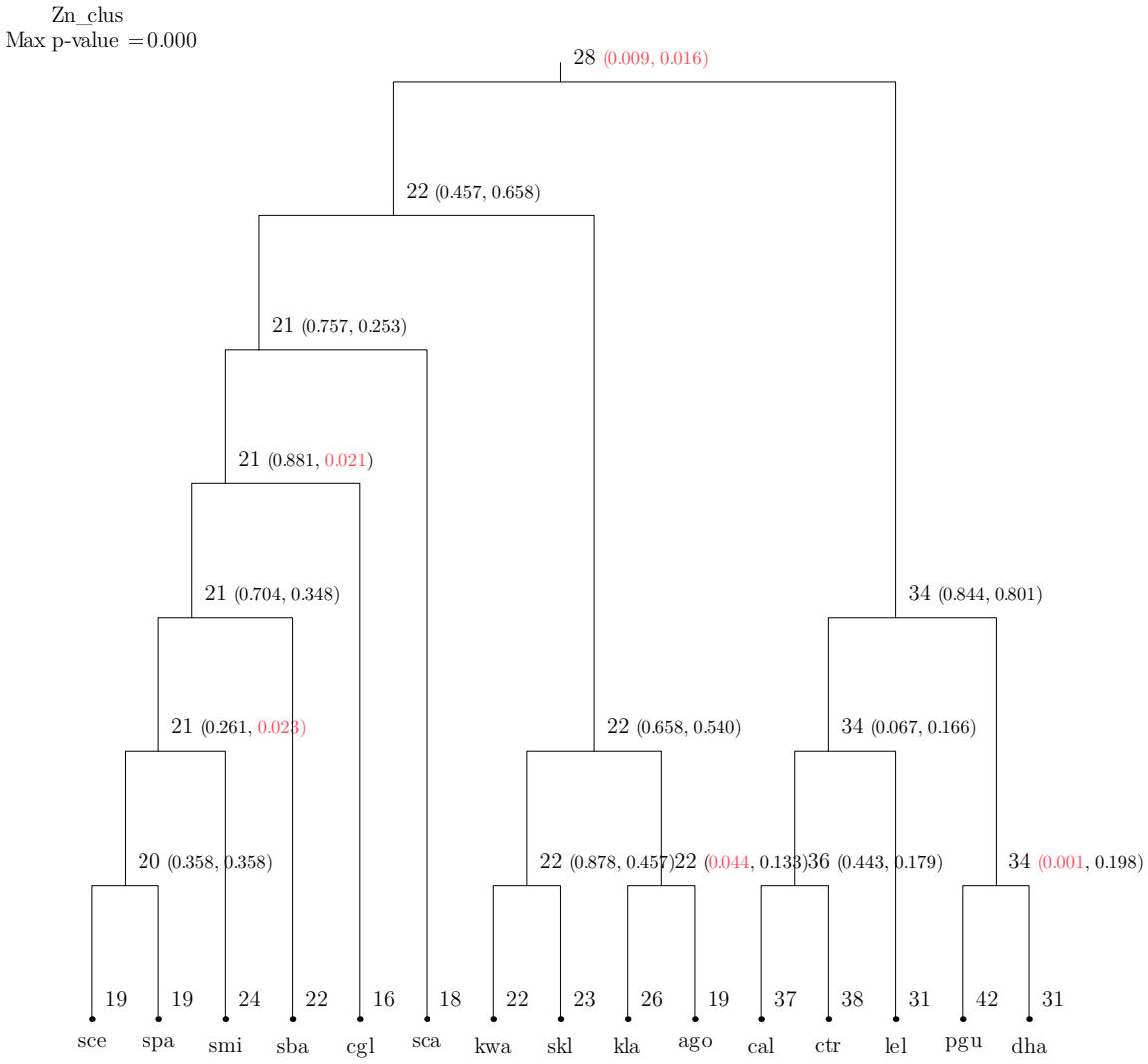


Figure A.5: CAFE output for Zn_clusTFs.

Whole-Genome Duplication TFs		
SGD ID	**Name**	**Annotation**
YML081W	—	unknown
YBR182C	SMP1	osmotic stress
YDR213W	UPC2	sterol biosynthesis
YNL068C	FKH2	cell cycle
YOR363C	PIP2	oleate-responsive; peroxisome proliferation
YDR501W	PLM2	DNA damage response
YDR146C	SWI5	cell cycle
YLR131C	ACE2	cell cycle
YHR056C	RSC30	osmotic stress; ribosomal proteins
YDL048C	STP4	unknown
YPL177C	CUP9	iron homeostasis
YBR049C	REB1	RNA pol I TF
YPL038W	MET31	sulfur metabolism
YKL062W	MSN4	general stress response
YHR006W	STP2	amino acid metabolism
YML027W	YOX1	cell cycle
YLR266C	PDR8	drug resistance
YGL071W	AFT1	iron homeostasis
YOR162C	YRR1	drug resistance
YDR303C	RSC3	osmotic stress; ribosomal proteins
YIL036W	CST6	oleate-responsive; utilisation of non-optimal carbon sources
YDR451C	YHP1	cell cycle
YDR096W	GIS1	nutrient limitation
YLR228C	ECM22	sterol biosynthesis
YOR028C	CIN5	drug resistance; salt tolerance; oxidative stress
YAL051W	OAF1	oleate-responsive; peroxisome proliferation
YPR008W	HAA1	weak acid stress
YLR375W	STP3	unknown
YBR066C	NRG2	glucose repression; filamentous growth suppressor
YDR026C	—	unknown
YJR127C	RSF2	glycerol-based growth; respiration
YKR034W	DAL80	nitrogen degradation
YDR463W	STP1	amino acid metabolism
YGL013C	PDR1	drug resistance
YMR072W	ABF2	mitochondrial DNA-binding protein
YPL230W	USV1	salt stress; non-fermentable carbon sources
YPL202C	AFT2	iron homeostasis
YGL166W	CUP2	copper homeostasis

SGD ID	Name	Annotation
YER045C	ACA1	utilisation of non-optimal carbon sources
YGL096W	TOS8	meiosis; cell damage
YDR253C	MET32	sulfur metabolism
YOL089C	HAL9	salt stress
YDR259C	YAP6	carbohydrate metabolism; salt stress
YDR423C	CAD1	iron homeostasis; drug resistance
YER169W	RPH1	DNA damage response
YBR150C	TBS1	unknown
YIR018W	YAP5	unknown
YJL110C	GZF3	nitrogen degradation
YIL131C	FKH1	cell cycle
YDR043C	NRG1	glucose repression; filamentous growth suppressor; pH response
YJR147W	HMS2	suppresses pseudohyphal growth
YER028C	MIG3	response to toxic agents
YMR016C	SOK2	PKA signalling
YLR183C	TOS4	unknown
YKL043W	PHD1	filamentation; cell cycle
YOL028C	YAP7	unknown
YKL032C	IXR1	respiratory chain
YPL089C	RLM1	cell integrity
YMR037C	MSN2	general stress response
YML007W	YAP1	oxidative stress
YHR206W	SKN7	oxidative stress
YMR182C	RGM1	unknown
YBL005W	PDR3	drug resistance

Table A.2: WGD duplicate TFs. Rows coloured in orange are stress-response regulators, blue nutrient-response regulators, green cell cycle regulators and purple drug resistance regulators.

CTG-amplified TFs		
OG ID	**# Duplications**	**Annotation**
222	10	lysine biosynthesis (?)
388	4	telomere accosiated
256	4	lysine biosynthesis (?)
317	4	drug resistance (?)
340	4	unknown function

OG ID	# Duplications	Annotation
318	3	unknown
329	3	Drug resistance
339	3	oxidative metabolism
370	3	Copper resistance
60	2	unknown
191	2	unknown
202	2	unknown
217	2	unknown
319	2	carbohydrate metabolism
326	2	unknown
330	2	homolog of CAT8
353	2	unknown
369	2	homolog of TEA1
22	1	methionine biosynthesis
30	1	iron acquisition
66	1	unknown
106	1	unknown
177	1	unknown
125	1	Purine biosynthesis
129	1	Drug resistance
154	1	Core stress response, proteasome
207	1	Calcium-responsive signalling
260	1	unknown
261	1	unknown
282	1	hyphal gene expression
290	1	unknown
325	1	unknown
336	1	Drug resistance
344	1	stress response
351	1	possible AA biosynthesis (homology)
368	1	unknown
372	1	unknown
374	1	GAL4, glycolysis
379	1	Morphogenesis, hyphal gene expression
382	1	Hyphal gene expression
388	1	Telomere accosiated

OG ID	# Duplications	Annotation

Table A.3: List of orthogroups that were amplified in the CTG clade and the number of inferred duplications per orthogroup.

References

Ahlroth, M.K., Ahlroth, P. & Kulomaa, M.S. (2001). Copy-number fluctuation by unequal crossing-over in the chicken avidin gene family. *Biochem Biophys Res Commun*, **288**, 400–6. 172

Ahnert, S.E., Fink, T.M.A. & Zinovyev, A. (2008). How much non-coding dna do eukaryotes require? *J Theor Biol*, **252**, 587–92. 13

Akerborg, O., Sennblad, B., Arvestad, L. & Lagergren, J. (2009). Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A*, **106**, 5714–5719. 141, 149, 150, 164, 165

Albert, Jeong & Barabasi (2000). Error and attack tolerance of complex networks. *Nature*, **406**, 378–82. 16

Almagro, A., Prista, C., Castro, S., Quintas, C., Madeira-Lopes, A., Ramos, J. & Loureiro-Dias, M.C. (2000). Effects of salts on debaryomyces hansenii and saccharomyces cerevisiae under stress conditions. *Int J Food Microbiol*, **56**, 191–7. 44

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**, 403–410. 103

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402. 61

Amoutzias, G.D., Robertson, D.L., Van de Peer, Y. & Oliver, S.G. (2008). Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci*, **33**, 220–9. 9

Amoutzias, G.D., He, Y., Gordon, J., Mossialos, D., Oliver, S.G. & Van de Peer, Y. (2010). Posttranslational regulation impacts the fate of duplicated genes. *Proc Natl Acad Sci U S A*, **107**, 2967–71. 125, 199

Anckar, J. & Sistonen, L. (2007). Sumo: getting it on. *Biochem Soc Trans*, **35**, 1409–13. 12

Aravind, L. & Koonin, E.V. (1999). Dna-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res*, **27**, 4658–70. 14, 92

Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M. & Iyer, L.M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev*, **29**, 231–62. 13, 91

Aravind, L., Anantharaman, V. & Venancio, T.M. (2009). Apprehending multicellularity: regulatory networks, genomics, and evolution. *Birth Defects Res C Embryo Today*, **87**, 143–64. 200

Artzy-Randrup, Y., Fleishman, S., Ben-Tal, N. & Stone, L. (2004). Comment on "network motifs: simple building blocks of complex networks" and "superfamilies of evolved and designed networks". *Science*, **305**, 1107. 16

Askew, C., Sellam, A., Epp, E., Hogues, H., Mullick, A., Nantel, A. & Whiteway, M. (2009). Transcriptional regulation of carbohydrate metabolism in the human pathogen candida albicans. *PLoS Pathog*, **5**, e1000612. 31, 33, 126, 129, 137, 240, 241

Babu, M., Teichmann, S.A. & Aravind, L. (2006a). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol*, **358**, 614–33. 13, 91

Babu, M., Teichmann, S.A. & Aravind, L. (2006b). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol*, **358**, 614–33. 196

Babu, M.M. (2010). Structure, evolution and dynamics of transcriptional regulatory networks. *Biochemical Society Transactions*, **038**, 1155–1178. 31, 40

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. & Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, **14**, 283–291. xi, 16, 17, 110

Babu, M.M., Iyer, L.M., Balaji, S. & Aravind, L. (2006c). The natural history of the wrky-gcm1 zinc fingers and the relationship between transcription factors and transposons. *Nucleic Acids Res*, **34**, 6505–20. 40, 113

Balaji, S., Babu, M.M., Iyer, L.M., Luscombe, N.M. & Aravind, L. (2006a). Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol*, **360**, 213–227. 13, 16, 20, 25

Balaji, S., Iyer, L.M., Aravind, L. & Babu, M.M. (2006b). Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. *J Mol Biol*, **360**, 204–12. 20

Banerjee, M., Thompson, D.S., Lazzell, A., Carlisle, P.L., Pierce, C., Monteagudo, C., López-Ribot, J.L. & Kadosh, D. (2008). Ume6, a novel filament-specific regulator of candida albicans hyphal extension and virulence. *Mol Biol Cell*, **19**, 1354–65. 197

Bauer, S., Grossmann, S., Vingron, M. & Robinson, P.N. (2008). Ontologizer 2.0–a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1. 233

Beggs, J.D. (2005). Lsm proteins and rna processing. *Biochem Soc Trans*, **33**, 433–8. 143

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300. 174, 233

Bergmann, S., Ihmels, J. & Barkai, N. (2004). Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*, **2**, E9. 27

Bevan, R.B., Bryant, D. & Lang, B.F. (2007). Accounting for gene rate heterogeneity in phylogenetic inference. *Syst Biol*, **56**, 194–205. 253

Bhardwaj, N., Kim, P.M. & Gerstein, M.B. (2010a). Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Sci Signal*, **3**, ra79. 202

Bhardwaj, N., Yan, K.K. & Gerstein, M.B. (2010b). Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels. *Proc Natl Acad Sci U S A*, **107**, 6841–6. xi, 18, 19, 20, 181, 182, 202, 210, 212, 218, 226, 227

Bininda-Emonds, O.R.P. (2005). Supertree construction in the genomic age. *Methods Enzymol*, **395**, 745–757. 52

Birney, E., Clamp, M. & Durbin, R. (2004). Genewise and genomewise. *Genome Res*, **14**, 988–995. 101

Björklund, S. & Gustafsson, C.M. (2005). The yeast mediator complex and its regulation. *Trends Biochem Sci*, **30**, 240–4. 5, 6

Blaiseau, P.L. & Thomas, D. (1998). Multiple transcriptional activation complexes tether the yeast activator met4 to dna. *EMBO J*, **17**, 6327–36. 268

Blaiseau, P.L., Isnard, A.D., Surdin-Kerjan, Y. & Thomas, D. (1997). Met31p and met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol Cell Biol*, **17**, 3640–8. 268

Blanquart, S. & Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol*, **25**, 842–858. 252

Boehr, D.D., Nussinov, R. & Wright, P.E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol*, **5**, 789–96. 12

Bofkin, L. (2005). *The Causes and Consequences of Variation in Evolutionary Processes Acting on DNA Sequences*. Ph.D. thesis, University of Cambridge. 56, 58

Bollback, J.P. (2002). Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol*, **19**, 1171–80. 253

Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M. & Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. *Science*, **317**, 815–819. 24, 27, 28, 29, 258

Brauer, M.J., Huttenhower, C., Airoldi, E.M., Rosenstein, R., Matese, J.C., Gresham, D., Boer, V.M., Troyanskaya, O.G. & Botstein, D. (2008). Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Mol Biol Cell*, **19**, 352–67. 46, 197, 247, 257

Brayer, K.J. & Segal, D.J. (2008). Keep your fingers off my dna: protein-protein interactions mediated by c2h2 zinc finger domains. *Cell Biochem Biophys*, **50**, 111–31. 108

Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–5. 22

Brinkmann, H., van der Giezen, M., Zhou, Y., de Raucourt, G.P. & Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol*, **54**, 743–757. 51, 55

Britten, R.J. & Davidson, E.H. (1971). Repetitive and non-repetitive dna sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol*, **46**, 111–38. 1

Brown, A. (2006). Integration of metabolism with virulence in candida albicans. In K. Esser & A. Brown, eds., *Fungal Genomics*, vol. 13 of *The Mycota*, 185–203, Springer Berlin Heidelberg. 48

Brown, V., Sabina, J. & Johnston, M. (2009). Specialized sugar sensing in diverse fungi. *Curr Biol*, **19**, 436–41. 33, 240, 241

Bruno, V.M. & Mitchell, A.P. (2005). Regulation of azole drug susceptibility by candida albicans protein kinase ck2. *Mol Microbiol*, **56**, 559–73. 133

Bruno, V.M., Wang, Z., Marjani, S.L., Euskirchen, G.M., Martin, J., Sherlock, G. & Snyder, M. (2010). Comprehensive annotation of the transcriptome of the human fungal pathogen candida albicans using rna-seq. *Genome Res*, **20**, 1451–8. 198

Buchler, N.E. & Louis, M. (2008). Molecular titration and ultrasensitivity in regulatory networks. *J Mol Biol*, **384**, 1106–19. 9, 269

Bulder, C.J. (1964). Induction of petite mutation and inhibition of synthesis of respiratory enzymes in various yeasts. *Antonie Van Leeuwenhoek*, **30**, 1–9. 122

Bullard, J.H., Mostovoy, Y., Dudoit, S. & Brem, R.B. (2010). Polygenic and directional regulatory evolution across pathways in saccharomyces. *Proc Natl Acad Sci U S A*, **107**, 5058–63. 22

Burnham, K.P. & Anderson, D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods Research*, **33**, 261–304. 66, 178, 253, 254

Bussereau, F., Casaregola, S., Lafay, J.F. & Bolotin-Fukuhara, M. (2006). The kluyveromyces lactis repertoire of transcriptional regulators. *FEMS Yeast Res*, **6**, 325–335. 41, 92

Butler, G., Kenny, C., Fagan, A., Kurischko, C., Gaillardin, C. & Wolfe, K.H. (2004). Evolution of the mat locus and its ho endonuclease in yeast species. *Proc Natl Acad Sci USA*, **101**, 1632–7. 113, 114

Butler, G., Rasmussen, M.D., Lin, M.F., Santos, M.A.S., Sakthikumar, S., Munro, C.A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J.L., Agrafioti, I., Arnaud, M.B., Bates, S., Brown, A.J.P.,

Brunke, S., Costanzo, M.C., Fitzpatrick, D.A., de Groot, P.W.J., Harris, D., Hoyer, L.L., Hube, B., Klis, F.M., Kodira, C., Lennard, N., Logue, M.E., Martin, R., Neiman, A.M., Nikolaou, E., Quail, M.A., Quinn, J., Santos, M.C., Schmitzberger, F.F., Sherlock, G., Shah, P., Silverstein, K.A.T., Skrzypek, M.S., Soll, D., Staggs, R., Stansfield, I., Stumpf, M.P.H., Sudbery, P.E., Srikantha, T., Zeng, Q., Berman, J., Berriman, M., Heitman, J., Gow, N.A.R., Lorenz, M.C., Birren, B.W., Kellis, M. & Cuomo, C.A. (2009). Evolution of pathogenicity and sexual reproduction in eight candida genomes. *Nature*, **459**, 657–62. xii, 42, 44, 45, 48, 134, 233

Byrne, K.P. & Wolfe, K.H. (2005). The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*, **15**, 1456–1461. 43, 108, 134, 145, 158, 164, 184

Byrne, K.P. & Wolfe, K.H. (2006). Visualizing syntenic relationships among the hemiascomycetes with the yeast gene order browser. *Nucleic Acids Res*, **34**, D452–D455. 61, 79

Byrne, K.P. & Wolfe, K.H. (2007). Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, **175**, 1341–1350. 38, 169, 202, 258

Carlson, M. (1999). Glucose repression in yeast. *Curr Opin Microbiol*, **2**, 202–7. 46, 129

Carroll, S.B. (2005). *Endless forms most beautiful*. W.W. Norton and Company, Inc., New York. 20, 21, 24, 26

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*, **17**, 540–552. 62, 96

Castrillo, J. & Oliver, S. (2006). *Metabolomics and Systems Biology in Saccharomyces cerevisiae*, vol. 13 of *The Mycota*, 3–18. Springer Berlin Heidelberg. 42

Chan, E.T., Quon, G.T., Chua, G., Babak, T., Trochesset, M., Zirngibl, R.A., Aubin, J., Ratcliffe, M.J.H., Wilde, A., Brudno, M., Morris, Q.D. & Hughes, T.R. (2009). Conservation of core gene expression in vertebrate tissues. *J Biol*, **8**, 33. 27

Chan, Y.F., Marks, M.E., Jones, F.C., Villarreal, G., Jr, Shapiro, M.D., Brady, S.D., Southwick, A.M., Absher, D.M., Grimwood, J., Schmutz, J., Myers, R.M., Petrov, D., Jónsson, B., Schluter, D., Bell, M.A. & Kingsley, D.M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitx1 enhancer. *Science*, **327**, 302–5. 20

Chang, Y.W., Robert Liu, F.G., Yu, N., Sung, H.M., Yang, P., Wang, D., Huang, C.J., Shih, M.C. & Li, W.H. (2008). Roles of cis- and trans-changes in the regulatory evolution of genes in the gluconeogenic pathway in yeast. *Mol Biol Evol*, **25**, 1863–75. 22, 23

Charoensawan, V., Wilson, D. & Teichmann, S.A. (2010a). Genomic repertoires of dna-binding transcription factors across the tree of life. *Nucleic Acids Res*. 7, 13, 14, 91, 92

Charoensawan, V., Wilson, D. & Teichmann, S.A. (2010b). Lineage-specific expansion of dna-binding transcription factor families. *Trends Genet*, **26**, 388–93. 7, 11, 13, 14, 15, 25, 41, 91, 92, 255

Chen, C.G., Yang, Y.L., Tseng, K.Y., Shih, H.I., Liou, C.H., Lin, C.C. & Lo, H.J. (2009). Rep1p negatively regulating mdr1 efflux pump involved in drug resistance in candida albicans. *Fungal Genet Biol*, **46**, 714–20. 116

Chen, K., Durand, D. & Farach-Colton, M. (2000). Notung: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol*, **7**, 429–447. 138, 151, 167

Chen, K., van Nimwegen, E., Rajewsky, N. & Siegal, M.L. (2010). Correlating gene expression variation with cis-regulatory polymorphism in saccharomyces cerevisiae. *Genome Biol Evol*, **2**, 697–707. 21

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87. 1

Cho, G., Kim, J., Rho, H.M. & Jung, G. (1995). Structure-function analysis of the dna binding domain of saccharomyces cerevisiae abf1. *Nucleic Acids Res*, **23**, 2980–7. 119, 121

Choi, J.K. & Kim, Y.J. (2008). Epigenetic regulation and the variability of gene expression. *Nat Genet*, **40**, 141–147. 23

Cipollina, C., Alberghina, L., Porro, D. & Vai, M. (2005). Sfp1 is involved in cell size modulation in respiro-fermentative growth conditions. *Yeast*, **22**, 385–99. 130

Conant, G.C. (2010). Rapid reorganization of the transcriptional regulatory network after genome duplication in yeast. *Proc Biol Sci*, **277**, 869–76. 182, 188, 199, 258

Conant, G.C. & Wolfe, K.H. (2006). Functional partitioning of yeast coexpression networks after genome duplication. *PLoS Biol*, **4**, e109. 38

Conant, G.C. & Wolfe, K.H. (2007). Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol Syst Biol*, **3**, 129. 47, 122, 129, 260

Conant, G.C. & Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet*, **9**, 938–50. 182, 260

Cornell, M.J., Alam, I., Soanes, D.M., Wong, H.M., Hedeler, C., Paton, N.W., Rattray, M., Hubbard, S.J., Talbot, N.J. & Oliver, S.G. (2007). Comparative genome analysis across a kingdom of eukaryotic organisms: specialization and diversification in the fungi. *Genome Res*, **17**, 1809–1822. 56, 58

Cosentino Lagomarsino, M., Jona, P., Bassetti, B. & Isambert, H. (2007). Hierarchy and feedback in the evolution of the escherichia coli transcription network. *Proc Natl Acad Sci U S A*, **104**, 5516–20. 18

Cowen, L.E. & Steinbach, W.J. (2008). Stress, drugs, and evolution: the role of cellular signaling in fungal drug resistance. *Eukaryot Cell*, **7**, 747–64. 131, 132

Davis, D., Edwards, J.E., Jr, Mitchell, A.P. & Ibrahim, A.S. (2000). Candida albicans rim101 ph response pathway is required for host-pathogen interactions. *Infect Immun*, **68**, 5953–9. 247

De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. (2006). Cafe: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–71. 177

De Deken, R.H. (1966). The crabtree effect: a regulatory system in yeast. *J Gen Microbiol*, **44**, 149–56. 46

Degnan, J.H. & Rosenberg, N.A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet*, **2**, e68. 53, 55

Degnan, J.H. & Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*, **24**, 332–340. 53, 54, 55, 68, 90, 138, 254

Delsuc, F., Phillips, M.J. & Penny, D. (2003). Comment on "hexapod origins: monophyletic or paraphyletic?". *Science*, **301**, 1482; author reply 1482. 51, 55

Delsuc, F., Brinkmann, H. & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, **6**, 361–375. 52

Dermitzakis, E.T., Bergman, C.M. & Clark, A.G. (2003). Tracing the evolutionary history of drosophila regulatory regions with models that identify transcription factor binding sites. *Mol Biol Evol*, **20**, 703–14. 28

Deutschbauer, A.M., Jaramillo, D.F., Proctor, M., Kumm, J., Hillenmeyer, M.E., Davis, R.W., Nislow, C. & Giaever, G. (2005). Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics*, **169**, 1915–25. 31

Dietrich, F.S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., P'ohlmann, R., Luedi, P., Choi, S., Wing, R.A., Flavier, A., Gaffney, T.D. & Philippsen, P. (2004). The ashbya gossypii genome as a tool for mapping the ancient saccharomyces cerevisiae genome. *Science*, **304**, 304–307. 42

Diezmann, S., Cox, C.J., Schönian, G., Vilgalys, R.J. & Mitchell, T.G. (2004). Phylogeny and evolution of medical species of candida and related taxa: a multigenic analysis. *J Clin Microbiol*, **42**, 5624–5635. xii, 56, 58, 59, 89

Doniger, S.W. & Fay, J.C. (2007). Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol*, **3**, e99. 27, 28

Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005). Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–4. 207

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O. & Arnold, F.H. (2005). Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A*, **102**, 14338–43. 201, 207

Du, J.X., McConnell, B.B. & Yang, V.W. (2010). A small ubiquitin-related modifier-interacting motif functions as the transcriptional activation domain of krüppel-like factor 4. *J Biol Chem*, **285**, 28298–308. 10

Dubacq, C., Chevalier, A. & Mann, C. (2004). The protein kinase snf1 is required for tolerance to the ribonucleotide reductase inhibitor hydroxyurea. *Mol Cell Biol*, **24**, 2560–72. 262, 264

Dujon, B. (2010). Yeast evolutionary genomics. *Nat Rev Genet*, **11**, 512–24. 42, 43, 44, 47, 140

Dujon, B., Sherman, D., Fischer, G. *et al.* (2004). Genome evolution in yeasts. *Nature*, **430**, 35–44. 42, 44, 61, 172

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. 96, 149

Eddy, S.R. (1998). Profile hidden markov models. *Bioinformatics*, **14**, 755–763. 149

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform*, **23**, 205–11. 96

Edgar, R.C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–7. 142

Edger, P.P. & Pires, J.C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res*, **17**, 699–717. 182

Eggimann, P., Garbino, J. & Pittet, D. (2003). Epidemiology of candida species infections in critically ill non-immunosuppressed patients. *Lancet Infect Dis*, **3**, 685–702. 131

Emerson, J.J., Hsieh, L.C., Sung, H.M., Wang, T.Y., Huang, C.J., Lu, H.H.S., Lu, M.Y.J., Wu, S.H. & Li, W.H. (2010). Natural selection on cis and trans regulation in yeasts. *Genome Res*, **20**, 826–36. 22, 23

Emerson, R.O. & Thomas, J.H. (2009). Adaptive evolution in zinc finger transcription factors. *PLoS Genet*, **5**, e1000325. 7, 111, 116

Erkine, A.M. (2004). Activation domains of gene-specific transcription factors: are histones among their targets? *Biochem Cell Biol*, **82**, 453–9. 10

Evangelisti, A.M. & Wagner, A. (2004). Molecular evolution in the yeast transcriptional regulation network. *J Exp Zool B Mol Dev Evol*, **302**, 392–411. 202, 229

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791. 68

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates Inc., Massachusetts. 64

Fernandes, A.R., Mira, N.P., Vargas, R.C., Canelhas, I. & Sá-Correia, I. (2005). Saccharomyces cerevisiae adaptation to weak acids involves the transcription factor haa1p and haa1p-regulated genes. *Biochem Biophys Res Commun*, **337**, 95–103. 265

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R. & Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Res*, **38**, D211–D222. 93, 96, 101, 103, 105, 107

Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L. & McCallion, A.S. (2006). Conservation of ret regulatory function from human to zebrafish without sequence similarity. *Science*, **312**, 276–9. 28

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool*, **19**, 99–113. 150

Fitzpatrick, D.A., Logue, M.E., Stajich, J.E. & Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol*, **6**, 99. xii, 56, 57, 58, 59, 60, 86

Fitzpatrick, D.A., O'Gaora, P., Byrne, K.P. & Butler, G. (2010). Analysis of gene evolution and metabolic pathways using the candida gene order browser. *BMC Genomics*, **11**, 290. 147, 158

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. & Feldman, M.W. (2002). Evolutionary rate in the protein interaction network. *Science*, **296**, 750–2. 202

Fraser, H.B., Moses, A.M. & Schadt, E.E. (2010). Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc Natl Acad Sci U S A*, **107**, 2977–82. 22

Fusco, D., Grassi, L., Bassetti, B., Caselle, M. & Cosentino Lagomarsino, M. (2010). Ordered structure of the transcription network inherited from the yeast whole-genome duplication. *BMC Syst Biol*, **4**, 77. 182, 199, 227, 258

Fuxreiter, M., Tompa, P., Simon, I., Uversky, V.N., Hansen, J.C. & Asturias, F.J. (2008). Malleable machines take shape in eukaryotic transcriptional regulation. *Nat Chem Biol*, **4**, 728–37. 26

Gao, F., Foat, B.C. & Bussemaker, H.J. (2004). Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31. 29

Gasch, A.P. (2007). Comparative genomics of the environmental stress response in ascomycete fungi. *Yeast*, **24**, 961–976. 248, 257

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D. & Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, **11**, 4241–4257. 33, 197, 244, 247

Gasch, A.P., Moses, A.M., Chiang, D.Y., Fraser, H.B., Berardini, M. & Eisen, M.B. (2004). Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol*, **2**, e398. 27, 31, 32, 248

Génolevures Consortium, Souciet, J.L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P.V., Cliften, P., Sherman, D.J., Weissenbach, J., Westhof, E., Wincker, P., Jubin, C., Poulain, J., Barbe, V., Ségurens, B., Artiguenave, F., Anthouard, V., Vacherie, B., Val, M.E., Fulton, R.S., Minx, P., Wilson, R., Durrens, P., Jean, G., Marck, C., Martin, T., Nikolski, M., Rolland, T., Seret, M.L., Casarégola, S., Despons, L., Fairhead, C., Fischer, G., Lafontaine, I., Leh, V., Lemaire, M., de Montigny, J., Neuvéglise, C., Thierry, A., Blanc-Lenfle, I., Bleykasten, C., Diffels, J., Fritsch, E., Frangeul, L., Goëffon, A., Jauniaux, N., Kachouri-Lafond, R., Payen, C., Potier, S., Pribylova, L.,

Ozanne, C., Richard, G.F., Sacerdot, C., Straub, M.L. & Talla, E. (2009). Comparative genomics of protoploid saccharomycetaceae. *Genome Res*, **19**, 1696–1709. 42, 43, 44, 114

Goldman, N. (1993). Statistical tests of models of dna substitution. *J Mol Evol*, **36**, 182–98. 253

Goldman, N. & Whelan, S. (2002). A novel use of equilibrium frequencies in models of sequence evolution. *Mol Biol Evol*, **19**, 1821–1831. 63, 67

Goldman, N., Thorne, J.L. & Jones, D.T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, **149**, 445–458. 54

Gompel, N., Prud'homme, B., Wittkopp, P.J., Kassner, V.A. & Carroll, S.B. (2005). Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in drosophila. *Nature*, **433**, 481–7. 20

Gonçalves, P.M., Maurer, K., Mager, W.H. & Planta, R.J. (1992). Kluyveromyces contains a functional abf1-homologue. *Nucleic Acids Res*, **20**, 2211–5. 119

Gordon, J.L., Byrne, K.P. & Wolfe, K.H. (2009). Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. *PLoS Genet*, **5**, e1000485. 147

Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *J Mol Biol*, **313**, 903–919. 105

Graham, I.R., Haw, R.A., Spink, K.G., Halden, K.A. & Chambers, A. (1999). In vivo analysis of functional regions within yeast rap1p. *Mol Cell Biol*, **19**, 7481–90. 41

Guelzim, N., Bottani, S., Bourgine, P. & Képès, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, **31**, 60–3. 16, 181

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, **52**, 696–704. 68

Haerty, W., Artieri, C., Khezri, N., Singh, R.S. & Gupta, B.P. (2008). Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution. *BMC Genomics*, **9**, 399. 41, 92

Hahn, M.W. (2007). Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol*, **8**, R141. xvi, 138, 139, 140, 155, 156, 166, 172

Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. & Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res*, **15**, 1153–1160. 177

Halder, G., Callaerts, P. & Gehring, W.J. (1995). Induction of ectopic eyes by targeted expression of the eyeless gene in drosophila. *Science*, **267**, 1788–92. 20

Hao, B., Clancy, C.J., Cheng, S., Raman, S.B., Iczkowski, K.A. & Nguyen, M.H. (2009). Candida albicans rfx2 encodes a dna binding protein involved in dna damage responses, morphogenesis, and virulence. *Eukaryot Cell*, **8**, 627–39. 116, 133

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E. & Young, R.A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104. 15

Hasegawa, M., Kishino, H. & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Mol Evol*, **22**, 160–174. 54, 63

Haw, R., Devi Yarragudi, A. & Uemura, H. (2001). Isolation of gcr1, a major transcription factor of glycolytic genes in saccharomyces cerevisiae, from kluyveromyces lactis. *Yeast*, **18**, 729–35. 126, 129

He, X. & Zhang, J. (2006). Higher duplicability of less important genes in yeast genomes. *Mol Biol Evol*, **23**, 144–51. 182

Heled, J. & Drummond, A.J. (2010). Bayesian inference of species trees from multilocus data. *Mol Biol Evol*, **27**, 570–80. 53, 254

Herbig, E., Warfield, L., Fish, L., Fishburn, J., Knutson, B.A., Moorefield, B., Pacheco, D. & Hahn, S. (2010). Mechanism of mediator recruitment by tandem gcn4 activation domains and three gal11 activator-binding domains. *Mol Cell Biol*, **30**, 2376–90. 10

Hibbett, D.S., Binder, M., Bischoff, J.F., Blackwell, M., Cannon, P.F., Eriksson, O.E., Huhndorf, S., James, T., Kirk, P.M., Lücking, R., Thorsten Lumbsch, H., Lutzoni, F., Matheny, P.B., McLaughlin, D.J., Powell, M.J., Redhead, S., Schoch, C.L., Spatafora, J.W., Stalpers, J.A., Vilgalys, R., Aime, M.C., Aptroot, A., Bauer, R., Begerow, D., Benny, G.L., Castlebury, L.A., Crous, P.W., Dai, Y.C., Gams, W., Geiser, D.M., Griffith, G.W., Gueidan, C., Hawksworth, D.L., Hestmark, G., Hosaka, K., Humber, R.A., Hyde, K.D., Ironside, J.E., Kõljalg, U., Kurtzman, C.P., Larsson, K.H., Lichtwardt, R., Longcore, J., Miadlikowska, J., Miller, A., Moncalvo, J.M., Mozley-Standridge, S., Oberwinkler, F., Parmasto, E., Reeb, V., Rogers, J.D., Roux, C., Ryvarden, L., Sampaio, J.P., Schüssler, A., Sugiyama, J., Thorn, R.G., Tibell, L., Untereiner, W.A., Walker, C., Wang, Z., Weir, A., Weiss, M., White, M.M., Winka, K., Yao, Y.J. & Zhang, N. (2007). A higher-level phylogenetic classification of the fungi. *Mycol Res*, **111**, 509–47. 42

Hittinger, C.T. & Carroll, S.B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, **449**, 677–81. 126

Hittinger, C.T., Rokas, A. & Carroll, S.B. (2004). Parallel inactivation of multiple gal pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci U S A*, **101**, 14144–14149. 33, 129

Ho, S.Y. & Jermiin, L. (2004). Tracing the decay of the historical signal in biological sequence data. *Syst Biol*, **53**, 623–637. 54

Hoekstra, H.E. & Coyne, J.A. (2007). The locus of evolution: evo devo and the genetics of adaptation. *Evolution*, **61**, 995–1016. 2, 21, 25

Hogues, H., Lavoie, H., Sellam, A., Mangos, M., Roemer, T., Purisima, E., Nantel, A. & Whiteway, M. (2008). Transcription factor substitution during the evolution of fungal ribosome regulation. *Mol Cell*, **29**, 552–62. 31, 32, 244

Holmberg, C.I., Tran, S.E.F., Eriksson, J.E. & Sistonen, L. (2002). Multisite phosphorylation provides sophisticated regulation of transcription factors. *Trends Biochem Sci*, **27**, 619–27. xi, 10, 11

Hsu, T., Gogos, J.A., Kirsh, S.A. & Kafatos, F.C. (1992). Multiple zinc finger forms resulting from developmentally regulated alternative splicing of a transcription factor gene. *Science*, **257**, 1946–50. 26

Huang, L., Guan, R.J. & Pardee, A.B. (1999). Evolution of transcriptional control from prokaryotic beginnings to eukaryotic complexities. *Crit Rev Eukaryot Gene Expr*, **9**, 175–82. 11

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. & Yeats, C. (2009). Interpro: the integrative protein signature database. *Nucleic Acids Res*, **37**, 211–215. 94, 95, 105, 111, 113, 121

Hurles, M. (2004). Gene duplication: the genomic trade in spare parts. *PLoS Biol*, **2**, E206. 172

Hynes, M.J., Murray, S.L., Duncan, A., Khew, G.S. & Davis, M.A. (2006). Regulatory genes controlling fatty acid catabolism and peroxisomal functions in the filamentous fungus aspergillus nidulans. *Eukaryot Cell*, **5**, 794–805. 33, 48

Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J. & Barkai, N. (2005). Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science*, **309**, 938–40. 31, 33, 46, 197

Ingram, P., Stumpf, M. & Stark, J. (2006). Network motifs: structure does not determine function. *BMC Genomics*, **7**, 108. 16

Iyer, L.M., Anantharaman, V., Wolf, M.Y. & Aravind, L. (2008). Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *Int J Parasitol*, **38**, 1–31. 200

Jacob, F. (1977). Evolution and tinkering. *Science*, **196**, 1161–6. 1

Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends Genet*, **22**, 225–231. xii, 55, 56, 57, 58, 59, 60, 86

Jiang, C., Gu, J., Chopra, S., Gu, X. & Peterson, T. (2004). Ordered origin of the typical two- and three-repeat myb genes. *Gene*, **326**, 13–22. 111

Jordan, I.K., Wolf, Y.I. & Koonin, E.V. (2003). No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol Biol*, **3**, 1. 202

Jothi, R., Balaji, S., Wuster, A., Grochow, J.A., Gsponer, J., Przytycka, T.M., Aravind, L. & Babu, M.M. (2009). Genomic analysis reveals

a tight link between transcription factor dynamics and regulatory network architecture. *Mol Syst Biol*, **5**, 294–294. xv, 18, 19, 105, 106, 107, 181, 182, 183, 186, 188, 189, 202, 207, 210, 212, 214, 218, 226, 227, 258

Jovelin, R. & Phillips, P.C. (2009). Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol*, **10**, R35. 202, 210, 213, 223, 225, 226, 227, 229, 255, 258

Jukes, T. & Cantor, C. (1969). *Mammalian protein metabolism*, chap. Evolution of protein molecules, 21–132. New York: Academic Press. 63

Kainou, T., Shinzato, T., Sasaki, K., Mitsui, Y., Giga-Hama, Y., Kumagai, H. & Uemura, H. (2006). Spsgt1, a new essential gene of schizosaccharomyces pombe, is involved in carbohydrate metabolism. *Yeast*, **23**, 35–53. 113, 126, 127

Kanehisa, M. & Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27–30. 233

Karababa, M., Valentino, E., Pardini, G., Coste, A.T., Bille, J. & Sanglard, D. (2006). Crz1, a target of the calcineurin pathway in candida albicans. *Mol Microbiol*, **59**, 1429–51. 133

Kasahara, K., Ohtsuki, K., Ki, S., Aoyama, K., Takahashi, H., Kobayashi, T., Shirahige, K. & Kokubo, T. (2007). Assembly of regulatory factors on rrna and ribosomal protein genes in saccharomyces cerevisiae. *Mol Cell Biol*, **27**, 6686–705. 31, 244

Katoh, K. & Toh, H. (2008). Recent developments in the mafft multiple sequence alignment program. *Briefings in Bioinformatics*, **9**. 62, 95, 142

Keeling, P.J. & Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*, **9**, 605–18. 111

Keller, G., Ray, E., Brown, P.O. & Winge, D.R. (2001). Haa1, a protein homologous to the copper-regulated transcription factor ace1, is a novel transcriptional activator. *J Biol Chem*, **276**, 38697–702. 265

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254. 42

Kellis, M., Birren, B.W. & Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, **428**, 617–624. 43, 57

Kent, W.J. (2002). Blat–the blast-like alignment tool. *Genome Res*, **12**, 656–664. 63

Kim, E., Magen, A. & Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Res*, **35**, 125–31. 2

Kim, J.H. & Johnston, M. (2006). Two glucose-sensing pathways converge on rgt1 to regulate expression of glucose transporter genes in saccharomyces cerevisiae. *J Biol Chem*, **281**, 26144–9. 129

King, M.C. & Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116. 1, 20

Kishino, H. & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *J Mol Evol*, **29**, 170–179. 80

Knop, M. (2006). Evolution of the hemiascomycete yeasts: on life styles and the importance of inbreeding. *Bioessays*, **28**, 696–708. 43, 44

Knowles, D.G. & McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Res*, **19**, 1752–9. 1

Kodadek, T., Sikder, D. & Nalley, K. (2006). Keeping transcriptional activators under control. *Cell*, **127**, 261–4. 12

Köhler, T., Wesche, S., Taheri, N., Braus, G.H. & Mösch, H.U. (2002). Dual role of the saccharomyces cerevisiae tea/atts family transcription factor tec1p in regulation of gene expression and cellular development. *Eukaryot Cell*, **1**, 673–86. 130

Koshi, J.M. & Goldstein, R.A. (1995). Context-dependent optimal substitution matrices. *Protein Eng*, **8**, 641–645. 54

Krishna, S.S., Majumdar, I. & Grishin, N.V. (2003). Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res*, **31**, 532–50. 105

Kubatko, L.S., Carstens, B.C. & Knowles, L.L. (2009). Stem: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, **25**, 971–3. 53

Kullas, A.L., Martin, S.J. & Davis, D. (2007). Adaptation to environmental ph: integrating the rim101 and calcineurin signal transduction pathways. *Mol Microbiol*, **66**, 858–71. 133, 247

Kuo, D., Licon, K., Bandyopadhyay, S., Chuang, R., Luo, C., Catalana, J., Ravasi, T., Tan, K. & Ideker, T. (2010). Coevolution within a transcriptional network by compensatory trans and cis mutations. *Genome Res*, **20**, 1672–8. 39

Kuramae, E.E., Robert, V., Snel, B., Weiss, M. & Boekhout, T. (2006). Phylogenomics reveal a robust fungal tree of life. *FEMS Yeast Res*, **6**, 1213–1220. 56, 57

Kuramae, E.E., Robert, V., Echavarri-Erasun, C. & Boekhout, T. (2007). Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom. *BMC Evol Biol*, **7**, 134–134. 55, 56

Kuras, L., Cherest, H., Surdin-Kerjan, Y. & Thomas, D. (1996). A heteromeric complex containing the centromere binding factor 1 and two basic leucine zipper factors, met4 and met28, mediates the transcription activation of yeast sulfur metabolism. *EMBO J*, **15**, 2519–29. 268

Kurtzman, C.P. & Robnett, C.J. (2003). Phylogenetic relationships among yeasts of the 'saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Res*, **3**, 417–432. xii, 56, 57, 58, 59, 89

Lambert, J.R., Bilanchone, V.W. & Cumsky, M.G. (1994). The ord1 gene encodes a transcription factor involved in oxygen regulation and is identical to ixr1, a gene that confers cisplatin sensitivity to saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, **91**, 7345–9. 116

Lartillot, N. & Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*, **21**, 1095–1109. 54, 56

Latchman, D.S. (2007). *Eukaryotic Transcription Factors*. Academic Press, 5th edn. 10

Lavoie, H., Hogues, H. & Whiteway, M. (2009). Rearrangements of the transcriptional regulatory networks of metabolic pathways in fungi. *Curr Opin Microbiol*, **12**, 655–63. xii, 33, 34, 126, 130

Lavoie, H., Hogues, H., Mallick, J., Sellam, A., Nantel, A. & Whiteway, M. (2010). Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS Biol*, **8**, e1000329. 27, 31, 32, 33, 130, 192, 197, 234, 245, 246, 250, 258, 270, 272

Le, S.Q. & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol*, **25**, 1307–20. 80

Le, S.Q. & Gascuel, O. (2010). Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst Biol*, **59**, 277–87. 54

Le, S.Q., Lartillot, N. & Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philos Trans R Soc Lond B Biol Sci*, **363**, 3965–3976. 56, 252

Lee, J.S., Smith, E. & Shilatifard, A. (2010a). The language of histone crosstalk. *Cell*, **142**, 682–5. 3

Lee, S.C., Ni, M., Li, W., Shertz, C. & Heitman, J. (2010b). The evolution of sex: a perspective from the fungal kingdom. *Microbiol Mol Biol Rev*, **74**, 298–340. 113, 114

Lee, T.A., Jorgensen, P., Bognar, A.L., Peyraud, C., Thomas, D. & Tyers, M. (2010c). Dissection of combinatorial control by the met4 transcriptional complex. *Mol Biol Cell*, **21**, 456–69. 9, 222, 261, 268, 270

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. & Young, R.A. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, **298**, 799–804. 15, 118, 256

Lee, W., Tillo, D., Bray, N., Morse, R.H., Davis, R.W., Hughes, T.R. & Nislow, C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet*, **39**, 1235–44. 3, 4

Lempiäinen, H. & Shore, D. (2009). Growth control and ribosome biogenesis. *Curr Opin Cell Biol*, **21**, 855–63. 244

Lespinet, O., Wolf, Y.I., Koonin, E.V. & Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res*, **12**, 1048–59. 200

Levine, M. & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, **424**, 147–151. 13

Li, C., Lu, G. & Ortí, G. (2008). Optimal data partitioning and a test case for ray-finned fishes (actinopterygii) based on ten nuclear loci. *Syst Biol*, **57**, 519–39. 253, 254

Li, C., Li, X., Miao, Y., Wang, Q., Jiang, W., Xu, C., Li, J., Han, J., Zhang, F., Gong, B. & Xu, L. (2009). Subpathwayminer: a software package for flexible identification of pathways. *Nucleic Acids Res*, **37**, e131. 233

Li, F.N. & Johnston, M. (1997). Grr1 of saccharomyces cerevisiae is connected to the ubiquitin proteolysis machinery through skp1: coupling glucose sensing to gene expression and the cell cycle. *EMBO J*, **16**, 5629–38. 261

Li, L., Stoeckert, C.J., Jr & Roos, D.S. (2003). Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res*, **13**, 2178–89. 150

Li, L., Huang, Y., Xia, X. & Sun, Z. (2006). Preferential duplication in the sparse part of yeast protein interaction network. *Mol Biol Evol*, **23**, 2467–73. 152, 182

Liao, B.Y. & Zhang, J. (2006). Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol*, **23**, 1119–28. 27

Lin, Z. & Li, W.H. (2010). Expansion of hexose transporter genes was associated with the evolution of aerobic fermentation in yeasts. *Mol Biol Evol*. 47

Lippman, S.I. & Broach, J.R. (2009). Protein kinase a and torc1 activate genes for ribosomal biogenesis by inactivating repressors encoded by dot6 and its homolog tod6. *Proc Natl Acad Sci U S A*, **106**, 19928–33. 245

Loch, C.M., Mosammaparast, N., Miyake, T., Pemberton, L.F. & Li, R. (2004). Functional and physical interactions between autonomously replicating sequence-binding factor 1 and the nuclear transport machinery. *Traffic*, **5**, 925–35. 120

Lockhart, P., Steel, M., Hendy, M. & Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence. *Mol Biol Evol*, **11**, 605–612. 54

Lohse, M.B. & Johnson, A.D. (2009). White-opaque switching in candida albicans. *Curr Opin Microbiol*, **12**, 650–4. 36, 37

Lopez, P., Casane, D. & Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Mol Biol Evol*, **19**, 1–7. 54

Löytynoja, A. & Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A*, **102**, 10557–62. 142

Löytynoja, A. & Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–1635. 55, 142

Lozada-Chávez, I., Janga, S.C. & Collado-Vides, J. (2006). Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res*, **34**, 3434–45. 202

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, **389**, 251–60. 3

Luscombe, N.M., Austin, S.E., Berman, H.M. & Thornton, J.M. (2000). An overview of the structures of protein-dna complexes. *Genome Biol*, **1**, REVIEWS001. 7, 8, 9

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A. & Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–12. 25, 118, 199, 271

Lynch, V.J. & Wagner, G.P. (2008). Resurrecting the role of transcription factor change in developmental evolution. *Evolution*, **62**, 2131–54. xii, 24, 25, 35

Ma, H.W., Buer, J. & Zeng, A.P. (2004). Hierarchical structure and modules in the escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics*, **5**, 199. 18

Ma, J. & Ptashne, M. (1987). Deletion analysis of gal4 defines two transcriptional activating segments. *Cell*, **48**, 847–53. 225

MacPherson, S., Larochelle, M. & Turcotte, B. (2006). A fungal family of transcriptional regulators: the zinc cluster proteins. *Microbiol Mol Biol Rev*, **70**, 583–604. 110, 115

Madera, M. (2008). Profile comparer: a program for scoring and aligning profile hidden markov models. *Bioinformatics*, **24**, 2630–2631. 149

Maduro, M. & Pilgrim, D. (1996). Conservation of function and expression of unc-119 from two caenorhabditis species despite divergence of non-coding dna. *Gene*, **183**, 77–85. 28

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. & Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A*, **102**, 5454–9. 182

Mangan, S. & Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A*, **100**, 11980–5. 16

Mangelsdorf, D.J., Thummel, C., Beato, M., Herrlich, P., Schütz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P. & Evans, R.M. (1995). The nuclear receptor superfamily: the second decade. *Cell*, **83**, 835–9. 11

Marcet-Houben, M. & Gabaldón, T. (2009). The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. *PLoS One*, **4**. 55, 56, 57, 58, 252

Martchenko, M., Levitin, A., Hogues, H., Nantel, A. & Whiteway, M. (2007). Transcriptional rewiring of fungal galactose-metabolism circuitry. *Curr Biol*, **17**, 1007–13. 31, 33, 137

Martin, D.E., Soulard, A. & Hall, M.N. (2004). Tor regulates ribosomal protein gene expression via pka and the forkhead transcription factor fhl1. *Cell*, **119**, 969–79. 31, 244

Merico, A., Sulo, P., Piskur, J. & Compagno, C. (2007). Fermentative lifestyle in yeasts belonging to the saccharomyces complex. *FEBS J*, **274**, 976–89. xii, 45, 46, 47, 122, 123, 125, 129, 197, 235, 241, 248, 249, 257, 259, 260

Messenguy, F. & Dubois, E. (2003). Role of mads box proteins and their cofactors in combinatorial control of gene expression and cell development. *Gene*, **316**, 1–21. 38

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–7. 15

Minezaki, Y., Homma, K. & Nishikawa, K. (2005). Genome-wide survey of transcription factors in prokaryotes reveals many bacteria-specific families not found in archaea. *DNA Res*, **12**, 269–80. 13, 14

Mintseris, J. & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A*, **102**, 10930–5. 202

Miranda, I., Silva, R. & Santos, M.A.S. (2006). Evolution of the genetic code in yeasts. *Yeast*, **23**, 203–13. 43

Mitsuda, N. & Ohme-Takagi, M. (2009). Functional analysis of transcription factors in arabidopsis. *Plant Cell Physiol*, **50**, 1232–1248. 92

Miyake, T., Loch, C.M. & Li, R. (2002). Identification of a multifunctional domain in autonomously replicating sequence-binding factor 1 required for transcriptional activation, dna replication, and gene silencing. *Mol Cell Biol*, **22**, 505–16. 120

Miyake, T., Reese, J., Loch, C.M., Auble, D.T. & Li, R. (2004). Genome-wide analysis of ars (autonomously replicating sequence) binding factor 1 (abf1p)-mediated transcriptional regulation in saccharomyces cerevisiae. *J Biol Chem*, **279**, 34865–34872. 111, 118

Miyazaki, T., Yamauchi, S., Inamine, T., Nagayoshi, Y., Saijo, T., Izumikawa, K., Seki, M., Kakeya, H., Yamamoto, Y., Yanagihara, K., Miyazaki, Y. & Kohno, S. (2010). Roles of calcineurin and crz1 in antifungal susceptibility and virulence of candida glabrata. *Antimicrob Agents Chemother*, **54**, 1639–43. 133, 197

Montcalm, L. & Wolfe, K. (2006). Genome evolution in hemiascomycete yeasts. In K. Esser & A. Brown, eds., *Fungal Genomics*, vol. 13 of *The Mycota*, 19–34, Springer Berlin Heidelberg. 43

Morse, R.H. (2000). Rap, rap, open up! new wrinkles for rap1 in yeast. *Trends Genet*, **16**, 51–3. 5

Moses, A.M., Pollard, D.A., Nix, D.A., Iyer, V.N., Li, X.Y., Biggin, M.D. & Eisen, M.B. (2006). Large-scale turnover of functional transcription factor binding sites in drosophila. *PLoS Comput Biol*, **2**, e130. 27

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigó, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J., Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott,

D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K., McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J.P., Von Niederhausern, A.C., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.P., Zdobnov, E.M., Zody, M.C. & Lander, E.S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–62. 1

Neduva, V. & Russell, R.B. (2005). Linear motifs: evolutionary interaction switches. *FEBS Lett*, **579**, 3342–5. 26

Neil, H., Lemaire, M. & Wésolowski-Louvel, M. (2004). Regulation of glycolysis in kluyveromyces lactis: role of klgcr1 and klgcr2 in glucose uptake and catabolism. *Curr Genet*, **45**, 129–39. 126, 129

Neil, H., Hnatova, M., Wésolowski-Louvel, M., Rycovska, A. & Lemaire, M. (2007). Sck1 activator coordinates glucose transport and gly-

colysis and is controlled by rag8 casein kinase i in kluyveromyces lactis. *Mol Microbiol*, **63**, 1537–1548. 126

Newcomb, L.L., Hall, D.D. & Heideman, W. (2002). Azf1 is a glucose-dependent positive regulator of cln3 transcription in saccharomyces cerevisiae. *Mol Cell Biol*, **22**, 1607–14. 129

Nishihara, H., Okada, N. & Hasegawa, M. (2007). Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biol*, **8**, R199. 51, 55, 56, 83, 252, 253

Nowick, K. & Stubbs, L. (2010). Lineage-specific transcription factors and the evolution of gene regulatory networks. *Brief Funct Genomics*, **9**, 65–78. xi, 17, 92, 200

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W., MacIsaac, K.D., Rolfe, P.A., Conboy, C.M., Gifford, D.K. & Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*, **39**, 730–732. 27, 28, 29

Onuma, Y., Takahashi, S., Asashima, M., Kurata, S. & Gehring, W.J. (2002). Conservation of pax 6 function and upstream activation by notch signaling in eye development of frogs and flies. *Proc Natl Acad Sci U S A*, **99**, 2020–5. 20

Osada, N., Kohn, M.H. & Wu, C.I. (2006). Genomic inferences of the cis-regulatory nucleotide polymorphisms underlying gene expression differences between drosophila melanogaster mating races. *Mol Biol Evol*, **23**, 1585–91. 22

Pace, J.K., 2nd, Gilbert, C., Clark, M.S. & Feschotte, C. (2008). Repeated horizontal transfer of a dna transposon in mammals and other tetrapods. *Proc Natl Acad Sci U S A*, **105**, 17023–8. 113

Pagel, M. & Meade, A. (2008). Modelling heterotachy in phylogenetic inference by reversible-jump markov chain monte carlo. *Philos Trans R Soc Lond B Biol Sci*, **363**, 3955–3964. 54, 56

Pál, C., Papp, B. & Lercher, M.J. (2006). An integrated view of protein evolution. *Nat Rev Genet*, **7**, 337–48. 201, 202

Pan, Y., Tsai, C.J., Ma, B. & Nussinov, R. (2009). How do transcription factors select specific binding sites in the genome? *Nat Struct Mol Biol*, **16**, 1118–20. 12, 24

Pan, Y., Tsai, C.J., Ma, B. & Nussinov, R. (2010). Mechanisms of transcription factor selectivity. *Trends Genet*, **26**, 75–83. 12

Papp, B., Pál, C. & Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, **424**, 194–7. 130, 182

Park, J., Park, J., Jang, S., Kim, S., Kong, S., Choi, J., Ahn, K., Kim, J., Lee, S., Kim, S., Park, B., Jung, K., Kim, S., Kang, S. & Lee, Y.H. (2008). Ftfd: an informatics pipeline supporting phylogenomic analysis of fungal transcription factors. *Bioinformatics*, **24**, 1024–5. 41

Pavletich, N.P. & Pabo, C.O. (1993). Crystal structure of a five-finger gli-dna complex: new perspectives on zinc fingers. *Science*, **261**, 1701–7. 7

Peña, M.M., Lee, J. & Thiele, D.J. (1999). A delicate balance: homeostatic control of copper uptake and distribution. *J Nutr*, **129**, 1251–60. 264, 265

Pennisi, E. (2005). Evolution 2005 meeting. wine yeast's surprising diversity. *Science*, **309**, 375–6. 44

Pérez-Rueda, E. & Janga, S.C. (2010). Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. *Mol Biol Evol*, **27**, 1449–59. 11, 92

Pérez-Rueda, E., Collado-Vides, J. & Segovia, L. (2004). Phylogenetic distribution of dna-binding transcription factors in bacteria and archaea. *Comput Biol Chem*, **28**, 341–50. 13, 14, 41, 91, 92, 255

Pfaller, M.A. & Diekema, D.J. (2007). Epidemiology of invasive candidiasis: a persistent public health problem. *Clin Microbiol Rev*, **20**, 133–63. 44

Philippe, H. & Douady, C.J. (2003). Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol*, **6**, 498–505. 55

Philippe, H., Delsuc, F., Brinkmann, H. & Lartillot, N. (2005). Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 541–562. 51, 55

Phillips, M.J., Delsuc, F. & Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol*, **21**, 1455–1458. 55, 56, 60, 86

Piskur, J. (1994). Inheritance of the yeast mitochondrial genome. *Plasmid*, **31**, 229–41. 122

Piskur, J., Rozpedowska, E., Polakova, S., Merico, A. & Compagno, C. (2006). How did saccharomyces evolve to become a good brewer? *Trends Genet*, **22**, 183–6. 46, 47

Posada, D. & Buckley, T.R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol*, **53**, 793–808. 66

Posada, D. & Crandall, K.A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol*, **54**, 396–402. 55

Postma, E., Verduyn, C., Scheffers, W.A. & Van Dijken, J.P. (1989). Enzymic analysis of the crabtree effect in glucose-limited chemostat cultures of saccharomyces cerevisiae. *Appl Environ Microbiol*, **55**, 468–77. 122

Price, M.N., Dehal, P.S. & Arkin, A.P. (2007). Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Comput Biol*, **3**, 1739–50. 202

Price, M.N., Dehal, P.S. & Arkin, A.P. (2008). Horizontal gene transfer and the evolution of transcriptional regulation in escherichia coli. *Genome Biol*, **9**, R4. 40, 111

Prud'homme, B., Gompel, N., Rokas, A., Kassner, V.A., Williams, T.M., Yeh, S.D., True, J.R. & Carroll, S.B. (2006). Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, **440**, 1050–3. 20

Prud'homme, B., Gompel, N. & Carroll, S.B. (2007). Emerging principles of regulatory evolution. *Proc Natl Acad Sci U S A*, **104 Suppl 1**, 8605–12. 20, 24

Raijman, D., Shamir, R. & Tanay, A. (2008). Evolution and selection in yeast promoters: analyzing the combined effect of diverse transcription factor binding sites. *PLoS Comput Biol*, **4**, e7. 27, 28

Rasmussen, M.D. & Kellis, M. (2010). A bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol*. 141, 142, 150, 152, 153, 156, 159, 164, 165, 166, 167

Reed, S.H., Akiyama, M., Stillman, B. & Friedberg, E.C. (1999). Yeast autonomously replicating sequence binding factor is involved in nucleotide excision repair. *Genes Dev*, **13**, 3052–8. 118

Ren, F., Tanaka, H. & Yang, Z. (2008). A likelihood look at the supermatrix-supertree controversy. *Gene*, **441**, 119–125. 53, 56

Rhode, P.R., Sweder, K.S., Oegema, K.F. & Campbell, J.L. (1989). The gene encoding ars-binding factor i is essential for the viability of yeast. *Genes Dev*, **3**, 1926–39. 118

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D.,

Sherman, B.K. & Yu, G. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110. 13, 14, 41, 91, 92, 108, 110, 111, 255

Rigoulet, M., Aguilaniu, H., Avéret, N., Bunoust, O., Camougrand, N., Grandier-Vazeille, X., Larsson, C., Pahlman, I.L., Manon, S. & Gustafsson, L. (2004). Organization and regulation of the cytosolic nadh metabolism in the yeast saccharomyces cerevisiae. *Mol Cell Biochem*, **256-257**, 73–81. 259

Rine, J. & Herskowitz, I. (1987). Four genes responsible for a position effect on expression from hml and hmr in saccharomyces cerevisiae. *Genetics*, **116**, 9–22. 264

Ripplinger, J. & Sullivan, J. (2008). Does choice in model selection affect maximum likelihood analysis? *Syst Biol*, **57**, 76–85. 67, 89

Ripplinger, J. & Sullivan, J. (2010). Assessment of substitution model adequacy using frequentist and bayesian methods. *Mol Biol Evol*, **27**, 2790–803. 254

Robbins, H. (1956). *An empirical Bayes approach to statistics.*, 157–163. Proc. 3rd Berkeley Sympos. Math. Statist. Probability 1. 152

Robinson, D.F. & Foulds, L.R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147. 67

Rodrigues-Pousada, C., Menezes, R.A. & Pimentel, C. (2010). The yap family and its role in stress response. *Yeast*, **27**, 245–58. 41

Rodríguez, F., Oliver, J.L., Marín, A. & Medina, J.R. (1990). The general stochastic model of nucleotide substitution. *J Theor Biol*, **142**, 485–501. 63, 204

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F. & Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol*, **56**, 389–399. 55, 56

Roetzer, A., Gabaldón, T. & Schüller, C. (2010). From saccharomyces cerevisiaeto candida glabratain a few easy steps: important adaptations for an opportunistic pathogen. *FEMS Microbiol Lett*. 48

Rokas, A. & Chatzimanolis, S. (2008). From gene-scale to genome-scale phylogenetics: the data flood in, but the challenges remain. *Methods Mol Biol*, **422**, 1–12. 76

Rokas, A., Williams, B.L., King, N. & Carroll, S.B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804. 52, 53, 56, 60, 61, 76, 86, 89, 90

Rolland, F., Winderickx, J. & Thevelein, J.M. (2002). Glucose-sensing and -signalling mechanisms in yeast. *FEMS Yeast Res*, **2**, 183–201. 46, 259, 261

Roth-Ben Arie, Z., Altboum, Z., Berdicevsky, I. & Segal, E. (1998). Isolation of a petite mutant from a histidine auxotroph of candida albicans and its characterization. *Mycopathologia*, **141**, 127–35. 122

Rudra, D., Zhao, Y. & Warner, J.R. (2005). Central role of ifh1p-fhl1p interaction in the synthesis of yeast ribosomal proteins. *EMBO J*, **24**, 533–42. 31, 244

Ruiz, A. & Ariño, J. (2007). Function and regulation of the saccharomyces cerevisiae ena sodium atpase system. *Eukaryot Cell*, **6**, 2175–83. 247

Rusche, L.N., Kirchmaier, A.L. & Rine, J. (2003). The establishment, inheritance, and function of silenced chromatin in saccharomyces cerevisiae. *Annu Rev Biochem*, **72**, 481–516. 3, 118, 264

Rutherford, J.C. & Bird, A.J. (2004). Metal-responsive transcription factors that regulate iron, zinc, and copper homeostasis in eukaryotic cells. *Eukaryot Cell*, **3**, 1–13. 11, 264, 265

Sabina, J. & Brown, V. (2009). Glucose sensing network in candida albicans: a sweet spot for fungal morphogenesis. *Eukaryot Cell*, **8**, 1314–20. 264

Sampaio, J.P. & Gonçalves, P. (2008). Natural populations of saccharomyces kudriavzevii in portugal are associated with oak bark and are sympatric with s. cerevisiae and s. paradoxus. *Appl Environ Microbiol*, **74**, 2144–52. 44

Santos, M.A. & Tuite, M.F. (1995). The cug codon is decoded in vivo as serine and not leucine in candida albicans. *Nucleic Acids Res*, **23**, 1481–6. 43, 80

Sato, T., Jigami, Y., Suzuki, T. & Uemura, H. (1999a). A human gene, hsgt1, can substitute for gcr2, which encodes a general regulatory factor of glycolytic gene expression in saccharomyces cerevisiae. *Mol Gen Genet*, **260**, 535–540. 126, 127

Sato, T., Lopez, M.C., Sugioka, S., Jigami, Y., Baker, H.V. & Uemura, H. (1999b). The e-box dna binding protein sgc1p suppresses the gcr2 mutation, which is involved in transcriptional activation of glycolytic genes in saccharomyces cerevisiae. *FEBS Lett*, **463**, 307–11. 126

Scannell, D.R. & Wolfe, K.H. (2008). A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res*, **18**, 137–147. 202, 215, 241, 248, 258

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S. & Wolfe, K.H. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, **440**, 341–345. 57, 86, 111, 115, 127, 135

Schierup, M.H. & Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics*, **156**, 879–891. 55

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., Talianidis, I., Flicek, P. & Odom, D.T. (2010). Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. *Science*. 27, 28, 29, 191

Schoch, C., Sung, G.H., Lopez-Giraldez, F., Townsend, J., Miadlikowska, J., Hofstetter, V., Robbertse, B., Matheny, B., Kauff, F., Wang, Z., Gueidan, C., Andrie, R., Trippe, K., Ciufetti, L., Wynns, A., Fraker, E., Hodkinson, B., Bonito, G., Groenewald, J., Arzanlou, M., Sybren de Hoog, G., Crous, P., Hewitt, D., Pfister, D., Peterson, K., Gryzenhout, M., Wingfield, M., Aptroot, A., Suh, S.O., Blackwell, M., Hillis, D., Griffith, G., Castlebury, L., Rossman, A., Lumbsch, T., Lucking, R., Budel, B., Rauhut, A., Diederich, P., Ertz, D., Geiser, D., Hosaka, K., Inderbitzin, P., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Mostert, L., Kerry, O., Sipman, H., Rogers, J., Shoemaker, R., Sugiyama, J., Summerbell, R., Untereiner, W., Johnston, P., Stenroos, S., Zuccaro, A., Dyer, P., Crittenden, P., Cole, M., Hansen, K., Trappe, J., Yahr, R., Lutzoni, F. & Spatafora, J. (2009). The ascomycota tree of life: A phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Systematic Biology*, **58**, 224–239. 56, 57, 58

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464. 66

Sekinger, E.A., Moqtaderi, Z. & Struhl, K. (2005). Intrinsic histone-dna interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell*, **18**, 735–48. 40

Sellam, A., Tebbji, F. & Nantel, A. (2009). Role of ndt80p in sterol metabolism regulation and azole resistance in candida albicans. *Eukaryot Cell*, **8**, 1174–83. 132

Sellam, A., Askew, C., Epp, E., Tebbji, F., Mullick, A., Whiteway, M. & Nantel, A. (2010). Role of transcription factor candt80p in cell separation, hyphal growth, and virulence in candida albicans. *Eukaryot Cell*, **9**, 634–44. 132, 197

Sellerio, A.L., Bassetti, B., Isambert, H. & Cosentino Lagomarsino, M. (2009). A comparative evolutionary study of transcription networks. the global role of feedback and hierachical structures. *Mol Biosyst*, **5**, 170–9. 196

Sempere, L.F., Cole, C.N., McPeek, M.A. & Peterson, K.J. (2006). The phylogenetic distribution of metazoan micrornas: insights into evolutionary complexity and constraint. *J Exp Zool B Mol Dev Evol*, **306**, 575–88. 2

SGD (2010). The saccharomyces genome database. *http://www.yeastgenome.org/*. 233

Shapiro, M.D., Marks, M.E., Peichel, C.L., Blackman, B.K., Nereng, K.S., Jónsson, B., Schluter, D. & Kingsley, D.M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, **428**, 717–23. 20

Sharp, P.M. & Li, W.H. (1987). The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, **15**, 1281–95. 207

Shelest, E. (2008). Transcription factors in fungi. *FEMS Microbiol Lett*, **286**, 145–151. 13, 14, 41, 92, 93, 110, 114, 255

Shiu, S.H., Shih, M.C. & Li, W.H. (2005). Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiol*, **139**, 18–26. 13, 14, 91, 92

Shore, D. (1994). Rap1: a protean regulator in yeast. *Trends Genet*, **10**, 408–12. 5

Silve, S., Rhode, P.R., Coll, B., Campbell, J. & Poyton, R.O. (1992). Abf1 is a phosphoprotein and plays a role in carbon source control of cox6 transcription in saccharomyces cerevisiae. *Mol Cell Biol*, **12**, 4197–208. 123

Singh, G.P. & Dash, D. (2007). Intrinsic disorder in yeast transcriptional regulatory network. *Proteins*, **68**, 602–5. 207, 213

Smets, B., Ghillebert, R., De Snijder, P., Binda, M., Swinnen, E., De Virgilio, C. & Winderickx, J. (2010). Life in the midst of scarcity: adaptations to nutrient availability in saccharomyces cerevisiae. *Curr Genet*, **56**, 1–32. 261, 262

Smolka, M.B., Albuquerque, C.P., Chen, S.h. & Zhou, H. (2007). Proteome-wide identification of in vivo targets of dna damage checkpoint kinases. *Proc Natl Acad Sci U S A*, **104**, 10364–9. 125

Soll, D.R., Lockhart, S.R. & Zhao, R. (2003). Relationship between switching and mating in candida albicans. *Eukaryot Cell*, **2**, 390–397. 114

Soll, D.R., Pujol, C. & Srikantha, T. (2009). Sex: deviant mating in yeast. *Curr Biol*, **19**, R509–11. 35, 114

Spoel, S.H., Tada, Y. & Loake, G.J. (2010). Post-translational protein modification as a tool for transcription reprogramming. *New Phytol*, **186**, 333–9. 12

Stajich, J.E., Berbee, M.L., Blackwell, M., Hibbett, D.S., James, T.Y., Spatafora, J.W. & Taylor, J.W. (2009). The fungi. *Curr Biol*, **19**, R840–5. 42

Stern, D.L. (2000). Evolutionary developmental biology and the problem of variation. *Evolution*, **54**, 1079–91. 20, 21, 24

Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–55. 27

Sugita, T. & Nakase, T. (1999). Nonuniversal usage of the leucine cug codon in yeasts: Investigation of basidiomycetous yeast. *J Gen Appl Microbiol*, **45**, 193–197. 80

Suh, S.O., Blackwell, M., Kurtzman, C.P. & Lachance, M.A. (2006). Phylogenetics of saccharomycetales, the ascomycete yeasts. *Mycologia*, **98**, 1006–17. 42, 43, 44

Sullivan, J. & Swofford, D.L. (1997). Are guinea pigs rodents? the importance of adequate models in molecular phylogenetics. *Journal of Mammalian Evolution*, **4**, 77–86. 60, 63, 67, 89

Sung, H.M., Wang, T.Y., Wang, D., Huang, Y.S., Wu, J.P., Tsai, H.K., Tzeng, J., Huang, C.J., Lee, Y.C., Yang, P., Hsu, J., Chang, T., Cho, C.Y., Weng, L.C., Lee, T.C., Chang, T.H., Li, W.H. & Shih, M.C. (2009). Roles of trans and cis variation in yeast intraspecies evolution of gene expression. *Mol Biol Evol*, **26**, 2533–8. 22, 23

Tadepally, H.D., Burger, G. & Aubry, M. (2008). Evolution of c2h2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. *BMC Evol Biol*, **8**, 176–176. 93, 100

Tan, K., Feizi, H., Luo, C., Fan, S.H., Ravasi, T. & Ideker, T.G. (2008). A systems approach to delineate functions of paralogous transcription factors: role of the yap family in the dna damage response. *Proc Natl Acad Sci U S A*, **105**, 2934–9. 39

Tanay, A., Regev, A. & Shamir, R. (2005). Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A*, **102**, 7203–7208. 27, 31, 32, 41, 100

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. In *Lectures on Mathematics in the Life Sciences*, vol. 17, 57–86. 63

Taylor, J.W. & Berbee, M.L. (2006). Dating divergences in the fungal tree of life: review and new analyses. *Mycologia*, **98**, 838–849. 43

Teichmann, S.A. & Babu, M.M. (2004). Gene regulatory network growth by duplication. *Nat Genet*, **36**, 492–6. 38

Teixeira, M.C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A.R., Mira, N.P., Alenquer, M., Freitas, A.T., Oliveira, A.L. & Sá-Correia, I. (2006). The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. *Nucleic Acids Res*, **34**, D446–51. 183, 207

Thomas, J.H. & Emerson, R.O. (2009). Evolution of c2h2-zinc finger genes revisited. *BMC Evol Biol*, **9**, 51. 41, 93, 100

Thomas, M.C. & Chiang, C.M. (2006). The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol*, **41**, 105–78. 5

Thompson, D.A. & Regev, A. (2009). Fungal regulatory evolution: cis and trans in the balance. *FEBS Lett*, **583**, 3959–3965. 24, 250

Thorne, J.L., Goldman, N. & Jones, D.T. (1996). Combining protein evolution and secondary structure. *Mol Biol Evol*, **13**, 666–673. 54

Tirosh, I., Reikhav, S., Levy, A.A. & Barkai, N. (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*, **324**, 659–662. 22, 23, 25

Tsankov, A.M., Thompson, D.A., Socha, A., Regev, A. & Rando, O.J. (2010). The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol*, **8**, e1000414. 40

Tsong, A.E., Miller, M.G., Raisner, R.M. & Johnson, A.D. (2003). Evolution of a combinatorial transcriptional circuit: a case study in yeasts. *Cell*, **115**, 389–99. 35, 36

Tsong, A.E., Tuch, B.B., Li, H. & Johnson, A.D. (2006). Evolution of alternative transcriptional circuits with identical logic. *Nature*, **443**, 415–20. xii, 35, 36, 114

Tsui, C.K., Daniel, H.M., Robert, V. & Meyer, W. (2008). Re-examining the phylogeny of clinically relevant candida species and allied genera based on multigene analyses. *FEMS Yeast Res*, **8**, 651–659. 56, 57, 58

Tu, B.P., Mohler, R.E., Liu, J.C., Dombek, K.M., Young, E.T., Synovec, R.E. & McKnight, S.L. (2007). Cyclic changes in metabolic state during the life of a yeast cell. *Proc Natl Acad Sci U S A*, **104**, 16886–91. 46

Tuch, B.B., Galgoczy, D.J., Hernday, A.D., Li, H. & Johnson, A.D. (2008a). The evolution of combinatorial gene regulation in fungi. *PLoS Biol*, **6**, e38. 24, 35, 37, 38, 192, 246, 258, 270, 272

Tuch, B.B., Li, H. & Johnson, A.D. (2008b). Evolution of eukaryotic transcription circuits. *Science*, **319**, 1797–9. xii, 24, 25, 35

Tuch, B.B., Mitrovich, Q.M., Homann, O.R., Hernday, A.D., Monighetti, C.K., De La Vega, F.M. & Johnson, A.D. (2010). The transcriptomes of two heritable cell types illuminate the circuit governing their differentiation. *PLoS Genet*, **6**, e1001070. 36

Uemura, H. & Jigami, Y. (1995). Mutations in gcr1, a transcriptional activator of saccharomyces cerevisiae glycolytic genes, function as suppressors of gcr2 mutations. *Genetics*, **139**, 511–521. 126

Upton, T., Wiltshire, S., Francesconi, S. & Eisenberg, S. (1995). Abf1 ser-720 is a predominant phosphorylation site for casein kinase ii of saccharomyces cerevisiae. *J Biol Chem*, **270**, 16153–9. 120, 121

Uzzell, T. & Corbin, K.W. (1971). Fitting discrete probability distributions to evolutionary events. *Science*, **172**, 1089–1096. 54

Van de Peer, Y., Maere, S. & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat Rev Genet*, **10**, 725–32. 182, 200

van Dyk, D., Pretorius, I.S. & Bauer, F.F. (2005). Mss11p is a central element of the regulatory network that controls flo11 expression and invasive growth in saccharomyces cerevisiae. *Genetics*, **169**, 91–106. 130

van Hoek, M.J.A. & Hogeweg, P. (2009). Metabolic adaptation after whole genome duplication. *Mol Biol Evol*, **26**, 2441–53. 47

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet*, **19**, 479–84. 13, 91

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, **10**, 252–263. 13, 14, 41, 91, 92, 94, 108

Venditti, P., Costanzo, G., Negri, R. & Camilloni, G. (1994). Abfi contributes to the chromatin organization of saccharomyces cerevisiae ars1 b-domain. *Biochim Biophys Acta*, **1219**, 677–89. 118

Venters, B.J. & Pugh, B.F. (2009). How eukaryotic genes are transcribed. *Crit Rev Biochem Mol Biol*, **44**, 117–41. xi, 4, 5, 6

Vershon, A.K. & Pierce, M. (2000). Transcriptional regulation of meiosis in yeast. *Curr Opin Cell Biol*, **12**, 334–9. 233

Vilela-Moura, A., Schuller, D., Mendes-Faia, A., Silva, R.D., Chaves, S.R., Sousa, M.J. & Côrte-Real, M. (2010). The impact of acetate metabolism on yeast fermentative performance and wine quality: reduction of volatile acidity of grape musts and wines. *Appl Microbiol Biotechnol*. 267

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. & Birney, E. (2009). Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, **19**, 327–35. xviii, 152, 156, 157

Vissing, H., Meyer, W.K., Aagaard, L., Tommerup, N. & Thiesen, H.J. (1995). Repression of transcriptional activity by heterologous krab domains present in zinc finger proteins. *FEBS Lett*, **369**, 153–7. 92

Wagner, A. & Wright, J. (2007). Alternative routes and mutational robustness in complex regulatory networks. *Biosystems*, **88**, 163–72. 202, 225, 226, 227, 229

Wall, D.P., Fraser, H.B. & Hirsh, A.E. (2003). Detecting putative orthologs. *Bioinformatics*, **19**, 1710–1. 150, 204

Wang, D., Sung, H.M., Wang, T.Y., Huang, C.J., Yang, P., Chang, T., Wang, Y.C., Tseng, D.L., Wu, J.P., Lee, T.C., Shih, M.C. & Li, W.H. (2007). Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors. *Genome Res*, **17**, 1161–9. 22

Wang, H., Xu, Z., Gao, L. & Hao, B. (2009). A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evol Biol*, **9**, 195–195. 56, 57, 58

Wang, W., Carey, M. & Gralla, J.D. (1992). Polymerase ii promoter activation: closed complex formation and atp-driven start site opening. *Science*, **255**, 450–3. 6

Wang, Y., Franzosa, E.A., Zhang, X.S. & Xia, Y. (2010). Protein evolution in yeast transcription factor subnetworks. *Nucleic Acids Res*. 202, 207, 210, 213, 223, 224, 225, 226, 227, 228, 229, 255, 258

Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. (2007a). Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, **23**, i549–i558. xv, 61, 95, 97, 98, 140, 147, 149, 164

Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. (2007b). Natural history and evolutionary principles of gene duplication in fungi. *Nature*, **449**, 54–61. xvii, 141, 171, 182

Wapinski, I., Pfiffner, J., French, C., Socha, A., Thompson, D.A. & Regev, A. (2010). Gene duplication and the evolution of ribosomal protein gene regulation in yeast. *Proc Natl Acad Sci U S A*, **107**, 5505–10. 39, 41

Ward, P.S., Brady, S.G., Fisher, B.L. & Schultz, T.R. (2010). Phylogeny and biogeography of dolichoderine ants: effects of data partitioning and relict taxa on historical inference. *Syst Biol*, **59**, 342–62. 253

Weaklim, D.L. (1999). A critique of the bayesian information criterion for model selection. *Sociological Methods Research*, **27**, 359–397. 66, 254

Weiner, A., Hughes, A., Yassour, M., Rando, O.J. & Friedman, N. (2010). High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res*, **20**, 90–100. 40

Weirauch, M.T. & Hughes, T.R. (2010). Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet*, **26**, 66–74. 27

Wendland, J. & Walther, A. (2005). Ashbya gossypii: a model for fungal developmental biology. *Nat Rev Microbiol*, **3**, 421–9. 44

Westholm, J.O., Nordberg, N., Murén, E., Ameur, A., Komorowski, J. & Ronne, H. (2008). Combinatorial control of gene expression by the three yeast repressors mig1, mig2 and mig3. *BMC Genomics*, **9**, 601. xxi, 261, 262, 263

Whelan, S. (2007). New approaches to phylogenetic tree search and their application to large numbers of protein alignments. *Syst Biol*, **56**, 727–740. 68, 151

Whelan, S. (2008). Spatial and temporal heterogeneity in nucleotide sequence evolution. *Mol Biol Evol*, **25**, 1683–1694. 54

Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*, **18**, 691–699. 80, 151

Whelan, S., de Bakker, P.I.W. & Goldman, N. (2003). Pandit: a database of protein and associated nucleotide domains with inferred trees. *Bioinformatics*, **19**, 1556–1563. 77

Wilson, D., Charoensawan, V., Kummerfeld, S.K. & Teichmann, S.A. (2008a). Dbd–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res*, **36**, 88–92. xv, 6, 7, 93, 105, 106, 255, 275

Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L., Tybulewicz, V.L., Fisher, E.M., Tavaré, S. & Odom, D.T. (2008b). Species-specific transcription in mice carrying human chromosome 21. *Science*, **322**, 434–438. 22

Wittkopp, P.J., Haerum, B.K. & Clark, A.G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature*, **430**, 85–8. 22, 23

Wittkopp, P.J., Haerum, B.K. & Clark, A.G. (2008). Regulatory changes underlying expression differences within and between drosophila species. *Nat Genet*, **40**, 346–50. 23

Wohlbach, D.J., Thompson, D.A., Gasch, A.P. & Regev, A. (2009). From elements to modules: regulatory evolution in ascomycota fungi. *Curr Opin Genet Dev*, **19**, 571–578. 24

Wolfe, K.H. & Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708–13. 43

Woo, D.K., Phang, T.L., Trawick, J.D. & Poyton, R.O. (2009). Multiple pathways of mitochondrial-nuclear communication in yeast: intergenomic signaling involves abf1 and affects a different set of genes than retrograde regulation. *Biochim Biophys Acta*, **1789**, 135–45. 123

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*, **8**, 206–216. 2, 20, 21, 24

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V. & Romano, L.A. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, **20**, 1377–419. xi, 4, 6, 20, 21, 24

Wu, M. & Eisen, J.A. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*, **9**, R151. 252

Xia, X. (2007). An improved implementation of codon adaptation index. *Evol Bioinform Online*, **3**, 53–8. 207

Xia, Y., Franzosa, E.A. & Gerstein, M.B. (2009). Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Comput Biol*, **5**, e1000413. 201, 202, 207

Xie, D., Chen, C.C., Ptaszek, L.M., Xiao, S., Cao, X., Fang, F., Ng, H.H., Lewin, H.A., Cowan, C. & Zhong, S. (2010). Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res*, **20**, 804–15. 31

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *J Mol Evol*, **39**, 306–314. 54, 63, 67, 68

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press. 67, 68

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**, 1586–1591. 64, 68, 79, 83, 152, 204

Yang, Z. (2009). Paml: Phylogenetic analysis by maximum likelihood. **http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf**. 63

Yang, Z., Goldman, N. & Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol Biol Evol*, **11**, 316–324. 60, 63, 82, 89

Ye, C., Galbraith, S.J., Liao, J.C. & Eskin, E. (2009). Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast. *PLoS Comput Biol*, **5**, e1000311. 23

Yu, H. & Gerstein, M. (2006). Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci U S A*, **103**, 14724–14731. 18, 19, 181

Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J. & Rando, O.J. (2005). Genome-scale identification of nucleosome positions in s. cerevisiae. *Science*, **309**, 626–30. 3

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. & Kruglyak, L. (2003). Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nat Genet*, **35**, 57–64. 22

Zachariae, W., Kuger, P. & Breunig, K.D. (1993). Glucose repression of lactose/galactose metabolism in kluyveromyces lactis is determined by the concentration of the transcriptional activator lac9 (k1gal4) [corrected]. *Nucleic Acids Res*, **21**, 69–77. 218, 240

Zaim, J., Speina, E. & Kierzek, A.M. (2005). Identification of new genes regulated by the crt1 transcription factor, an effector of the dna damage checkpoint pathway in saccharomyces cerevisiae. *J Biol Chem*, **280**, 28–37. 130, 133

Zeitlinger, J., Simon, I., Harbison, C.T., Hannett, N.M., Volkert, T.L., Fink, G.R. & Young, R.A. (2003). Program-specific distribution of a transcription factor dependent on partner transcription factor and mapk signaling. *Cell*, **113**, 395–404. 130

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L.M. & Snyder, M. (2010). Genetic analysis of variation in transcription factor binding in yeast. *Nature*, **464**, 1187–91. 27, 29

Ziegler, E.C. & Ghosh, S. (2005). Regulating inducible transcription through controlled localization. *Sci STKE*, **2005**, re6. 11

Zmasek, C.M. & Eddy, S.R. (2001). A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, **17**, 821–828. 138