Thesis submitted for the degree of Doctor of Philosophy
at the University of Oxford

Inferring the selection on genes of yeasts, schistosomes and amniotes

Tania Wei Ling Oh

BRASENOSE COLLEGE
University of Oxford

MICHAELMAS TERM
2008

ABSTRACT

Inferring the selection on genes of yeast, schistosomes and aminiote

Tania Wei Ling Oh
Brasenose College

A thesis submitted for the degree of Doctor of Philosophy
Michaelmas Term 2008

Research presented in this thesis focuses on using evolutionary rates to detect the fingerprint of natural selection. I have undertaken four studies that consider different branches of the eukaryotic tree: *1*. Predicting positive selection in genes from five species of *Saccharomyces* genes. *2*. Studying the constraints on both synonymous and non-synonymous sites of *Saccharomyces* genes. *3*. Using evolutionary rate studies to investigate how schistosomes have adapted to their environment. *4*. Associating gene evolution with the non-linear increase between amniotes' brain and body sizes. In the first project, I investigated protein coding sequences to identify genes and their amino acids that have been subject to positive selection. I show that *Saccharomyces* genes exhibiting strong evidence for positive selection are enriched in defence and growth functional categories. In the second study, I find that a set of *Saccharomyces* genes are both constrained at the non-synonymous sites and synonymous site, presumably to ensure the correct folding of the protein, irrespective of their expression levels. I also find that the majority of yeast genes are not biased with respect to codon usage, and investigate constraints on substitution rates at synonymous sites. The third project examines the evolutionary rates of schistosome genes. I show that stage-specific genes are under weaker selective pressures than genes expressed in many life stages. I predict that the fastest evolving stage-specific genes enable the pathogen to better adapt to its ever-changing environment. In the final project of this thesis, I ask whether evolutionary rate variation for brain-expressed/specific genes correlates with brain size (allometric) change in amniotes. I find that lineages with relatively large brains have faster evolutionary rates. However, I find that there is no region of the brain in which faster evolving genes of relatively large-brained animals are preferentially expressed.

Authorship Declaration

The author, Tania Wei Ling Oh, conducted all the work described in this thesis, unless otherwise stated. Where information has been derived from other sources, the author confirms that this has been indicated in the thesis. This work was conducted at the MRC Functional Genomics (formerly Genetics) Unit, Department of Physiology, Anatomy and Genetics. No part of this thesis has been submitted for any other degree or any other university or institute of learning. The work for this degree was supervised by both Prof. Chris P. Ponting and Prof. Dame Kay E. Davies (co-supervisor).

Acknowledgements

By the grace and mercy of God, this DPhil has finally come to a close and I would like to thank those who have travelled this journey with me:

Prof. Chris P. Ponting – my main supervisor. Thank you very much for all your time, patience and understanding, especially during the most trying periods of the DPhil. Thank you for seeing me through this degree. Thank you very much also for reading my chapters and returning them so promptly. You have made me the envy of other DPhil students. Above all, thank you for giving me the chance to be in your lab. The experiences and knowledge gained from members of your lab have been very fulfilling.

Prof. Dame Kay E. Davies – my co-supervisor. Thank you for your time.

Members of the CPP lab – Thank you for everything. If I were to list all that I was grateful for, it would push the thesis over the word limit. Thank you especially to the post docs, in particular, Luis Sánchez-Pulido, without whose help with the trees in Chapter 5, the finishing of this thesis would not have been possible. I am deeply grateful (and indebted to him) for his time, guidance, patience and skill in filling gaps in my knowledge about domains and alignments. Andreas Heger who showed me what good programming was and who always generously helped me with his code and data; Caleb Webber who taught me to ask two questions instead of one and always answered them and Gerton Lunter, for his clear statistical explanations.

Dr Richard Boyd – my college advisor. Thank you for giving me the chance to be in Brasenose College. Thank you for all your support, advice and wisdom.

My family – Thank you to my parents who have supported me unconditionally, especially my dear mother who would make the 14-hour flight at the drop of a pen. Thank you so much for everything and for providing abundantly. Thank you to my husband, Adrian who never stopped believing in me and Charlotte my little daughter, who taught me discipline and whose smiles at the end of a hard day gave me hope. I am proud to be a Mondry.

Friends – thank you for all the memories. The friendships have enriched my time at Oxford. A special word thanks goes out to Adai Ramasamy for all his help with R codes. To say I am grateful would be a severe understatement.

Finally, I would like to thank and gratefully acknowledge my maternal grandmother, Soh Seok Cheng for sponsoring my DPhil at Oxford, the Tan Kah Kee foundation for their generous scholarship and the UK Medical Research Council for research equipment and financial assistance.

TABLE OF CONTENTS

ABSTRACT 2

AUTHORSHIP DECLARATION 3

ACKNOWLEDGEMENTS 4

TABLE OF CONTENTS 5

CHAPTER 1: INTRODUCTION 9

Summary 9

Introduction to yeast genomics 9

Introduction to Schistosomiasis 13
Genetics of Schistosomes 18

Introduction to amniote brain evolution 19

Introduction to sequencing methods 23
Trends in genomic analysis 23
Sequencing a genome- Whole-genome shotgun sequencing 24
Sequencing transcripts of expressed genes – EST sequencing 26

Introduction to sequence similarity and homology 28
Homologues, Orthologues and Paralogues 29

Introduction to mutation and natural selection – definitions and estimations 31
Point mutations – Transition, Transversion mutations 31
Synonymous and non-synonymous substitutions 34
Measurements of substitution rates 35

CHAPTER 2: METHODS AND MATERIALS 38

Summary 38

Sequence Similarity Tools 38
BLAST (Basic Local Alignment Search Tool) 39
PSI-BLAST 45
Exonerate 46
CLUSTALW 47

Gene prediction using GeneWise 48

Evolutionary analyses – algorithms and statistics 50
Neighbour joining 50
Bootstrapping 52

Programs used to infer positive selection **52**
Estimating selection: the ω (K_A/K_S, dN/dS) ratio. 52
CODEML 53
SLR 57
Evolver 58

General statistics used in the project **59**
Hypergeometric distribution function 59
Kolmogorov-Smirnov (K-S) test 60
Kruskal-Wallis test 60
Fisher's exact test 61
Multiple testing 61

CHAPTER 3: PREDICTING POSITIVE SELECTION AMONG *SACCHAROMYCES* GENES **63**

Summary **63**

Introduction **64**

Methods and Materials **66**
Obtaining sequence data 66
Data storage 68
Data Merger of MIT and WashU gene sets for *S. bayanus* and *S. mikatae* 68
Orthology assignments 71
Identification of truncated genes 72
Predicting positive selection 74
Simulations using Evolver 74
Functional annotation of genes 75
Finding sequences in the PDB 76
Relative solvent accessibility of positively selected residues 76

Results and Discussion **77**
Homing in on a final dataset – unreported work that led to this chapter 77
SLR predictions 80
Functional enrichments for GO-slim categories 93
Tertiary structure of the proteins of the 27 genes 101

Discussion **104**

CHAPTER 4: CONSTRAINT ON *SACCHAROMYCES* GENES **111**

Summary **111**

Introduction **111**
Codon usage bias of genes 112

Methods and Materials **115**
Gene selection and group distinction 115
Calculating mRNA abundance and translational activity 116
Estimating the codon usage bias of the genes 117

Results and Discussion **118**
Constraints at synonymous sites 118
Constraints at non-synonymous sites 122
Other possible constraints at non-synonymous sites 124
Highly expressed genes are constrained at non-synonymous sites 124
Some lowly expressed genes show constraints at non-synonymous sites 131

CHAPTER 5: SCHISTOSOME GENE EVOLUTION 136

Summary 136

Introduction 136
Nuclear receptors 137

Methods and Materials 139
Obtaining schistosome sequences 139
Clustering of the EST sequences 141
Obtaining stage specific data 142
Methods used to account for short sequences 143
Mapping of GO terms to GO-slim terms 150
Multiple testing correction 151

Results 151
Results of clustering *S. japonicum* EST and mRNA sequences 151
Do stage-specific sequences exhibit elevated dN/dS values? 152
Are fast evolving genes particularly expressed in certain stages? 161
Functional annotations for fast evolving genes 162
Annotating genes without InterPro annotations 184

Discussion 186
Enrichment within fastest evolving genes 187
Annotation of fast evolving genes 187
Vaccines and potential candidates for future drugs 191

Future work 192

CHAPTER 6: BRAIN EVOLUTION 194

Summary 194

Introduction 194

Methods and Materials 197
Determining large and small brains 197
Obtaining sequences for 1:1 orthologues 199
Calculating lineage-specific rates 201
Obtaining brain expressed and brain specific data 203

Results 205
Lineage-specific genes 205
Preferential expression 208

Discussion 209

CHAPTER 7: DISCUSSION 214
Future work using evolutionary studies 217

LIST OF ABBREVIATIONS 219

APPENDIX A 221

APPENDIX B 226
Adult whole brain 227

Fetal adult brain 228
Sub-brain tissue specific 229

REFERENCES 237

CHAPTER 1: Introduction

Summary

The focus of the research presented in this thesis is on selection. The evolutionary rates of three different groups of genes were analysed – those of yeasts, schistosomes (blood flukes or flatworms) and amniotes, specifically genes expressed in the brains. For each group, I investigated protein-coding sequences to identify genes and their amino acids that have been subject to positive or purifying selection. Most amino acid sites are under purifying selection to maintain the integrity of protein function. Mutations at these sites may be deleterious to the cell and even to the organism as a whole. Occasionally, positive selection occurs at sites where new mutations benefit a species to better adapt to its environment. Detecting these mutation rate differences between genes of different species would indicate the fingerprints of natural selection. This requires the use of various bioinformatics tools. In this chapter, to set the stage of the research carried out, I first define the background of the projects undertaken and the biological terms used throughout the thesis. The methods and tools used for this thesis are discussed in **Chapter 2**, followed by results of analyses in the following chapters.

Introduction to yeast genomics

In 1996, *Saccharomyces cerevisiae* (commonly called baker's or budding yeast) was the first eukaryotic model organism to have its genome fully sequenced [1, 2]. It was found to contain 16 chromosomes and consists of approximately 5,800 protein-coding

genes. It has one of the smallest eukaryotic genomes (1.2×10^7 base pairs (bp)), which is just 3.5 times larger than the genome of the common bacterium *Escherichia coli* [3]. Only 3.8% of yeast genes contain introns and when present, one small intron is usually close to the start of the coding sequence. *S. cerevisiae* has a compact genome. 70% of the genome is composed of genes, whilst the remainder comprises of a limited amount of repetitive DNA. *S. cerevisiae* is known to adopt a variety of morphologies, ranging from its ellipsoid shape as a unicellular organism in the laboratory, to more complex colonies forming multicellular pseudohyphae or biofilms, or 'fluffy' or 'stalk-like' structures (reviewed in [4]).

Yeast has been an excellent model for studying complex eukaryotic organisms. The advantages of using yeast as a model organism have been expounded in countless articles, papers and books (for example, [1, 2, 5-18]). Virtues such as its easy manipulation in the lab, rapid growth and non-pathogenicity, and both the stability of its haploid and diploid states (unlike most microorganisms) [7], give the experimentalist the opportunity to harvest the desired state (haploid/diploid) easily. Furthermore, the "yeast 2-hybrid system" technique has made yeast a valuable tool as it tests for the association of two proteins that are believed to interact [19].

Yeasts are also very appropriate models for evolutionary studies for five reasons. Firstly, over 95% of their genes only contain single protein-coding exons and are thus more accurately determined than those in, for example, mammals [20]. Secondly, the evolution of these unicellular organisms is not under the additional selective constraints imposed on multicellular organisms [21-23]. Thirdly, apart from chromosome III of *S. cerevisiae*, "isochores", which are extended regions of relatively

homogeneous base composition usually found in tetrapods, are not well represented in yeasts [24, 25]. These "isochores" strongly influence substitution rates at the synonymous site (third position in a codon) [26]. Fourthly, yeasts in the wild are likely to experience frequent and efficient selection of advantageous substitutions in part due to their large effective population sizes [27]. Finally, because yeast has long been used as an experimental model organism, much is known about its genes' functions.

Information from the yeast genome has been very well curated and is available from three complementary databases: *Saccharomyces* genome database (SGD) (http://genome-www.stanford.edu), Munich Information Centre for Protein Sequences (MIPS) (http://mips.gsf.de/genre/proj/yeast/), and the Yeast Proteome Database (YPD) (http://www.proteome.com). Non-redundant and non-ambiguous nomenclature about the genes, open reading frames (ORFs) and proteins are available within these databases. However, the total number of ORFs encoded by the yeast genome and the numbers of ORFs with unknown function remain to be determined. At the current count by SGD, there are 4,689 verified ORFs, 1,106 uncharacterised ORFs and 813 dubious ORFs [28].

Other yeast genomes

The fission yeast *Schizosaccharomyces pombe* has been sequenced and annotated [6]. Its 4,824 protein coding genes places it as the eukaryote with the smallest number of genes recorded. The human fungal pathogen, *Candida albicans* was sequenced at a high coverage of 10.9X [29] and a recent assembly was published [30]. *Achbya*, the cotton pathogen which has a genome size of 8.8MB has also been sequenced to near

completion [31]. Finally, the yeast genomes that have been sequenced and are used for this thesis (in **Chapter 3** and **4**) are: *Saccharomyces bayanus, S. paradoxus*, *S. kudriavzevii* and *S. mikatae* [20, 32] (**Figure 1**). These are an isolated and monophyletic group of yeasts that exhibit strong phenotypic similarities [14]. In addition, four distantly related species were also used in initial studies (**Chapter 3**); two (*S. kluyveri* and *S. castellii*) from the above monophyletic group of yeasts and the other two (*S. servazzii* and *S. exiguus*) from the 14 shotgun sequenced Hemiascomycete genomes sequenced by a French consortium [33, 34].

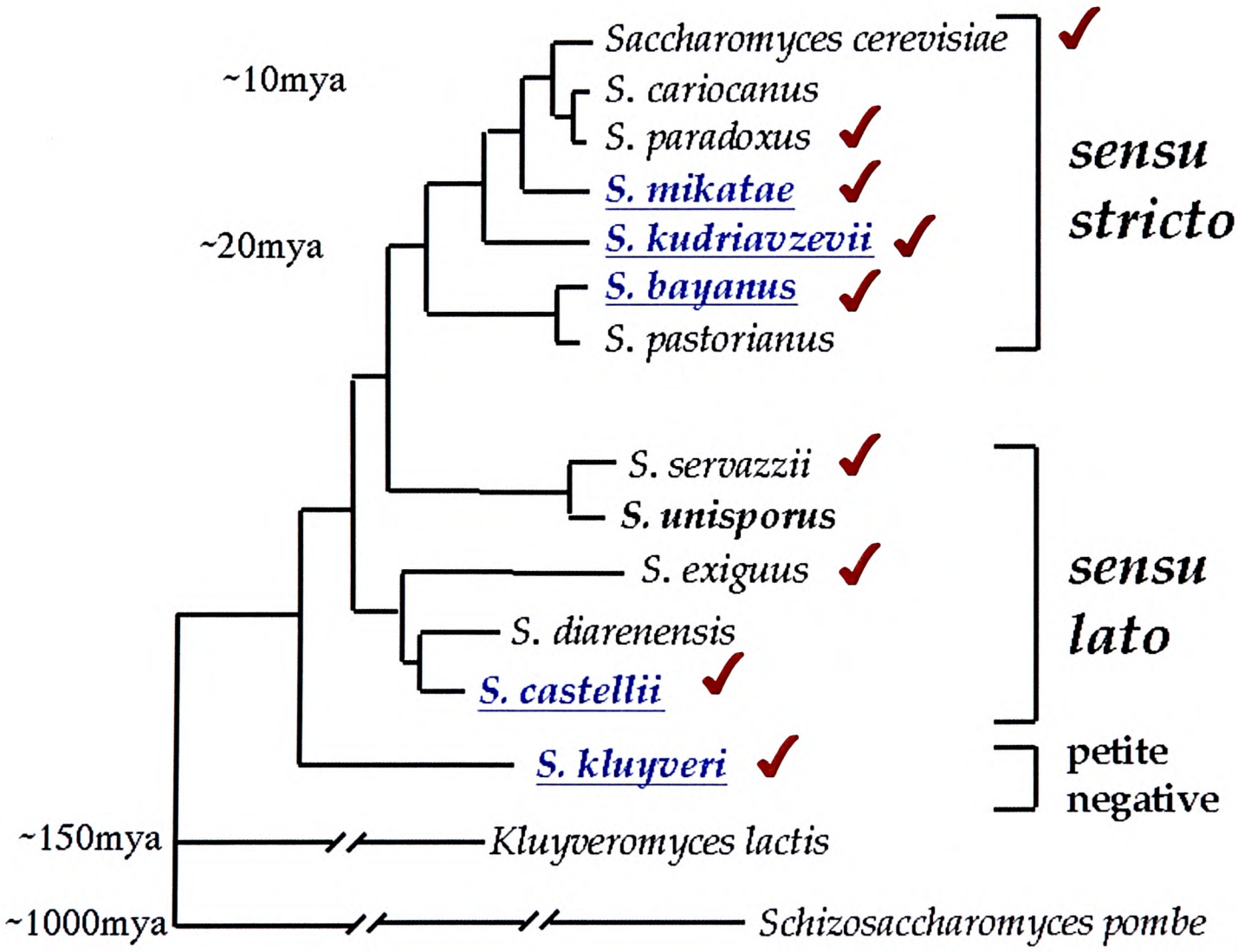


Figure 1. Phylogenetic tree of the nine species of yeast from the *Saccharomyces* genus indicating approximate time scales of evolutionary divergence. Species used in this project are indicated by a red tick next to their names. *Sensu stricto* ("in the strict sense") species are considered closely related and *sensu lato* ("in the broad sense") are more evolutionarily diverged sequences. "Petite negative" species are those which are incapable of growth in the absence of mitochondrial DNA [35]. "mya" indicates "million years ago". Figure taken from http://www.genome.wustl.edu/projects/yeast/index.php?map=1.

Introduction to Schistosomiasis

Schistosomiasis, or bilharzia, is one of the major diseases affecting developing regions such as Africa, Asia, Middle East and South America. Theodor Bilharz, a German pathologist, after whom the disease was named, first scientifically described the disease in 1851. Currently, schistosomiasis is endemic in 74 tropical developing countries, and potentially affects some 600 million people. Currently, 200 million people are infected, 120 million show symptoms, and 20 million have severe illness [36].

Schistosomiasis continues to spread to new geographic areas despite substantial progress in control and decreased morbidity and mortality. This has been attributed to environmental changes due to creation of new water supplies (for example, building of dams) and to the movement of infected populations. Extreme poverty, and lack of knowledge regarding the health risks as well as a lack of public health funding are further predisposing factors for the spread of infection.

The life cycle of the schistosome is depicted in **Figure 2** below and is roughly similar among all the schistosomes that infect humans. The parasite goes through a typical trematode life cycle involving various, quite distinct life forms that target different hosts. When coming into contact with water containing the cercaria life form, humans are infected. Cercaria can penetrate healthy skin within minutes to reach the blood stream. They then change form to become schistosomula, and then continue their life cycle within the infected individual's blood vessels. Within 30-45 days, transformation into a worm, either male or female occurs. Female worms produce a daily average of 200 to 2000 eggs, over a period of up to five years, depending on the

subspecies. These eggs are usually shed with human faeces or urine, and owing to poor hygiene, will come into open waters. There, the eggs hatch and release miracidia, which infect water snails through their feet [37]. Inside the snail, close to the penetration site, the miracidia transforms into a primary sporocyst that will then divide into secondary sporocysts. These secondary sporocysts migrate to the snail's hepatopancreas where they divide again producing thousands of cercariae. (The hepatopancreas is the snail's digestive organ that is the equivalent of the liver and pancreas in mammals). These cercaria are the larvae which, when released into the water, can re-infect humans.

This parasitic disease causes chronic malaise, often manifesting itself by blood in the urine, but frequently also by serious complications involving the liver (hepatitis, liver failure) and spleen (rupture). In the case of intestinal schistosomiasis, the parasites live in blood vessels lining the intestine. In the case of urinary schistosomiasis, they live in blood vessels lining the urinary bladder. About half of the eggs are shed, the others remain in the body, where they can lead to further organ damage. Of interest to note is that it is the damage caused by the eggs, not the worm, which makes carriers feel ill.

Another interesting point about the schistosomes, relevant to my findings in **Chapter 5**, is that the outer covering (called the tegument), of both females and males consists of a *double* lipid bilayer (**Figure 3**). This is in contrast to the single lipid bilayer common to most other living cells. The double outer membrane is believed to be an adaptation to living in a blood environment, driven by the exposure to the immune defence of the host [38]. In contrast, flukes (or parasites) inhabiting the gut or other

body cavities all only possess a single lipid bilayer.

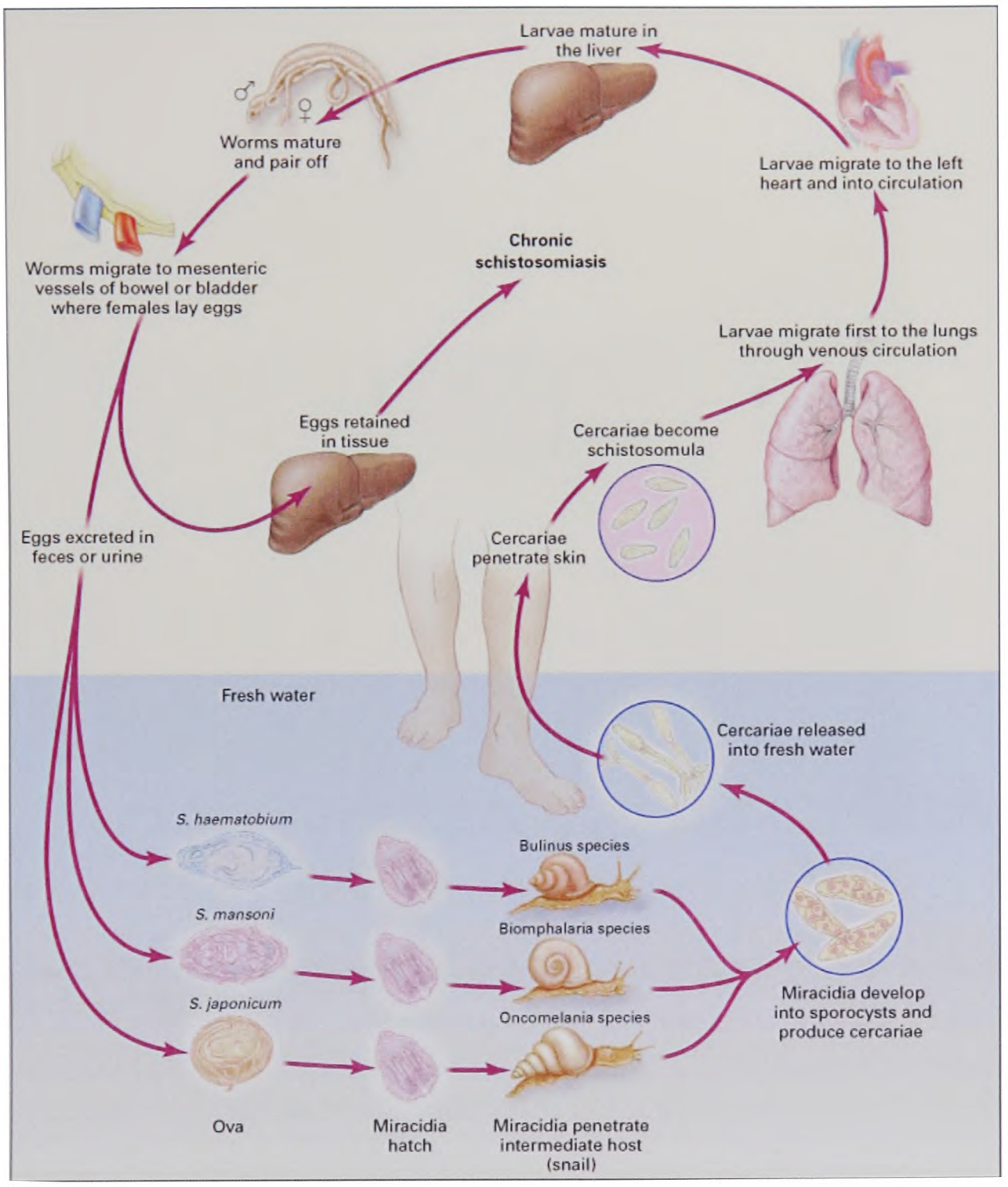


Figure 2. Life cycle of *Schistosoma*. There are 5 species of schistosoma known to affect humans. All schistosoma infections follow the same mode of infection – through direct contact with fresh water harbouring the free-swimming larval form of the parasite known as cercariae. After penetration of humans or other mammalian hosts, cercariae shed their tail to become schisotosomula. The schistosomula migrate through several tissues and life stages before taking up residence in the veins. After several days, the worms migrate to the veins of the liver where they mature and pair off. Adult worms in human reside in the mesenteric veins at various locations, which appear to be specific for the various species. For instance, *S. japonicum* is more frequently found in the inferior mesenteric and superior haemorrhoidal veins draining the small intestine, *S. mansoni* tends to reside in the superior mesenteric veins of the large intestine and *S. haematobium* is frequently found in the veins draining the ureters or vesical plexus of the bladder. These locations are not exclusive to each species and have been found to be interchangeable. The eggs produced throughout the life of the worm pass from the

lumen of blood vessels into adjacent tissues (lumen of intestines for *S. japonicum* and *S. mansoni*, bladder and ureters for *S. haematobium*). The eggs are shed in the faeces or urine of the infected individual. They hatch; releasing miracidia that in turn, infect freshwater snails. After two generations within the snail, cercariae are released. Picture taken from Ross *et al.* [39].

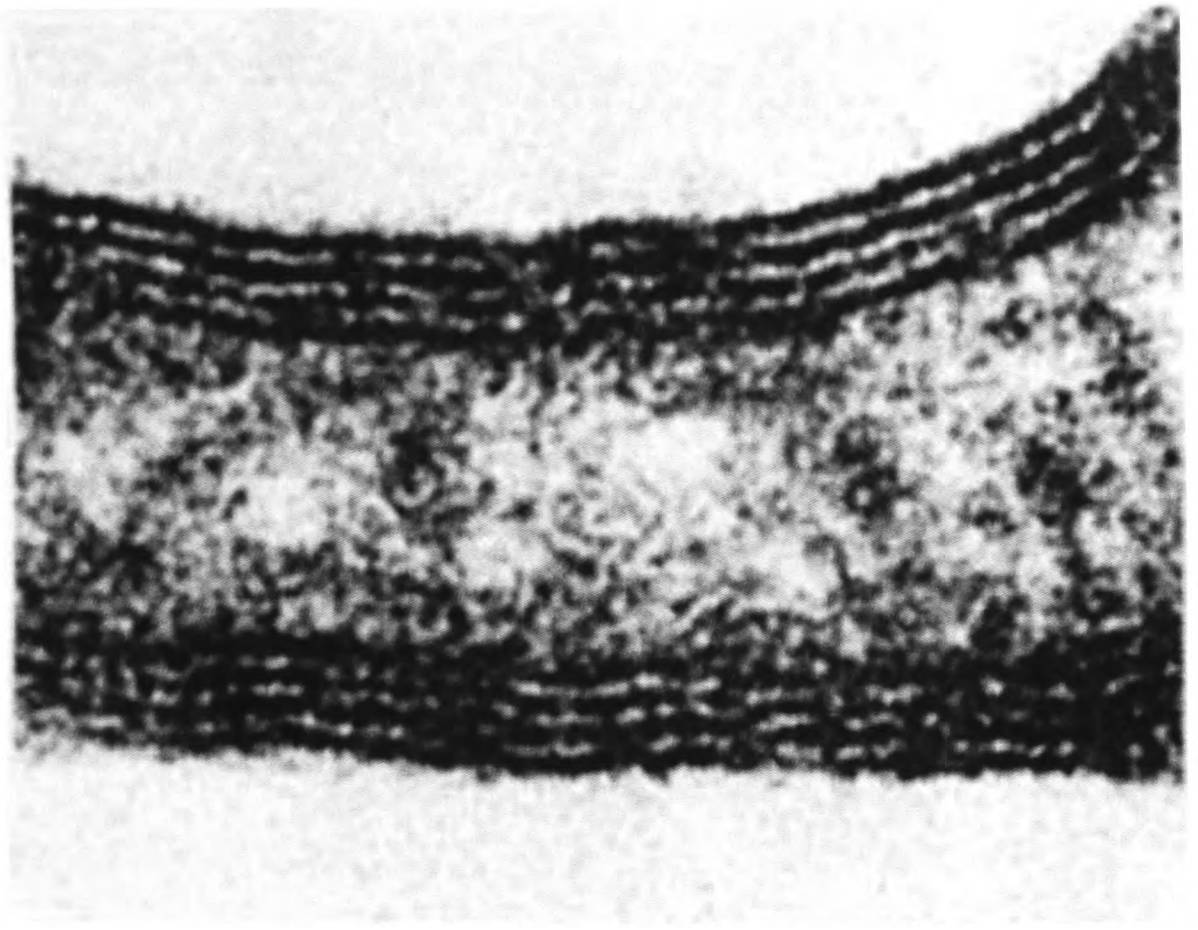


Figure 3. Electron micrograph of the double outer membrane of *S. japonicum*. Picture taken from [38].

Of the 19 known species of *Schistosoma, Schistosoma mansoni* and *S. haematobium* are primarily human parasites. Some species are anthropozoonoses, meaning that the disease can be transferred from animals to humans and thus species like *S. japonicum* affect wild animals, domestic livestock and man. Some species, like *S. bovis and S. curassoni* are generally veterinary pathogens [40].

The main forms of human schistosomiasis are caused by five species of schistosomes–*S. mansoni, S. japonicum, S. mekongi, S. intercalatum* and *S. haematobium*. Infections caused by the species *S. haematobium*, *S. mansoni* and *S. intercalatum* are found in sub-Saharan Africa, whereas *S. mansoni* infection is in the Caribbean, Venezuela and parts of Brazil. *S. japonicum* is endemic in China, Indonesia and the Philippines and *S. mekongi* infects individuals along the Mekong River, in Cambodia and Laos.

The current drugs used to treat schistosomiasis may provide a clue of where to direct our analyses. There are currently three drugs approved for the treatment of schistosomiasis. Of these, praziquantel, is the only effective drug against all types of schistosomiasis. It is most commonly used to treat human schistosomiasis due to its high efficacy and minimal side effects (WHO report, 2007). Oxaminiquine is the only alternative drug to treat *S. mansoni* but it has limited availability. Metrifonate, which was the alternative to praziquantel for *S. haematobium*, is no longer available commercially.

Praziquantel works only on adult worms. Although its precise action is still not known, it seems to cause vacuoles in the tegument and tetanic contractions (continuous muscle contractions caused by steady stream of nerve impulses) by causing a Ca^{2+} influx into the worm. This causes the worms to detach from the wall of the vein and die. The Ca^{2+} influx also induces tegumental vacuolization and surface blebbing causing an increased exposure of schistosomes' antigens on the body surface and triggers a local host immune response. The efficacy of praziquantel, in animal models, is shown to be dependent on the presence of host antibodies [41]. (A detailed

review on praziquantel and its mechanism of anti-schistosomal activity has been published by Andrews [42]). An antimalarial drug, arthemether, is lethal to schistosomula, the migrating larvae, if it is administered within 21 days of it being in the host. Used together, both arthemether and praziquantel work synergistically and the combination is especially useful in endemic areas.

Praziquantel has been in use for more than 20 years, and resistance to the drug has started to emerge in these parasites [39]. Widespread resistance is expected to occur within the next 10 to 20 years [43] and looking to new anti-schistosomal drug developments may prevent it becoming a clinical and public health issue.

Genetics of Schistosomes

Both *S. mansoni* and *S. japonicum* have 7 pairs of autosomes and one pair of sex chromosomes (female = ZW, male = ZZ). Chromosomes range between 18 to 73Mb and can be distinguished by shape, size and C banding. The number of genes in *S. japonicum* is estimated to be 15,000 [44] and *S. mansoni* is predicted to contain 14,000 genes [45]. The total expected size of each genome is around 270Mb. The guanine-cytosine (GC) content of the *S. mansoni* genome is about 34% and it has a highly repetitive genome (45%). The sequence identity between both genomes is about 84% [46].

Introduction to amniote brain evolution

"Greater understanding of brain evolution depends on studying and interrelating evolutionary changes at a variety of levels, from microscopic to macroscopic anatomy and from neural systems to behavioural ecology" – Robert A. Barton [47].

The brain has been a relatively conservative organ in evolution [48]. All across the vertebrates, the most striking change has been to its size (**Figure 4A**) and its disproportionate increase relative to body size. The brain is essentially made up of three main parts – forebrain, midbrain and hindbrain. The forebrain contains the cerebrum (or cortex). In mammals, the neocortex occupies the bulk of the cortex that is related to the forebrain structures found in other vertebrate classes. The neocortex can be further divided into four sections – frontal lobe, temporal lobe, occipital lobe and parietal lobe. In the "higher" and more advanced mammals, such as humans, the neocortex is a highly developed six-layered structure with deep grooves and wrinkles to increase its surface area to allow evolution of enhanced cognitive skills such as working memory, speech and language. More specifically, it is the frontal lobe and most parts of the left temporal lobe which have long been associated with speech and language [49]. In rodents and smaller mammals, the neocortex is smooth (**Figure 4B**).

Implications of a large brain

The neocortex is a feature particular to mammals and is suggested to have arisen from the need to adapt to changing temperatures in the environment [49]. Having large brains means an increase in cortex size that in turns allows an increase in cortical neurons [49, 50]. This has been linked to increased cognitive abilities [51], which helps the animal to adapt to changing environments. Yet having a large brain is a

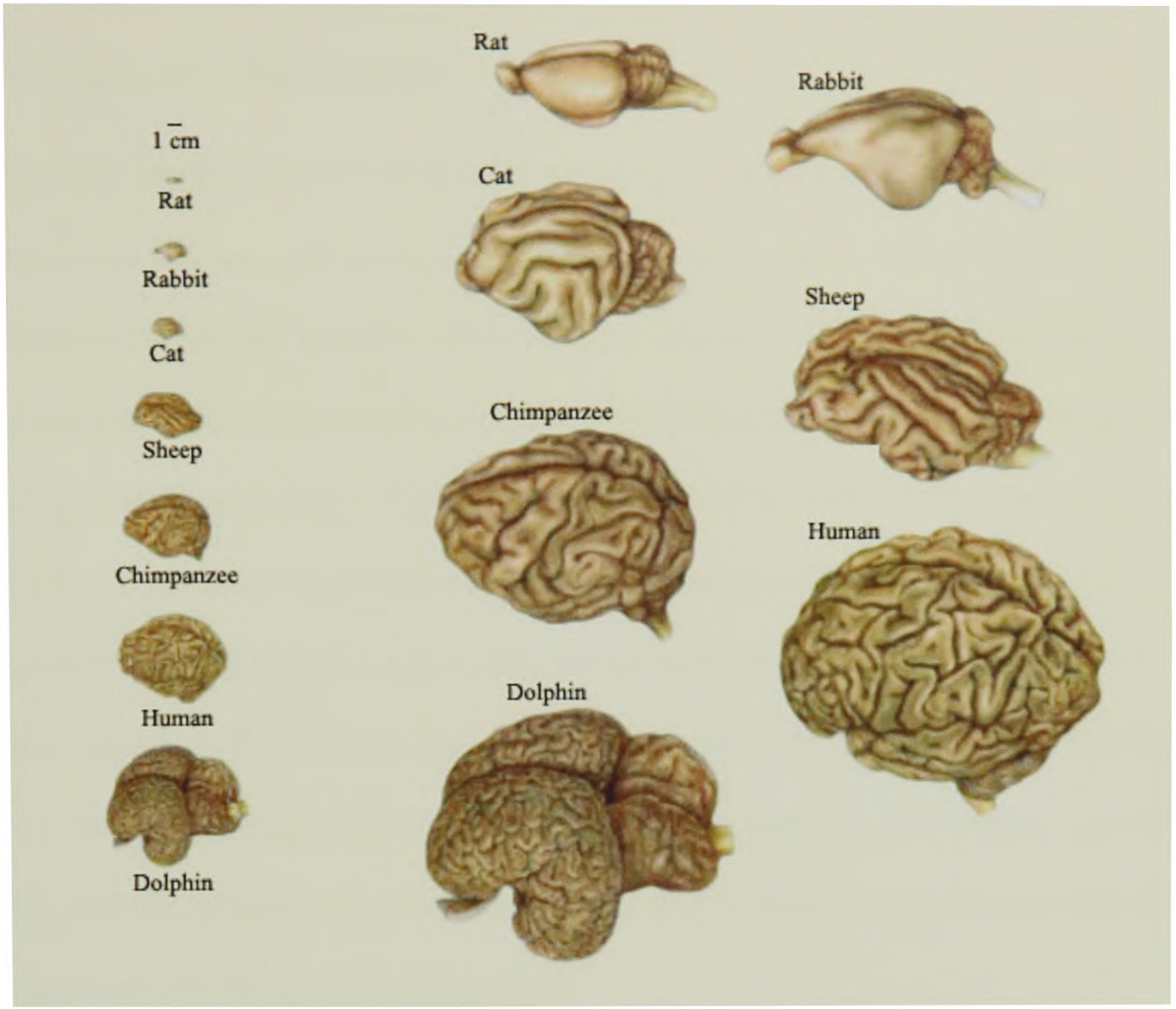
costly affair. Large brains take a long time to mature, which means that large-brained animals are dependent on their parents for a long time [49, 52]. Having a large brain also means that there is competition with other organs for energy. For instance, in the "expensive tissue hypothesis" proposed by Aiello and Wheeler [53], the human gut has been shown to be small relative to body size and was suggested to have decreased in size to offset the energy cost. This brain-gut correlation which is characteristic of primates is not seen in bats [54]. However, this "expensive tissue hypothesis" still has its applications in bats. A recent study showed that male bats with large brains had small testes [55]. Again, a lot of metabolic energy is required to produce and maintain both brain tissue and sperm cells. Preference for developing one organ more than the other seems to have evolved in different species of bats, presumably determined by which will optimise reproductive success. These resulting trade-offs based on the "expensive tissue hypothesis" can occur even within the same organ. For instance, the increase in neocortex size has resulted in the decreased sizes of other structures in the mammalian brains, for example, the hippocampus, septum, schizocortex, piriform cortex and olfactory bulbs [56].

Studies of small brains

The study of small brains is in its own right as important as understanding the evolution of large brains. Examining small brains provides evidence of the basic necessary components for sensory inputs and the appropriate motor outputs [57]. The smallest brain (absolute size) of all the mammals used in this study of brain evolution is that belonging to the lesser hedgehog tenrec (*Echinops telfari*). Among all extant mammals, tenrecs have the least neocortex. Their brains are said to be similar to those of long extinct mammals which roamed the earth 75-80 million years ago [57]. This

has placed the tenrec as an interesting model for determining forebrain evolution in mammals [57]. The only non-mammalian organism used in this study is the chicken. Studies have shown that evolution of brain size in this small-brained bird is influenced by maternal effects [52] and sperm competition [58]. Briefly, maternal effects such as the deposition of androgens (testosterone) and antioxidants (carotenoids, vitamins A and E) into the egg yolk influence the maturation and growth of the brain of their offspring. Furthermore, brain size evolution is independent in both male and female birds leading to brain size dimorphism which has been linked to sperm competition [58], a trait not seen in primates [59]. With intense sperm competition, the females birds have relatively larger brains (than male birds) which may aid in comparing and choosing copulation partners, thus emphasizing the role of sexual selection in the evolution of brain size in birds.

A)



B)

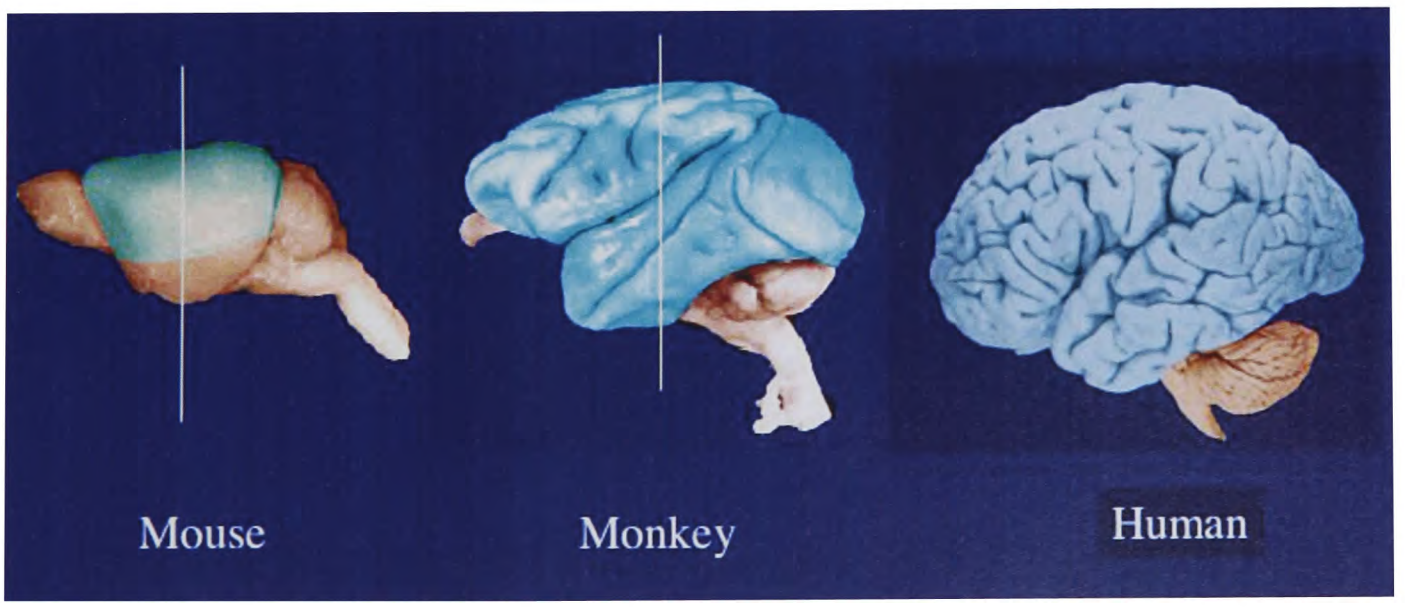


Figure 4. A) Differing brain sizes (absolute sizes) across the vertebrates. Picture taken from (http://thebrain.mcgill.ca/flash/i/i_05/i_05_cr/i_05_cr_her/i_05_cr_her.html). B) Smooth neocortex in mouse compared with the wrinkled primate neocortex of monkey and human. Neocortex is shaded in blue. The brain sizes of mouse and monkey in this picture are not shown to scale relative to the human brain. Picture taken from (http://www.nibb.ac.jp/brish/Gallery/cortexE.html)

Introduction to sequencing methods

Trends in genomic analysis

Sequence annotations and analyses have evolved from intensive small-scale procedures to large-scale projects spanning multiple countries and groups. The fully sequenced eukaryotic genome of *Saccharomyces cerevisiae* in 1996 [2] and the completion of the human draft genome in 2000 [60, 61] heralded the beginning of an era of genome sequencing. This has resulted in a barrage of different genomes being sequenced in the last few years with many others approaching completion. This included mammalian, non-mammalian, fungal, plant and pathogenic genomes, to name but a few. For example, the genomes of mouse [62], rat [63], puffer fish (fugu) [64], fruit fly (*Drosophila*) [65, 66], common pink bread mould (*Neurospora crassa)* [67], rice [68, 69] and blood flukes (or flatworms): *Schistosoma japonicum* [44] and *S. mansoni* [45, 46] have been sequenced.

Following the publication of these genome sequences, comparative studies are often performed between individual genomes, for example, between human and mouse, mouse and fly, human and fish. We are now able to analyze sequences at amino acid and nucleotide levels across broad evolutionary scales, deriving information about genes and genomes based on conservation and evolutionary selection. Furthermore, our ability to functionally annotate these sequences can lead to important medical discoveries and may ultimately have relevance to the pharmaceutical industry. In addition, comparing different genomes also point to species-specific genes which are of interest because they contribute to the individuality of the organism.

There is now a paradigm shift away from undertaking comparisons between distantly related species towards comparisons between closely related species. For instance, 12 species of *Drosophila* were sequenced for comparative analyses and published in 2007 [70], and multiple mammalian genomes have followed suit.

In this thesis, analyses on yeast were a pilot project aimed at multiple species comparisons among species of the genus *Saccharomyces*. Nine species of *Saccharomyces* have been sequenced; of these, five have been sequenced at a high (7-fold) coverage [20] and the rest at low coverage (0.2 -3 fold) [33, 34, 71]. These provide informative sequences with which to perform evolutionary rate analyses.

The techniques used to generate the yeast and schistosome sequences were whole-genome shot gun sequencing and EST sequencing respectively. Different sequencing methods produce different statistical coverage of the genome assembly that provides a rough estimation of the quality of sequencing data produced.

Sequencing a genome- Whole-genome shotgun sequencing

Haemophilus influenzae was the first microbial organism to be sequenced using the fast whole-genome shotgun sequencing technique [72]. This revolutionary method eliminated the need for initial mapping efforts in what was commonly known as the map-based or "BAC by BAC" method. This technique required a crude physical map of the whole genome before DNA sequencing could be carried out [73]. This physical map would be constructed by cutting up chromosomes into large pieces, ranging from hundreds of kilo basepairs to whole chromosomes and ordering them. Several copies

of the genome would then be cut up into fragments before being inserted into bacterial artificial chromosomes (BACs) for further processing and sequencing. (A BAC is a man-made piece of DNA that can replicate inside a bacterial cell, for example, *Escherichia coli*.) This accurate, but slow, method produced genetic maps with few gaps.

Whole-genome shotgun sequencing on the other hand, obviated the rate-limiting step of sequence-based maps. This technique was developed by J. Craig Venter and its speed and accuracy was demonstrated when it challenged the Human Genome Project consortium's effort in sequencing the human genome [60] [61]. The strategy involved randomly breaking DNA up into several pieces of various sizes and cloning them into vectors. These vectors are plasmids which are pieces of DNA that can replicate in bacteria. Multiple cloning systems are used to reduce the effect of sequences which are not clonable (example, repeat elements) or not present in any of the clone libraries (a library is a collection of clones). The clones (clones comprise of a vector inserted with a sequence fragment of DNA) are then sequenced from both ends that are essential for assembling the entire chromosome. Computer algorithms then assemble these sequenced fragments of DNA into the correct order and assembly.

How many copies of the genomic DNA need to be sequenced before a full coverage is obtained? In general, the more copies of DNA sequence generated, the higher the coverage of the genome. For instance, for the human genome project, an estimated 9X coverage of the genome (*i.e.* nine times the estimated 3.5 billion basepairs of human DNA) was thought to enable the clones to include most of the human DNA [74]. An

8-fold (or 8X) coverage of the human genome was published back in 2001 and this covered approximately 95% of the genome.

In this thesis, the sequence coverage of the yeast project ranged from 0.2X (very low coverage) to 7X (high coverage). Sequences for both the *S. japonicum* and *S. mansoni* genome were generated by expressed sequence tag (ESTs) sequencing (see below). Concurrently, 7-fold coverage of the *S. mansoni* genome is being sequenced by a joint project between the Wellcome Trust Sanger Institute (UK) and The Institute for Genomic Research (now the J Craig Venter Institute) (USA) (http://www.sanger.ac.uk/Projects/S_mansoni/).

Sequencing transcripts of expressed genes – EST sequencing

The initial emphasis of both the *S. japonicum* [44] and *S. mansoni* projects [75] was on gene discovery and comparative analyses between the 2 genomes [46]. Likewise, analyses of the *S. mansoni* transcriptome (set of all messenger RNAs (mRNAs) expressed in a cell) were also undertaken by a Brazilian group [45] with the aim of gene discovery. Both the Chinese [44] and Brazilian [75] groups used the technique of partially sequencing complementary DNAs (cDNAs) to produce ESTs. This has led to the rapid identification of expressed genes.

ESTs are short, partial, sequences (200-500 nucleotides in length) of cDNA clones [76, 77]. These ESTs are obtained by sequencing either one or both ends of genes found transcribed in a cell. Each short fragment acts as a "tag" for the full-length gene. As these sequences have been transcribed, the presence of an EST may indicate that a sequence is a real and functioning gene. In order to achieve a very high

throughput, these sequences are often only subjected to a single pass of sequencing, which can lead to a high error rate [78]. The biological value of these error-prone sequences is greatly enhanced once they are assembled or re-constructed into transcripts.

From cDNA to ESTs

The process of generating ESTs is summarized in **Figure 5** below. ESTs are sequenced from cDNA which in turn is reverse transcribed from the mRNA. The reason that cDNAs are used instead of mRNAs is that mRNAs are very unstable outside of a cell. Hence, the mRNA acts as a template on which the enzyme reverse transcriptase is used to produce the complementary DNA strand. Like mRNA, cDNA contains only the expressed DNA sequence (exonic sequence) and none of the intronic sequence. This is key to gene identification.

Normalisation of cDNA libraries

The more copies of mRNA in a cell, the more highly expressed the corresponding gene is. This sometimes leads to large variations in the numbers of mRNAs produced between highly expressed and rarely expressed genes [79]. By counting the copies of mRNAs (via the cDNAs sequenced from them), an estimation of the expression level can be obtained. However, if the abundance of the clones is required in almost equal numbers to each other, normalisation of the cDNA libraries would need be carried out. The *S. japonicum* ESTs used in the schistosome project were from unnormalised cDNA libraries. Hence, they are able to be used for EST abundance studies.

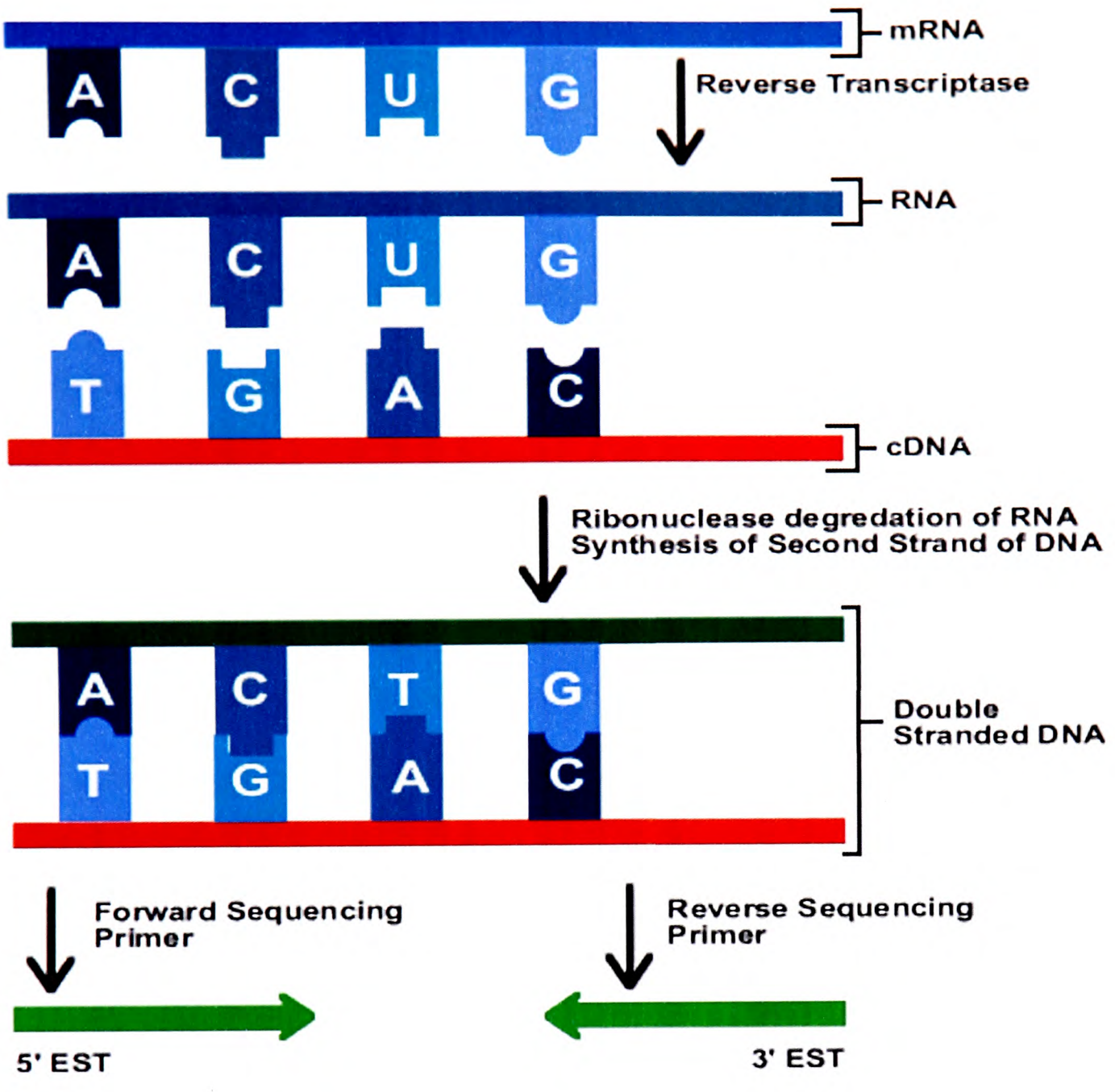


Figure 5. From cDNA to ESTs. cDNAs are generated by the enzyme reverse transcriptase, from mRNAs expressed in a cell. As cDNAs are stable outside a cell, they can then be used as a template to generate ESTs. 5' or (and) 3' ESTs are produced depending on whether one or both ends of the cDNA are sequenced. Picture taken from http://www.ncbi.nlm.nih.gov/About/primer/est.html.

Introduction to sequence similarity and homology

When genes from different organisms are available, interesting questions such as "which sequences are homologous to each other?", "how similar are the genomes?", "which are species-specific genes?" maybe asked. Methods (example, BLAST [80])

used to detect if similar sequences were homologous are discussed in **Chapter 2**. Here I briefly introduce the terminology used later on.

Homologues, Orthologues and Paralogues

'Homology' is a widely used term in biology. Following Darwin's evolutionary theory, homology requires that characters be similar by descent, *i.e.* descended from a common ancestor, usually with divergence [81].

The term "homology" is often wrongly used in place of similarity. Homology is indivisible. A character is either homologous with another or it is not. It cannot be 70% (for example) homologous, instead, maybe 70% identical. Similarity alone cannot provide information on ancestral relationships. Homology is inferred from similarity [82]. The only situation in which the term "percent homology" may be used between genes is if one or the other has been the product of gene fusion or fission events [83, 84]. For example, through mutation, a region in a particular gene may have been formed by the accidental joining of 2 DNA sequences through translocation (interchange of DNA between non-homologous chromosomes), interstitial deletion (segment of DNA deleted from a chromosome, thus bringing together previously distant genes) or inversion (reversal in orientation of a segment of DNA). This domain may then be partly homologous with that of another gene sharing the same ancestral derivation. However, even in such cases, "percentage homology" is open to such misunderstanding that its use should be generally avoided [84].

When homology is applied to genes, a further separation into two distinct subclasses occurs: orthologous and paralogous gene sequences. Again, homologous genes are

those that descended from a common ancestral gene, regardless of functional differences or degree in sequence identity. Perceived sequence similarity does not imply a common evolutionary origin [83]. Orthologous sequences (*i.e.* orthologues) are genes that arose from a speciation event. Paralogous sequences (*i.e.* paralogues) are genes arising from a duplication event within the genome. These relationships are illustrated in **Figure 6**. In **Figure** 6, members of groups A and B are all homologues; examples of orthologues are A_1 and B, or A_2 and B; members within group A are paralogues.

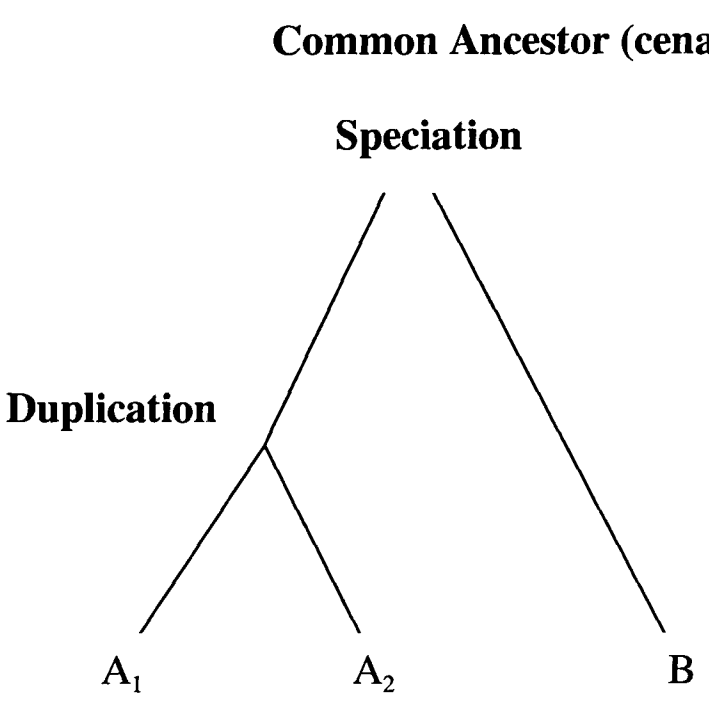


Figure 6. Homologues, orthologues and paralogues. Members of group A and B are homologues as they share a common ancestor. A speciation event leads to orthologues being formed. Orthologue pairs are between either of the As and B, for example, A_1 and B, or A_2 and B. Intragenome duplication gives rise to paralogues A_1 and A_2.

Importance of using orthologues

It is important to use only the sequences of orthologues in the reconstruction of organisms' phylogeny [82]. Only when orthologues are used would the phylogeny of the organisms be the same as that of the sequences. If paralogues were mixed in with

orthologues, the sequences' (gene) tree would differ from the species' tree [84]. The study of orthologues is of particular importance because often they retain the function of the ancestral gene. They may also have similar physiological or developmental roles, thereby sharing conserved functional and regulatory domains [77]. Paralogues usually diverge functionally following gene duplication or else only one copy is retained [82]. However, functional similarity does not imply orthology. Only through phylogenetic reconstruction and comparative functional analyses can orthology and functional similarity be determined respectively.

Introduction to mutation and natural selection – definitions and estimations

The mutations considered here are *substitutions*, *insertions* and *deletions*. Substitutions occur when nucleotides in sequences are changed whereas insertions and deletions (collectively called indels but called gaps in sequence alignments) add or remove bases. Selection acts on these mutations and preferentially fixes these changes in the population or removes them.

Point mutations – Transition, Transversion mutations

Point mutations causing a substitution to the DNA are either transitions or transversions. They are often caused by chemicals or a malfunction in DNA replication [85]. DNA is composed of purines (adenine and uuanine) and pyrimidines (cytosine and thymine). Purines are two ring structured bases and pyrimidines one-

ring structures. Transition is the replacement of purine for a purine (A ↔ G) or of a pyrimidine for a pyrimidine (C ↔ T). Transversion is the interchange between a pyrimidine and a purine (C/T ↔ A/G), which would involve the exchange of a one ring for a two-ring structure (**Figure 7**). Although at a mutation site there are twice as many possible transversions (four possible combinations) as transitions (two possible combinations), transitions occur at a higher frequency than transversions.

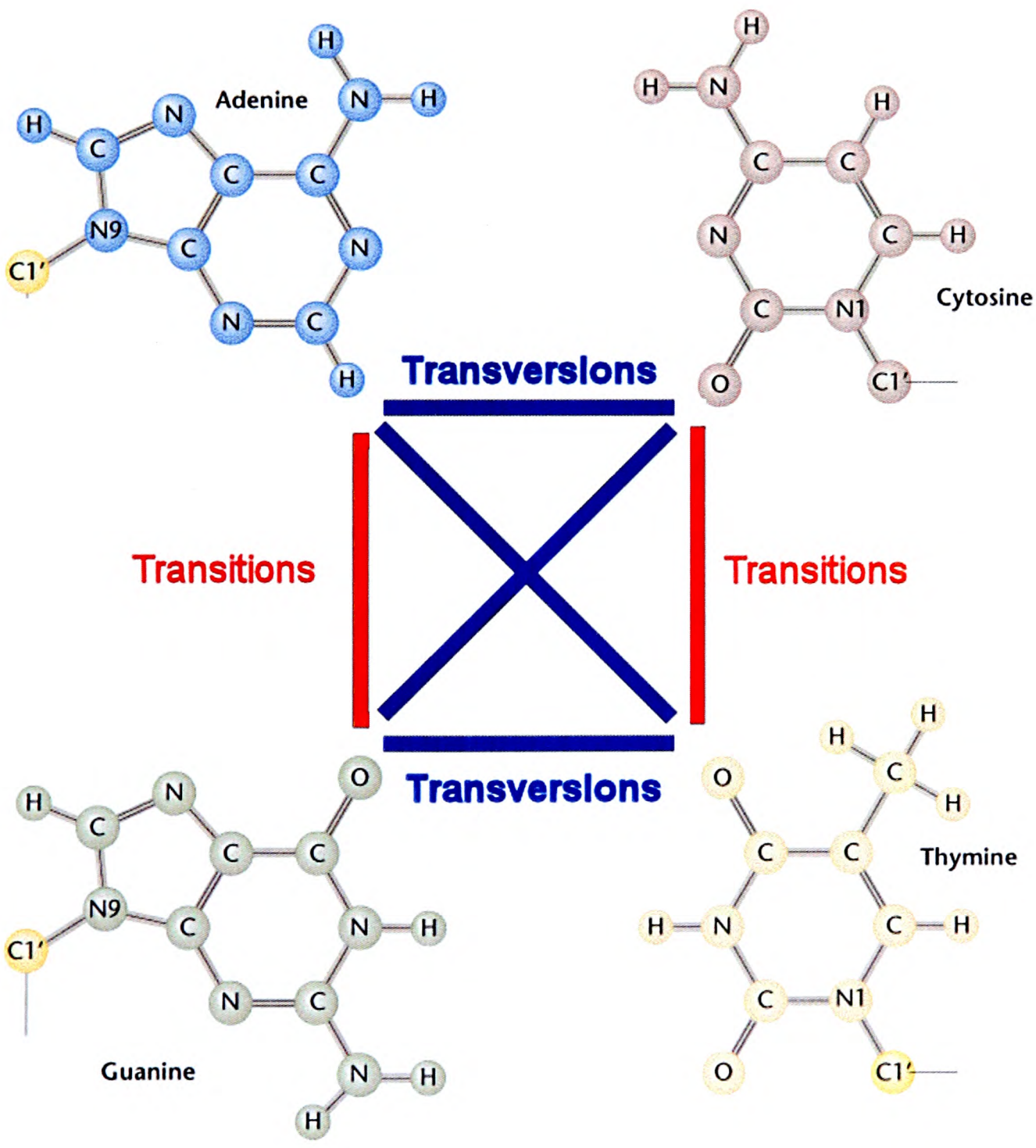


Figure 7. Transition and Transversion nucleotide substitution. Picture taken from http://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html.

Transition-transversion bias

This transition-transversion bias has been attributed to the hypermutability of cytosine-guanine (CpG) dinucleotides and its mutational bias towards adenine and thymine. Methylation of CpG sites have been well characterized in vertebrates [86] and involves the addition of a methyl group to the cytosine pyrimidine ring. The deamination of these methylated cytosines (to a thymine) leads to elevated rates of transition at these methylated sites [87, 88].

The observed bias can also be attributed to the protein coding regions due to constraints on non-synonymous (amino-acid changing) substitutions. Both transitions and transversions can occur at non-synonymous sites. However, selection may act against transversions by favouring DNA repair systems which prevent them [87]. This may in turn influence the observed substitution patterns across the genome (including non-coding regions). The difference in physicochemical (charge, polarity and volume) change caused by a transversion introduces more radical changes compared to changes caused by a transition [89]. Alteration in biochemical properties of the ensuing amino acid and its function comes at a greater cost for transversions than transitions [89]. Transversions would thus be more likely to be subjected to greater purifying selection. The intensity of purifying selection may vary between genes due to factors like codon usage bias and constraints on protein structures [87]. Later in **Chapter 4**, I discuss the effects of codon usage bias and constraints on protein sequences within yeast.

Usage in phylogeny reconstruction

The ratio of transitions to transversions is also known as *kappa* (κ). This value can be estimated using different methods (distance based, parsimony and maximum-likelihood methods) (reviewed in [90]). Ignoring the transition-transversion bias can result in the overestimation of the number of non-synonymous sites [89, 91]. Thus I account for κ in the methods for calculating evolutionary distances as discussed in **Chapter 2**.

Synonymous and non-synonymous substitutions

When a mutation occurs in the protein-coding region of a gene, both the DNA and the subsequent mRNA are mutated. However, this may or may not lead to a different protein, owing to the genetic code being *degenerate*. For instance, if one letter in a codon of a coding sequence was mutated from TTA to TTG, the amino acid "leucine" would still be translated from both codons. Such mutations are called silent, same-sense or synonymous, as they do not affect the protein sequence in any way.

If however, the mutation caused the change of TTA to TTT, the amino acid "phenylalanine" would be translated instead of "leucine". Such substitutions are commonly called missense or non-synonymous mutations. Although non-synonymous mutations alter the ensuing protein sequence, it may or may not affect the properties of the protein. *Non-synonymous* mutations can be further sub-categorized into either *conservative* or *non-conservative*. In the example above, the two (both the original and the altered) amino acids "leucine" and "phenylalanine", are chemically similar, thus this substitution is termed *conservative*. Alternatively,

chemically different amino acids would result in a *non-conservative* substitution. If a non-synonymous mutation results in a stop codon, for instance, TTA to TGA, such *non-sense* mutation will induce premature translation termination or result in a partial protein. This truncated protein may be partially functional or not at all, or it may be toxic to the cell.

When nucleotides are removed and/or inserted, deletions and/or insertions occur respectively. Large-scale rearrangements such as duplications or inversions may result in gene fusion or gene deletions or destruction. The fate of such mutations depends on many factors, mainly selection and genetic drift. If mutations are beneficial for the organism and are selected for, these changes are preferentially propagated through the population and become fixed. Conversely, if by random chance genetic drift occurs, the frequency of a mutation not affected by natural selection is increased and over many generations if the trend continues, this mutation may eventually become fixed [92].

Measurements of substitution rates

We need to understand why some mutations occur more frequently than others if we are to interpret the differences that have accumulated between individuals and species. Estimation of the evolutionary time since species diverged and detecting the fingerprints of natural selection on the DNA sequences may shed light on the mutational biases involved. Although selection has eliminated most of the mutations which have arisen, some have persisted. Hence, by using sequence alignments, the

evolutionary rates of change caused by substitutions and insertions or deletions (indels) may be measured [20].

dN and dS

dN (K_A) is defined as the number of observed non-synonymous substitutions for each non-synonymous site. This is used to measure amino acid sequence change. dS (K_S) is the observed number of synonymous substitutions for each synonymous site. Synonymous sites are under weak selection hence dS can be used to estimate the neutral mutation rate.

dN/dS

The rate ratio ω (K_A/K_S or dN/dS) is a measure of natural selection at the protein level. By calculating the ratio of the rates of non-synonymous (amino-acid changing) to synonymous (silent) nucleotide substitutions on pairwise sequence comparisons, we can estimate how much protein-coding sequences have evolved for a given underlying mutation rate. $\omega = 1$ indicates neutral evolution for both sequences. This is commonly seen in pseudogenes which are largely free from evolutionary constraints [93]. $\omega < 1$ indicates purifying selection with most amino acid changes being deleterious and thus few being fixed in the population. $\omega > 1$ indicates positive selection where the number of amino acid changing substitutions is greater than substitutions that are selectively neutral. Hence, the lower the ω, the greater the conservation of the encoded amino acid sequence. The higher the ω, the faster the amino acid sequence is diversifying. In general, most genes, for example, housekeeping genes, are well conserved with ω ratios less than 0.1 [94].

An important point to note is that the above method of obtaining ω is averaged over the entire protein sequence; it often lacks the power to detect positive selection. This is because adaptive evolution usually only occurs at small numbers of sites because most amino acids are subject to purifying selection and changes to them would adversely affect both structure and function. Hence, site-specific information and not averaging rates over the entire protein would enable these small numbers of amino acids to be detected [95]. By measuring the mutation rate at each site, we are able to gain a better insight into selection and understanding of the mechanisms of DNA sequence evolution. I discuss these tools in **Chapter 2** and show their application in all the projects undertaken in this thesis (**Chapters 3** to **6**).

CHAPTER 2: Methods and Materials

Summary

In this chapter, I describe the algorithms and the basic concepts involved in the various tools used for analyses in this thesis. First, I discuss sequence similarity search tools used to infer homology among sequences. I then summarize the concepts behind these tools and their application in my analyses. Secondly, I detail techniques that were used in deriving evolutionary rates of sequences. Finally, I highlight the statistical methods and concepts which I used in **Chapters 3** to **6**, with a focus on their logic and characteristics.

Sequence Similarity Tools

Throughout this thesis, database homology searches have been used to search protein and DNA databases for sequence similarities. The goal of deriving a sequence alignment from such searches is to infer the true evolutionary relationship between the query sequence and its matching database sequences. A sequence alignment is the pairing up of two sequences in such a way so as to identify regions of similarity (due to their functional, structural or evolutionary relationship). It is noteworthy that the inverse is not always true: homologous sequences do not necessarily imply a common function, rather that they share a common ancestor [96].

In bioinformatics, the two most commonly used types of alignment tools are those of global and local alignment algorithms. A global alignment approach attempts to optimally align a pair of sequences along their entire lengths. In **Chapter 3**, I used the global alignment approach (the CLUSTALW program) when aligning orthologues from the closely related *Saccharomyces stricto sensu* species. This was because sequences were similar and roughly of equal size and I wanted to align all the residues in the sequence. By contrast, a local alignment approach identifies regions of similarity within long (often widely divergent) sequences. In **Chapters 3** and **6**, I used local alignment tools, Basic Local Alignment Search Tool (BLAST), Blast-2-Sequences (bl2seq locally aligns pairwise sequences) and PSI-BLAST when orthologues from evolutionarily more diverged species (for instance, outside of the *stricto sensu* group in **Chapter 3**) were aligned. However, it is noteworthy that when sequences are sufficiently similar, there are essentially no differences between the results of a global or local alignment approach [97].

BLAST (Basic Local Alignment Search Tool)

Local similarity approaches are usually preferred in database searches, for instance where cDNA is compared to incompletely sequenced genes or when distantly related proteins share only small regions of similarity corresponding to structural motifs or active sites [80]. The basic local alignment search tool (BLAST) [80] is a rapid heuristic algorithm that aligns two sequences efficiently. BLAST does not need to explore the entire search space (defined as the set of all possible alignments) between sequences before detecting statistically significant alignments. BLAST employs three processes in the following order: seeding, extension and evaluation.

Seeding

During seeding, BLAST assumes that significant alignments contain regions of exact matches (word hits) or near matches (word neighbourhood hits) and looks for the location of these regions in the search space. In other words, BLAST locates short sequence matches (without gaps) between the query sequence and the database sequences which seed the search. The length of this initial match and the number of word hits are both determined by the word size (W). A larger word size would allow BLAST to run faster but at a higher chance of missing an alignment. For the protein matches used in this thesis, I used the default word size of 3. Thus a word "RGD" would register in its neighbourhood words such as "KGD", "QGD", "RGE" and so forth.

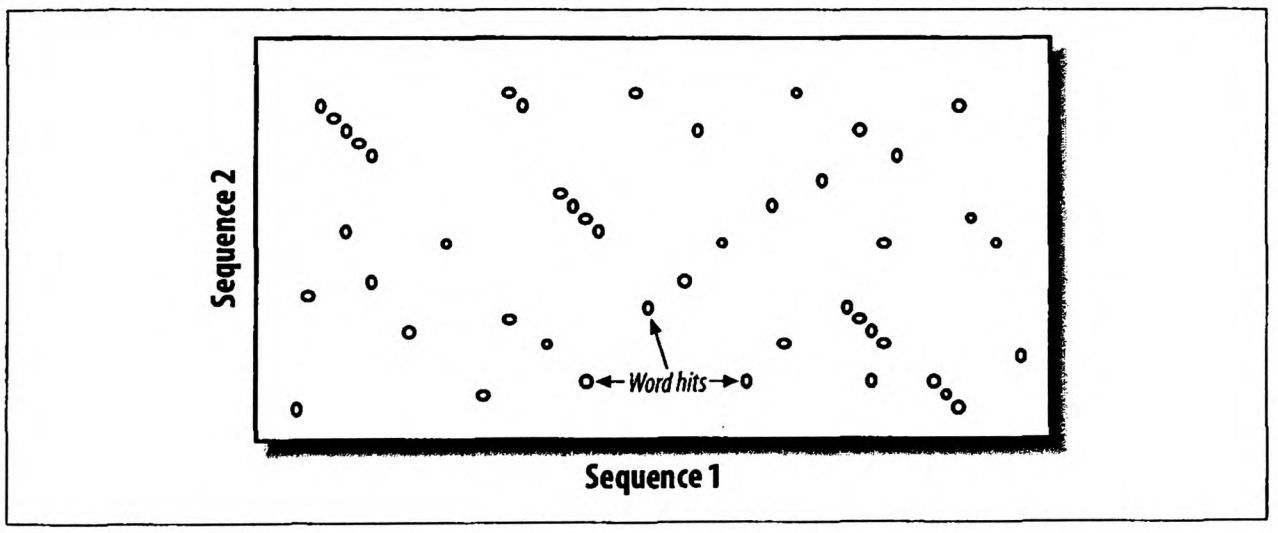


Figure 8. An example of word hits (represented by the dots), taken from [98]. BLAST identifies exact word matches and near-exact matches, which obviates the need to explore the entire search space between two sequences to find significant alignments. The axes represent the two sequences being aligned and the search space is the set of all possible alignments.

BLAST uses a matrix of similarity scores derived for all possible pairs of residues. Here, I used the BLOSUM (BLOcks SUbstitution Matrix [99] scoring matrix which is derived from "blocks" of aligned regions from homologous protein sequences. Specifically, I used the BLOSUM-62 amino acid substitution matrix (**Figure 9**), which was constructed from "blocks" (or regions) of alignments whose sequences had at least 62% identity to another member [98]. The log odds ratio of the probabilities that any two amino acids would replace each other over evolutionary time is tabulated into a scoring matrix (**Equation 1**). Hence, this scoring matrix or substitution matrix represents the relative rate of substitution between two amino acids.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1

Figure 9. BLOSUM 62 scoring matrix (from [98]).

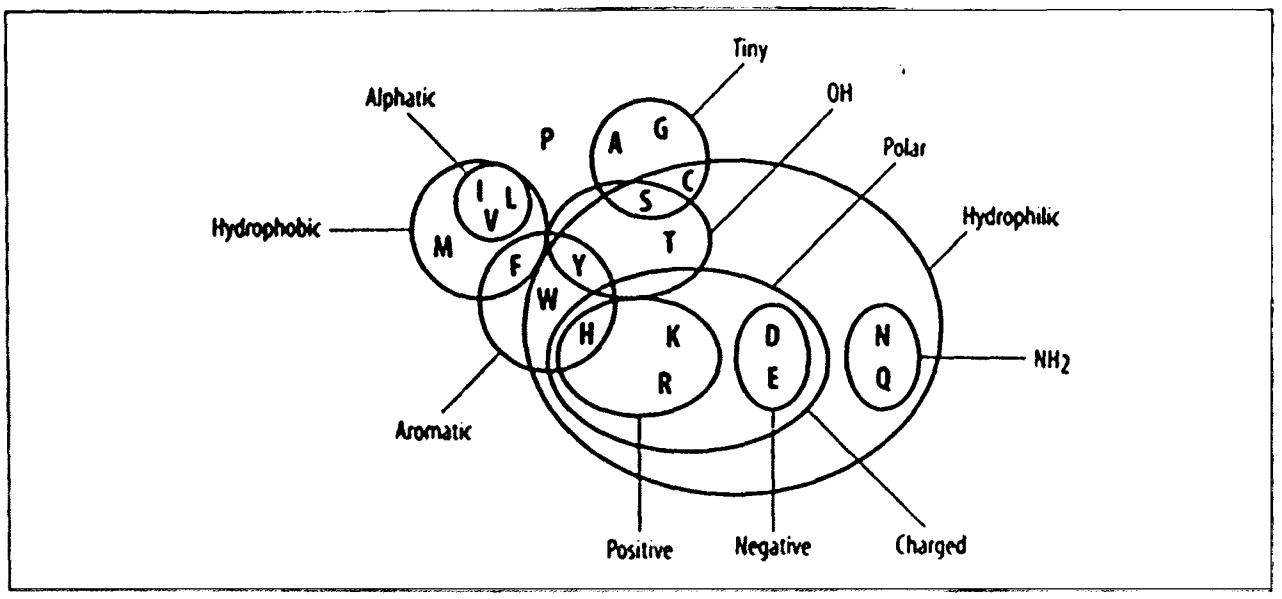


Figure 10. Amino acid chemical relationships. Phenylalanine (F) is most frequently paired to itself or with other aromatic ring structure amino acids - Tyrosine (Y) or Tryptophan (W), which have similar chemical properties. It may also occasionally pair to other hydrophobic amino acids (M, V, I, L). Pairings with hydrophilic amino acids like R, K, D and others are rarer. Taken from [98].

We expect substitutions between amino acids which alter the chemical properties of the resulting protein to be rare as these would more frequently disrupt structure and/or function. Thus, we expect more substitutions between amino acids sharing similar chemical properties (**Figure 10**). These are calculated as log odds ratios defined as the logarithm of the ratio of the observed number of substitutions between the two amino acids divided by the probabilities of observing these residues by chance. The general formula for any pair of amino acids is shown in **Equation 1**.

$$S_{ij} = \log(q_{ij} / p_i p_j)$$

Equation 1. General formula for any pair of amino acid in a scoring matrix. *S* is the log odds ratio of 2 probabilities: the probability that residues i and j are aligned by evolutionary descent and the probability that they align by chance. The frequencies of occurrence of a pair of amino acids residues, i and j, are p_i and p_j. The frequencies that residues i and j are observed to align in related sequences is given by q_{ij}. These are derived from a transition probability matrix and provide an estimate of how likely two amino acids are aligned in an alignment of homologous sequences.

In new versions of BLAST, a "*two-hit algorithm*" has been implemented [100] which increases the speed of the search. This algorithm exploits the observation that closely related word hits are usually located near to each other, and hence uses two hits instead of a one word hit as seeds to start the extension of a hit. Variations of BLAST except for BLASTN (because large identical word hits are rare) all employ this algorithm. For instance, BLASTP, used in **Chapter 3**, compares protein queries to a protein database by using this "*two-hit algorithm*".

Extension

After finding a short match between the sequences, a hit is extended in both directions to assess whether it lies within a high-scoring alignment. Identities and conservative replacements would score positively while unlikely replacements score negatively. The similarity score associated with the hit is the sum of the S_{ij} (**Equation 1**) values for each pair of aligned residues. This log score is additive – BLAST will continue to extend this alignment so long as the score remains positive and does not drop below a user defined threshold (T). As the current version of BLAST (version 2.2.14) accounts for gaps (by using negative values), the cost of gap initiation, extension cost and substitution scores all contribute to the alignment score. Once the extension step is terminated, the alignment is trimmed back to the maximal score and statistically evaluated to test for significance (see below). Segment pairs within the alignment whose scores cannot be improved further by extension or trimming are called "high-scoring pairs" or HSPs.

Evaluation

The BLAST statistics used throughout this thesis to evaluate the significance of an alignment match are the *E*-value and the bit-score. These values are explained below.

E-value

Every HSP is associated with a score. The statistical significance of these alignment scores can be estimated by the Expect (*E*) value. The *E*-value allows us to estimate the number of different HSPs which score better or equivalent than the highest HSP score expected by chance, given the size of the database searched and the scoring matrix used [80, 101]. This holds true for both gapped and ungapped alignments. Hence the lower the *E*-value, the closer it is to zero, the more likely the alignment contains homologous sequences.

Bit-score

To compare alignment scores from different searches in **Chapter 3**, I used the bit-score calculated by the BLAST program. The bit-score, S′, is derived from the raw alignment score (S_{ij} from **Equation 1**) and normalized according to the scoring matrix used. By normalising the raw alignment score, the bit-score may be used to compare alignment scores from different searches. Hence the higher the similarity, the higher the bit-score.

PSI-BLAST

PSI-BLAST (Position-Specific Iterative BLAST) [100, 102] is a program developed to increase the sensitivity of database searches. It is a protein sequence comparison tool that exploits multiple alignment profiles in order to achieve sensitivity at detecting distant relationships better than BLASTP. It uses a position-specific scoring matrix (PSSM) to conduct database searches, see **Figure 11**. In a first round (i = 1), PSI-BLAST takes the input/query sequence and compares it to a protein database using BLAST whilst allowing for gaps. A multiple alignment and then a profile (or PSSM) are generated using any significant alignments found, based on a user defined threshold. The query sequence then forms a template for the multiple alignment and PSSM. This PSSM has a length identical to that of the query sequence. PSI-BLAST then uses this PSSM to search the protein database for local alignments in the next round of iteration (i +1). This PSSM contains scores to each position of the query sequence based on alignments generated in round i. The scores of the positions in these motifs are added iteratively on each consecutive run until homologous protein matches from within the database are exhausted. These scores are calculated based on the number of observed substitutions between the query sequence and the corresponding position in the homologues detected, just like those of a substitution matrix.

This iterative search improves its sensitivity in searching for homologues in consecutive runs [100]. The use of these PSSMs, which encode conservation of residues, enable better detection of more distant evolutionary relationships between proteins than pairwise comparison methods [100].

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
2	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
3	P	-1	-2	-2	-2	-3	-2	-1	-2	-2	-3	-3	-1	-3	-4	8	-1	-1	-4	-3	-3
4	S	1	-1	0	-1	-1	0	0	-1	-1	-3	-3	0	-2	-3	-1	5	1	-3	-2	-2
5	C	-1	-4	-3	-4	9	-3	-4	-3	-3	-2	-2	-3	-2	-3	-3	-1	-1	-3	-3	-1
6	T	0	-1	0	-1	-1	-1	-1	-1	-2	-2	-3	-1	-2	-3	-1	4	3	-3	-2	-2
7	Y	-2	-3	-3	-4	-3	-2	-3	-4	1	-1	-1	-3	-1	5	-4	-2	-2	1	7	-2
8	Y	-1	-1	-1	-1	-2	0	-1	-2	6	-2	-1	-1	-1	1	-1	-1	-1	0	5	-2
9	V	-1	-2	-2	-2	-1	-2	-2	-2	-2	1	2	-2	0	-1	-2	-2	-1	-2	-1	4
10	S	-1	-1	-1	-1	-3	3	3	-2	-1	-2	1	0	-1	-2	-2	2	-1	-3	-2	-2

Figure 11. PSSM for the first 10 amino-acids of the coelacanth HoxA11 protein (taken from [98]).

In **Chapter 3** and **6**, I used **PSI-BLAST** to search various databases (example, UniProt [103] and/or NCBI's GenBank non-redundant database [104]) to functionally annotate my protein sets of unknown function.

Exonerate

Exonerate [105] is a generic tool for pairwise sequence comparison. This unique tool can be used to align sequences using either exhaustive dynamic programming, or a variety of heuristics. This provides the flexibility necessary to incorporate different alignment models; with some models approximating to the complexity of the models used in gene prediction programs such as GeneWise [106]. Exonerate uses the same seeding strategy as BLAST to obtain HSPs which are then linked to form alignments. In **Chapter 5**, I used Exonerate to align ESTs to translated cDNA using the "protein2dna" alignment model. An added advantage to Exonerate is the "roll-your-own" (ryo) option which allows the user to specify the desired output of the results. This eliminates the need for an additional parser of the Exonerate output. Due to its speed and accuracy, Exonerate is often used in genomic annotations [105].

CLUSTALW

There have been over 50 multiple sequence alignment programs described in the last 10 years [107]. Some of the more popular programs for multiple sequence alignment are MULTALIGN [108], DiAlign [109], CLUSTALW [110], T-Coffee [111] and MUSCLE [112]. A review of these and other programs has been provided by Notredame [113]). The performance of these different programs varies with the number of sequences, the degree of identity of sequences and the number of indels in the alignment.

Later, in **Chapter 3**, I describe how closely related (*stricto sensu*) yeast sequences were globally aligned. These were highly similar and roughly of equal size, hence it did not really matter which alignment program was used. The program chosen, CLUSTALW [110], is a global alignment tool used to align multiple sequences. A column of aligned residues within an alignment implies that they sit within equivalent 3-D structural locations and that they diverged from a common ancestral residue [114]. It uses the heuristic progressive alignment method which first constructs a distance matrix of all $N(N-1)/2$ pairs of sequences which gives the evolutionary distances between each pair. From this distance matrix, a guide tree is constructed using the neighbour-joining algorithm [115] (see below), which determines the order in which the progressive alignment is carried out. Sequences are progressively aligned, at each branch point, starting with the most closely related species. Specific rules within CLUSTALW contribute to the generation of accurate alignments. For instance, sequences are down-weighted according to how closely related the sequences are. This aims to compensate if a large subfamily dominates an alignment. In addition, the substitution matrix used to score the alignment also depends on how

closely related the sequences are. More closely related sequences may incur the use of a conservative matrix, for example the BLOSUM-80 matrix that requires sequences to have at least 80% identity with other members in the set. Distant sequences may be scored based on the more relaxed matrix, for instance, the BLOSUM-50 substitution matrix. Hydrophobic residues have higher gap-penalties than the surface accessible hydrophilic (flexible) residues. Also, gaps in the alignment are forced to occur in the same place by the increased gap-extend and gap-open penalties imposed. This occurs if there are gaps in the alignment near a fully aligned column of residues, for instance, in the loop regions between secondary-structure elements.

The major disadvantage in using CLUSTALW is that the algorithm is "greedy" and that errors made in the first alignments cannot be rectified later as the rest of the sequences are added. Recently, a new program, M-coffee [107], was made available. This program is an extension of T-coffee and is able to combine the output from different multiple sequence alignment programs into one single multiple sequence alignment. The authors claim that M-coffee is able to deliver a better alignment than any of the individual methods it incorporates. It might be worthwhile testing this algorithm in subsequent projects.

Gene prediction using GeneWise

GeneWise was used for gene prediction in **Chapter 3**. There are two forms of evidence for a gene – placement of cDNA and EST on the genome of the same species and the evidence of homologous genes in other species [106]. I used GeneWise to do the latter. (GenomeWise is another program developed by Birney *et al.* [106] which does the former).

Although the computational cost of GeneWise is high, this program has been robustly tested within Ensembl on many genomes [60, 62, 64]. (Ensembl is a joint project between the EMBL-EBI and the Wellcome Trust Sanger Institute in Hinxton, Cambridge, which aims at developing a system which maintains automatic annotation of large eukaryotic genomes (http://www.ensembl.org/)).

As GeneWise predicts the intron-exon structure of a gene using similar protein sequences, I supplied regions of homology (*E* values of less than $1x10^{-10}$) between a specific yeast genomic sequence and its protein sequences. Through dynamic programming, GeneWise uses the homologous protein sequence to guide the gene prediction process as each protein sequence is being directly compared to its genomic DNA whilst taking into account the statistical properties of a gene structure and the presence of sequencing errors.

Whilst there is always the classic trade-off between specificity and sensitivity at the exon level and in the GeneWise algorithm, the authors have chosen specificity [106]. This results in the loss of coverage by GeneWise for less similar genes, which may be the reason why I obtained low numbers of full length transcripts when I used genomic sequences and protein sequences with large evolutionary distances between them (**Chapter 3**).

Evolutionary analyses – algorithms and statistics

Neighbour joining

Neighbour joining (NJ) [115] is a heuristic "star-decomposition" method for reconstructing phylogenetic trees from evolutionary data. The algorithm takes in as raw input a matrix containing the pairwise evolutionary distances between all sequences, and attempts to build a tree with the minimal internal branch length. It starts with a "star-like" tree that has no internal branches (Step 1 in **Figure 12**), and as it adds its first internal branch it calculates the length of the resulting tree and takes the pair of external nodes which results in the smallest sum of branch lengths (Step 2 in **Figure 12**). This then becomes the starting point for the next step (Step 3 in **Figure 12**) and pairs of external nodes are joined such that each pair has the least distance to each other and to the rest of the tree nodes. The algorithm sequentially connects all the external nodes of the tree by choosing the smallest sum of branch lengths, thus resulting in the shortest tree. As the tree does not assume an evolutionary clock, it is an unrooted tree. NJ trees thus provide a good estimate of the "minimum evolution" tree [115]. Minimum evolution trees have a topology which gives a minimal tree size due to the least total sum of branch lengths.

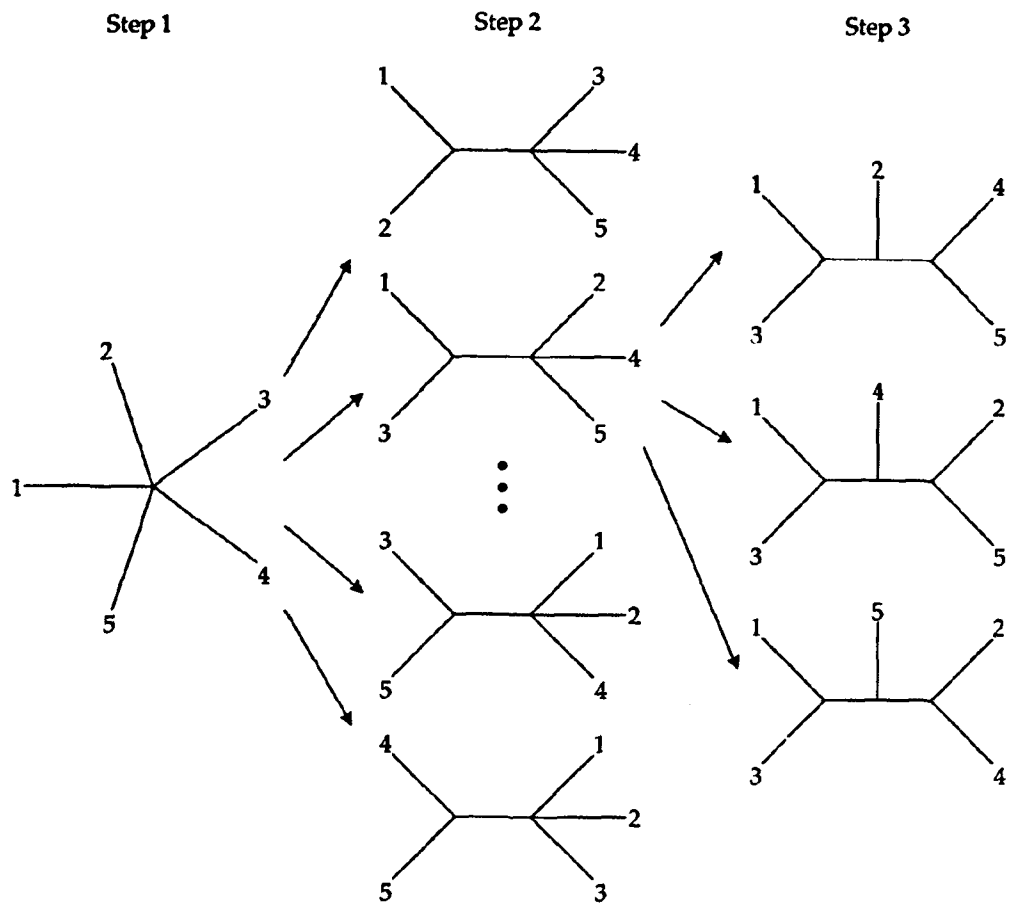


Figure 12. Heuristic tree selection using the "star decomposition" method. At each step, pairs of external nodes are joined only if they give the smallest sum of branch lengths. The best tree found at each step becomes the starting point for the next (figure taken from [116]).

Due to its "greedy" algorithm that constructs trees in a step-wise fashion, the NJ method may not find the true tree topology with least total branch length. But because this neighbour-joining method is able to handle large number of sequences in a short computational time, it is one of the methods most commonly used to construct a tree close to the optimal tree. In addition, the trees derived from the joining-joining method are used in progressive sequence alignments, for instance in the multiple sequence alignment tool, CLUSTALW (see above). In **Chapter 5**, I use the NJ method to construct phylogenetic trees from orthologous sequences.

Bootstrapping

In **Chapter 5**, I used the common technique of bootstrapping to evaluate the reliability of the inferred phylogenetic trees [117, 118]. The inferred phylogenetic tree is compared to a large number of trees generated from simulated multiple alignments derived from the original data set. Based on the assumption that columns (sites) in the alignment are independent, sites in the new simulated alignments are chosen randomly from the original with replacement. Thus a site can be chosen more than once or not at all. For each of the simulated alignments, which are of the same length as that of the original, a tree is constructed using the same method. The topology of each simulated tree is compared to the inferred tree by checking if the interior branches partition similarly, and the proportion of this occurrence is used to provide a statistical confidence supporting the interior branches of the tree. Usually called the *bootstrap confidence value*, a value of 0.95 (or 95%) [117] and above indicates that an interior branch is considered significant.

Programs used to infer positive selection

Estimating selection: the ω (K_A/K_S, dN/dS) ratio.

Using the definitions of Yang and Nielsen [95] and ω as described in **Chapter 1**, the ω (dN/dS, K_A/K_S) ratio estimates the selective pressure at the protein level by measuring the proportion of non-synonymous (amino-acid changing) substitutions per non-synonymous site to synonymous (silent) changes per synonymous site in protein-coding genes. By comparing the rates of fixation of these two mutations, the effect of

natural selection on DNA sequence evolution can be better understood [95]. There are several tools currently available which attempt to estimate mutation rates, with much controversy over each method's performance (see review by Nei [119]). In this thesis, two different methods were used to calculate ω – CODEML [95] and Sitewise Likelihood Ratio (SLR) [120]. Both CODEML and SLR predict the selective forces at individual codon (amino acid) sites. In **Chapters 3** and **6**, I used CODEML to estimate the selection strength on groups of genes across different lineages. In **Chapter 3**, I also used SLR as an alternative to the conservative CODEML to detect positive selection among closely related yeast genes. The basic concepts behind each of the three methods are discussed below.

CODEML

Codons are subject to a variety of selective pressures resulting in diverse ω values in a single protein. CODEML [95], found in the "Phylogenetic Analysis by Maximum Likelihood" (PAML) package (version 3.14), implements a number of models whose statistical distributions are used to detect heterogeneity of ω ratios among sites. Positive selection at individual sites is detected by CODEML often only when many of the lineages in the phylogeny are affected. If only a few lineages are affected by positive selection, CODEML may not predict this if purifying selection dominates.

To explain the underlying theory of CODEML, the probability theory of the Markov process of codon substitution will be discussed briefly. This theory forms the basis for the maximum likelihood estimations of dN and dS rates in a phylogeny.

Markov model of codon substitution

CODEML uses a (continuous-time) Markov process to model substitutions among the sense codons in a protein-coding sequence [95, 121, 122]. Here, the states of the Markov process are the 61 amino-acid coding codons. (The three stop codons are not considered as substitutions as these are usually lethal to the protein and hence would not be observed.) The Markov process is characterized by the matrix $Q = \{q_{ij}\}$, where q_{ij} is the substitution rate from sense codon i to sense codon j ($i \neq j$). The codon substitution model within CODEML is thus described by a 61 x 61 rate matrix where $Q_{ij}\triangle t$ represents the transition probabilities that codon i would change to codon j over a small interval of time $\triangle t$. The property of a Markov model is that the progression from one state to another depends only on the current state and not on past states. Mutations are thus assumed to occur independently among the three codon positions and only one position is allowed to change instantaneously.

The Markov model used in CODEML accounts for the transition-transversion bias, biases in codon and base frequencies and in unequal synonymous and non-synonymous substitution rates. As transitions (A to G or C to T) are known to occur more frequently than transversions (see above), the substitution rate (μ) is multiplied by the parameter κ, the transition/transversion rate ratio, if the change involves a transition. Likewise, to account for the codon usage bias and nucleotide composition in general, the substitution rate of change from i to j is multiplied by the equilibrium frequency of codon j (π_j). π_j is calculated using the observed frequencies at the three codon positions. I used the F61 option available in the PAML software package where all the codon frequencies are used as free parameters [123], thus the codon usage bias is calculated from the sequences by CODEML. Finally, to account for unequal

synonymous and non-synonymous substitution rates, the rate is multiplied by the non-synonymous/synonymous rates ratio ω. These parameters under the codon substitution model are estimated by the maximum likelihood method (see below). The above biases (from the different types of transition/transversion rate ratios), and how they are accounted for are summarized in **Figure 13**, with μ representing the substitution rate.

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions} \\ \mu\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transversion} \\ \mu\omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transition} \end{cases}$$

Figure 13. Transition probabilities in the Markov model of codon substitution (taken from [121]). ($i \neq j$) represents the rate of substitution from codon i to j. π_j and κ are parameters which characterize processes at the DNA level, and selection at the protein level is modified by the parameter ω.

Detection of positive selection using maximum likelihood estimations

The goal of maximum likelihood (ML) estimation is to find the best fit of the data to a given model. Model comparison can then be carried out using the log-likelihood values of the two nested models being compared. This is known as the likelihood ratio test (LRT). In CODEML (version 3.14), the test for positive selection is a two-step process. First, an LRT is used to test the entire gene for positive selection. The LRT compares a null model which does not allow for $\omega > 1$ with an alternative model which does. Finally, if both the alternative models (see below) fit the data

significantly better, then identification of putative positive sites is carried out using the Bayes empirical Bayes approach [124].

In the first step, a nearly neutral model (model M1a) that allows ω to vary between 0 and 1 ($0< \omega <1$) is compared with a selection model (model M2a) by the LRT. Model M2a allows an additional category of positively selected ($\omega > 1$) sites. Positive selection is inferred when M1a (nearly neutral model) is a poorer fit to the data than is M2a (selection model). The second LRT compares M7 which is the null model where ω is assumed to be beta-distributed among sites ($0 < \omega < 1$) and M8, the alternative selection model which allows an extra category of positively selected sites ($\omega > 1$). By allowing a beta distribution, the ω distribution is allowed to be flexible and to take any shape. This accommodates the ω distribution shape that is likely to occur in real data sets [95]. A normal distribution, in contrast, requires data to be distributed symmetrically about a point.

Several other models are also implemented within CODEML but models M1a, M2a and another pair, M7 and M8 are sufficient to infer positive selection [95, 125, 126]. Once the LRT is deemed significant at the appropriate degrees of freedom, the Bayes empirical Bayes (BEB) statistical procedure is then used to estimate the probability at each site that it belongs to the $\omega > 1$ category. Sites containing a posterior probability of a predetermined cut-off, for example, > 95% (i.e. $p > 0.95$) of belonging to this category are referred to as "positively selected" (or ω^+) sites. By using an empirical Bayes approach, identifying sites which are under positive selection is possible even if the average ω over all sites is much less than one [95, 125-127]. The added advantage of using the BEB procedure method is that sampling errors which occur in

small data sets during the maximum likelihood estimation of the parameters (in the codon model) are better accounted for [124]. These errors are negligible when large data sets are used.

SLR

New methods of identifying positive selection are continually being published. The most recent tool used in this thesis is the "Sitewise Likelihood Ratio" (SLR) method from Massingham & Goldman [120], which the authors claim represents an improvement over CODEML.

Like CODEML, SLR uses a ML method that estimates the likelihood of a particular phylogenetic tree based on the Markov codon substitution model. CODEML and SLR are both based on the same probabilistic model that assumes codons evolve independently of each other through nucleotide substitutions. The main difference between CODEML and SLR is that CODEML first conducts a likelihood ratio test between a null model and a selection model; only if the selection model shows a significantly higher ML value than the null model is positive selection detected at each site. This is achieved by finding if the Bayesian posterior probability of $\omega > 1$ is greater than, for example, 95%. SLR, on the other hand, makes no assumption about the overall distribution of ω but simply conducts a site-wise likelihood ratio test, testing the null model of neutrality ($\omega = 1$) against the alternative model of $\omega \neq 1$ at each site. Then sites with $\omega > 1$ associated with a significant probability (p-value < 0.05) against the null hypothesis of neutrality are accepted as positively selected residues.

SLR thus does not depend on the use of different models to estimate how ω varies along the sequence, which reduces the errors and leads to fewer false positive assignments [120]. SLR also claims to be more robust in handling data that would otherwise fall outside of the models in CODEML [120].

It is worthwhile to note that the probabilities obtained from CODEML and SLR are different and cannot be compared directly. CODEML produces a Bayesian posterior probability whilst SLR produces a *p*-value which provides an upper bound to the true probability that the result is a false positive [120]. Also, corrections for comparisons (or tests) on multiple codons which were used for CODEML (Wong, Yang et al. 2004) were adopted for the SLR method but these resulted in SLR being overly conservative. These multiple testing corrections assume that all sites evolve neutrally and thus have the same probability of falsely being assigned as evolving by positive selection. This does not hold because a purifying site has a lower chance of being falsely assigned as being under positive selection than a neutral site. Hence, the results presented in this thesis were not subjected to the multiple correction step available in SLR.

Evolver

Computer simulations have been a common method used to assess the accuracy of ω^+ site prediction programs [120, 125, 127]. These simulations estimate the amount of error one would expect from the probabilistic prediction programs and gives an indication of how conservative the predictions are.

Evolver can generate datasets of nucleotide, amino acid and codon sequences using a user-defined tree topology and branch lengths. In **Chapter 3**, I used Evolver (PAML package) to determine the false positive prediction rate on the *Saccharomyces* sets of genes. This was done by generating neutrally evolving ($\omega = 1$) codon sequences. Evolver first generated a root sequence from the codon equilibrium frequency specified by user input. Then, the program "evolves" each site along the branches of the tree independently according to the parameter values that were defined (for instance, values of κ, ω, tree branch length of the phylogeny examined and tree topology) (PAML documentation http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf).

General statistics used in the project

Hypergeometric distribution function

The hypergeometric distribution was used in **Chapters 3** to **5** to test for the over-representation of certain characteristics, for instance, Gene Ontology [128] categories or ω^+ sites within a sample size. The hypergeometric distribution models the number of successes in a fixed sample size drawn without replacement from a finite population. This is calculated as the discrete probability of obtaining k items with a certain characteristic from a sample size of n items. This sampling occurs in a population of size N containing D number of items with that particular characteristic [129]. The probability of obtaining k items is given in **Equation 2**.

$$\Pr(k \mid N, D, n) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}$$

Equation 2. Probability of obtaining *k* items from a sample size of *n* items without replacement. The binomial coefficient $\binom{D}{k}$ means "*D* choose *k*", or choosing a sample size of *k* items from a population of *D* items.

Kolmogorov-Smirnov (K-S) test

For a single sample of data, the non-parametric Kolmogorov-Smirnov (K-S) test can be used to test if the sample data are consistent with a specified distribution function. When there are two samples of data, the K-S test can be used to test if the two samples had been drawn from the same distributions. This useful test makes no assumption about the underlying data distribution and is powerful in detecting differences in the sample population. I used the K-S test in **Chapter 5** to determine if underlying distributions of stage-specific data of the schistosomes were significantly different.

Kruskal-Wallis test

The non-parametric Kruskal-Wallis (K-W) test is used to compare three or more samples of data. The null hypothesis is that all the sample populations have identical distribution functions and this is tested against the alternative hypothesis that at least two of the samples differ only with respect to their medians [130]. The K-W test is used in **Chapter 5** to test if the medians of each of the schistosome stages differed significantly.

Fisher's exact test

The Fisher's exact test is used to test the significance of the association between two categorical values in a two by two (2x2) contingency table. This is under the null hypothesis that there is no association between the two variables (*i.e.* they are independent) and is an alternative to the Chi-square test. Whereas the Chi-square test relies on a large sample approximation, the Fisher's exact test is based on exact probabilities from a specific distribution (the hypergeometric distribution). As there is no lower bound for the data that are needed in a Fisher's exact test, this test is useful for highly imbalanced tables. For instance, if one or two cells of the table have counts in the thousands and the other cells have numbers less than five, this test can be used. As long as there is a data value in each row and in each column, this test can still be used even when one of the cells of the table has a zero in it. The usefulness of this test is shown in **Chapter 6** where I test for the enrichment of fast evolving genes expressed in specific regions of the brain.

Multiple testing

When several dependent or independent statistical tests are being performed on the same data set, I used both the Bonferroni correction and the False Discovery Rate to account for multiple testing. Multiple testing corrections adjust the *p*-value to account for the number of comparisons being performed to decrease the number of false-positives. In **Chapters 3** to **6**, I show examples of multiple testing correction being carried out.

Bonferroni correction

The Bonferroni correction is the simplest and most conservative approach to account for multiple testing. It accounts for the number of tests performed by adjusting the pre-determined *p*-value cut off. This is done simply by dividing the *p*-value by the number of tests *n* done (*p*-value / *n*) to obtain a new significance threshold, α. Hence, only if the *p*-value for each test is now less than α is it deemed significant.

FDR

Controlling the false discovery rate (FDR) [131] is another statistical method to account for the multiple comparisons problem. Hence, instead of controlling the chance of any false positive like the Bonferroni method approach, FDR controls the expected proportion of false positives in a set of predictions. For example, if an algorithm returns 100 genes with an FDR of 0.30, then 70 genes would be expected to be correct. The FDR is very different from a *p*-value and a high FDR can be tolerated and may be more meaningful than a high *p*-value. In the above example, if there were 1000s of genes on an array, 100 predictions in which 70 genes were correct may be useful. A *p*-value of 0.30 is generally not acceptable in most circumstance. In all, controlling the FDR keeps the number of false negatives low while controlling for the number of false positives.

CHAPTER 3: Predicting positive selection among *Saccharomyces* genes

Summary

This chapter discusses a comparative genomics study of *Saccharomyces* genes that aims to determine the different selective pressures that have acted on different regions of its sequences, with a particular focus on positively selected sites. This study makes use of four genomes from the *sensu stricto Saccharomyces* species that have been sequenced to high coverage [20, 32] and also that of *S. cerevisiae* which has been well annotated. These resources allow detailed functional interpretations of evolutionary analyses.

I have identified functionally related groups of genes that exhibit positive selection at single nucleotide sites. These groups are over-represented in two categories of cellular functions: growth and defence. Equivalent categories in the mammalian genomes have also been reported to exhibit positive selection [62, 132-134], although for the category of growth, different lineages may show positive selection whilst other lineages may not [134, 135]. I present evidence that the rapid sequence divergence seen among these yeast genes is the result of adaptation rather than being a chance observation arising from the large number of statistical tests employed.

Introduction

Detecting positive selection in large scale studies

Nucleotide substitutions in eukaryotic genomes, in particular for species of large effective population sizes, have frequently been subject to positive selection [136, 137]. For different species, positive selection affects different biological processes [138]. This reflects the diversity of their ecological niches and of the different challenges for survival and reproduction. Among multicellular organisms, the substrates of positive selection have often been found among genes involved in reproduction, chemosensation, toxin degradation and response to infection [139-141]. Such rapidly evolving genes have usually been identified by studying either gene families or single genes. Small-scale studies rely heavily on prior scientific knowledge of these genes and often insignificant associations are generated when corrections are applied for all possible statistical tests on all genes in a genome.

Conversely, large (or whole) genome studies (for example, comparing among bacteria [142], fungi [20, 71] or mammals [62]) focus on detecting positive selection across the genome by using an entire gene sequence, rather than each codon in turn. This is done by estimating the rates of non-synonymous substitutions per non-synonymous site (dN), and of synonymous substitutions per synonymous site (dS), and their ratio ($\omega = dN/dS$) between pairs of gene alignments. An unfortunate consequence of this is the failure to infer episodes of positive selection when only a minority of codons are affected. Moreover, only a tiny fraction of genes exhibit ω values significantly >1 (indicating positive selection) when estimated across their entire sequences [139].

Only by examining individual sites of genes is one able to detect where positive selection has occurred.

Detecting positive selection in *Saccharomyces sensu* stricto yeasts

I was interested in identifying genetic substrates of adaptation within the *Saccharomyces* genus. This study was heavily reliant on the high quality genomic sequence of *S. cerevisiae* [1] and on four sequenced genomes of the *sensu stricto* yeasts: *S. bayanus*, *S. kudriavzevii*, *S. mikatae* and *S. paradoxus* [20, 71]. The large-scale estimation of evolutionary rates across the five *sensu stricto* yeasts needed to consider both the most appropriate methods to apply and possible confounding effects in their application. In particular, the synonymous sites of some yeasts' genes maybe under strong translational selection, due to an increased pressure for translational accuracy [143, 144]. Thus inferring positive selection at synonymous sites under translational selection may be unreliable [145].

In order to address this problem, I first used CODEML [146], which takes into account codon usage bias, for inferring positive selection among the yeast sequences. However, with five sequences at the divergences shown by these *Saccharomyces* genes, detection of positive selection by CODEML, although reliable, is known to be extremely conservative [125, 127, 147]. Indeed, CODEML predicted positively selected codons, with posterior probability > 0.95) for only two of 1014 *Saccharomyces* genes tested (data not shown). Instead, I chose to apply the "Sitewise Likelihood Ratio" (SLR) method of Massingham and Goldman [120] which uses the same codon substitution model as CODEML. The main reason for choosing SLR over

other methods, such as those of Nielsen and Yang [122], was because of its statistical power to predict positive selection combined with an effective control of the false positive prediction rate.

As SLR was designed to test for non-neutral evolution at single sites, I used it to test for positive selection at each aligned codon from 2,869 multiple alignments of orthologous genes from the five *Saccharomyces sensu stricto* species. Defining positively selected (ω^+) sites as codons predicted by SLR with $p_{SLR} < 0.01$, an assumption that false positive predictions of positively selected sites are randomly distributed among all sites was used. This led to the identification of 27 alignments that were significantly enriched in ω^+ sites. The genes in which the ω^+ sites were detected were significantly enriched in functions related to growth and defence. Due to the numerous *p*-values used in this chapter, I use subscripts (SLR in this case) to distinguish among them.

Methods and Materials

Obtaining sequence data

Sequence data for the five species of *Saccharomyces stricto sensu* yeasts – *S. bayanus*, *S. paradoxus*, *S. cerevisiae*, *S. kudriavzevii* and *S. mikatae* were taken from analyses performed by Cliften and colleagues [71] at Washington University (WashU), and Kellis and colleagues at the Massachusetts Institute of Technology (MIT) [20] (**Table 1**). Both groups sequenced and studied the genome sequences of

the *Saccharomyces* species to varying degrees to elucidate regulatory elements, identify genes and perform nucleotide and protein analyses.

Species	Data Source	Data Type
S. cerevisiae	SGD	DNA, protein
S. paradoxus	SGD, MIT	DNA, protein
S. kudriavzevii	SGD, WashU	DNA, protein
S. bayanus	SGD, WashU, MIT	DNA, protein
S. mikatae	SGD, WashU, MIT	DNA, protein

Table 1. Summary of data sources for the five *Saccharomyces sensu stricto* species of yeast. Sequences from the MIT and WashU groups are deposited at SGD (ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes). Data types from SGD consisted of the DNA and the amino acid sequences of the genes. WashU is the abbreviation for Washington University, MIT is the Massachusetts Institute of Technology, and SGD is the acronym for the *Saccharomyces* Genome Database.

Results from the two groups have revised the original gene count found in the *Saccharomyces* Genome Database (SGD) maintained at Stanford (http://genome-www.stanford.edu/Saccharomyces/). **Table 2** summarises the starting and final gene count from the two groups and provides information of when the data were obtained from SGD. Kellis *et al.* ended with a lower gene count mainly owing to their criterion that open reading frames (ORFs) contain at least 100 amino acids.

	Cliften *et al.* (MIT)	Kellis *et al.* (WashU)
Version of SGD sequences	May 2002	July 2002
Starting number of sequences	6331	6275 (ORFs encoding ≥ 100 amino acids)
Final gene count	5773	5538 (ORFs encoding ≥ 100 amino acids)

Table 2. Overview of the data from the WashU and MIT group. Kellis *et al.* (WashU) considered only *S. cerevisiae* genes whose Open Reading Frames (ORFs) encoded at least 100 amino acids.

Data storage

All sequence data for this study were stored in a MySQL database. This is an open source relational database management system (http://www.mysql.com). Data stored in these tables are retrieved through the Structured Query Language (SQL). SQL is a standard and interactive programming language used for obtaining information from, and updating, a database. Throughout this project, SQL was used from within the Practical Extraction and Reporting Language (PERL), a robust programming language, used to retrieve data from the database automatically. Numerous object orientated PERL modules were built to support the insertion and retrieval of data from the database.

Data Merger of MIT and WashU gene sets for *S. bayanus* and *S. mikatae*

To prevent redundancy, the MIT and WashU data sets were merged to obtain a union of the two groups and only ORFs that were ≥ 100 amino acids were retrieved. This applied specifically to two species, *S. bayanus* and *S. mikatae*, whose genomes were sequenced by both groups. Treating both the MIT and WashU data sets of *S. bayanus* (and *S. mikatae*) as separate species was found to be problematic as the number of paralogues and orthologues between the MIT and WashU data sets differed greatly.

This is shown in **Table 3** where a sample of different *S. cerevisiae* genes have different gene copy numbers from their *S. bayanus* homologue.

S. cerevisiae gene name	Number of copies in *S. bayanus* (MIT)	Number of copies in *S. bayanus* (WashU)
YAL013W	2	1
YBL017C	4	1
YBR023C	3	0
YBR266C	2	0
YBR273C	1	0
YBR275C	2	0
YBR277C	1	0
YBR297W	0	1
YBR298C	0	1

Table 3. A sampling of genes. Genes from the *S. bayanus* species are shown with their orthologues in *S. cerevisiae*. For example, the *S. cerevisiae* gene YAL013W has two homologues in *S. bayanus* of the MIT group and one orthologue in the WashU group. Some genes for *S. bayanus* (for example, YBR277C) are predicted by the MIT group but not by the WashU group, and *vice versa*.

In order to merge the MIT and WashU *S. bayanus* and *S. mikatae* gene sets, a representative orthologue of each *S.* cerevisiae gene was chosen from either the MIT or WashU sets using the highest bit scores from the alignment to the *S. cerevisiae* counterpart as a criterion. Using *S. bayanus* and its *S. cerevisiae* homologue as an example, the criteria are as follows:

1. Highest bit score density of the alignment between the *S. cerevisiae* and corresponding *S. bayanus* orthologue. Bit score density is the bit score from the BLAST alignment divided by the number of alignment positions. The alignment position length is obtained by counting the number of bases which align between the two sequences. This excludes gaps in the alignment and hence prevents large insertions from being counted.

2. If both bit score densities are identical from step 1, the longer sequence is chosen.
3. If both bit score densities and sequence lengths are identical, the MIT gene set is preferentially chosen to be the *S. bayanus* orthologue to the *S. cerevisiae* gene.

By merging the two groups of genes, the gene counts for both *S. bayanus* and *S. mikatae* within the two groups increased and redundancy of the data was reduced. For *S. bayanus*, by merging both the MIT and WashU groups, there was an increase of 11% in the gene count. The increase in gene count was greater for *S. mikatae*, where there was a 38% increase from just the WashU gene set and an increase of 11% from the MIT gene set (**Table 4**). For both species, there were about 16-17% of genes from both the MIT and WashU gene sets that had no homologues in *S. cerevisiae*.

	S. bayanus	*S. mikatae*
Total no. of genes	5502 (82 %)	5581 (83%)
No. of identical genes between MIT and WashU groups	4058 (61%)	2492 (37%)
No. of genes with no orthologue in *S. cerevisiae*	1180 (18%)	1083 (16%)
Ratio of MIT: WashU genes	4732 (71%) : 83 (1%)	4724 (71%) : 916 (14%)

Table 4. Summary of merging the MIT and WashU group of genes for the species *S. bayanus* and *S. mikatae*. Percentages are relative to the number of *S. cerevisiae* genes (6,703). The percentage of WashU genes used was relatively smaller than their MIT counterparts because the MIT gene was preferentially chosen over the WashU gene set.

Orthology assignments

Orthology relationships between *S. cerevisiae* and the four other *stricto sensu Saccharomyces* species (*S. bayanus*, *S. kudriavzevii*, *S. mikatae* and *S. paradoxus*) had previously been established by the MIT and WashU groups. I extended these orthology assignments to infer orthology relationships among the four *stricto sensu Saccharomyces* yeasts.

Using the 6,703 *S. cerevisiae* genes obtained from SGD as a reference set, sequences from the other four *Saccharomyces* species were assigned orthology relationships if they each exhibited orthology with their *S. cerevisiae* counterpart. **Figure** 14 illustrates this point further. Genes "Sm_bay_7081" in *S. bayanus* and "Sm_mik_6174" in *S. mikatae*, are both orthologues of the *S. cerevisiae* YER168C gene which then implies, in turn, that Sm_bay_7081 and Sm_mik_6174 are also orthologues of each other. The orthology relationships across all pairs for the five species, obtained using this method, were used for all subsequent experiments discussed in this thesis.

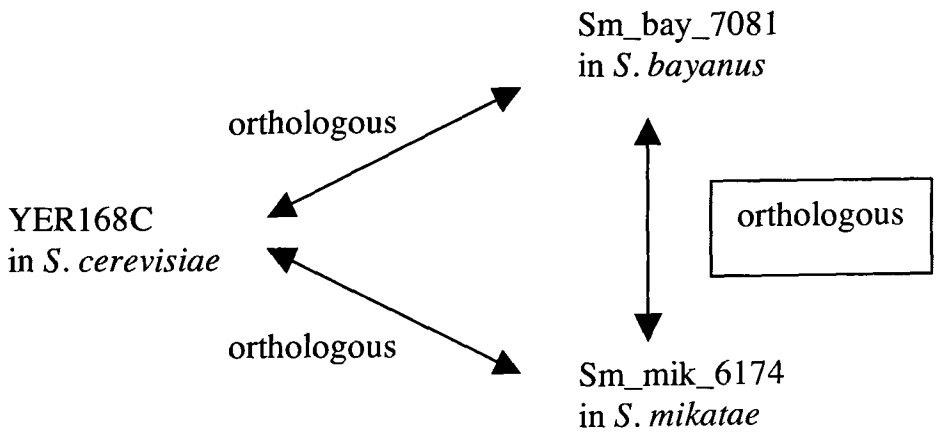


Figure 14. Illustration of an orthology inference. As both the *S. mikatae* gene (gene name in the database: Sm_mik_6174) and the *S. bayanus* gene (gene name in the database: Sm_bay_7081) are orthologues of *S. cerevisiae* YER168C, this implies that both Sm_mik_6174 and Sm_mik_7081 are also orthologous to one another.

Table 5 shows the results of the orthology assignment. For each of the species, the numbers of orthologues, paralogues and sequences without homology to any of the *S. cerevisiae* genes are shown. The percentages based on the 6,703 *S. cerevisiae* gene set are given in parentheses. Gene duplicates either arose through duplication events or else are artefacts of the assembly step. Orthologues in this table are sequences that had only a single gene orthologous to a *S. cerevisiae* gene.

Species	No. of orthologues to *S. cerevisiae* genes	No. of gene duplicates	No. of genes without homology to *S. cerevisiae* genes
S. kudriavzevii	3643 (54 %)	-	3060 (46%)
S. paradoxus	5299 (79%)	408 (6%)	996 (15%)
S. bayanus	5502 (82%)	-	1201 (18%)
S. mikatae	5581 (83%)	-	1122 (17%)

Table 5. Summary of orthology assignment. All percentages have been rounded to the nearest whole number.

Identification of truncated genes

The original sequence data of the five species were known to contain truncations, due to frame-shifts, indels or sequencing errors, thereby resulting in non-full length transcripts. A method was designed to detect truncated sequences and to target them for removal. This was important, as full-length transcripts could then be used to maximise information derived from estimations of nucleotide substitution rates.

In order to detect truncated sequences, the BLAST-2-Sequences (bl2seq) [148] program was used to align the sequences between all species pair combinations. The observed number of aligned positions $l_G(i,s)$ between orthologues from species i and s was compared to an expected number of such positions. This expected number was

derived from the average number of aligned positions in comparisons of orthologues from all other species.

More specifically, for each gene in species s, the average observed alignment position length L_O is given as

$$L_O^s = \frac{\sum_{i \neq s}^{N} l_G(i,s)}{(N-1)}$$ where N is the number of species.

The expected alignment position length L_E is given as

$$L_E^s = \frac{\sum_{i \neq s, i=1}^{N-1} \sum_{j \neq s, j=i+1}^{N} l_G(i,j)}{(N-1)(N-2)/2}$$

Hence, the ratio of observed : expected lengths is given by

$$\frac{L_O^s}{L_E^s} = \frac{\sum_{i \neq s}^{N} l_G(i,s)}{(N-1)} \Bigg/ \frac{\sum_{i \neq s, i=1}^{N-1} \sum_{j \neq s, j=i+1}^{N} l_G(i,j)}{(N-1)(N-2)/2}$$

The genes are sorted in descending order by their $\frac{L_O}{L_E}$ ratios. In the absence of truncations, this ratio is expected to be approximately 1. Truncations and imperfect gene predictions, however, result in ratios that are less than 1. Ratios greater than 1 are due to the truncated sequences affecting the average L_E. A threshold K, on this

ratio, was thus imposed to identify and then remove truncations and incomplete transcripts.

To select a suitable threshold K for removal of truncated sequences, multiple sequences alignments at various thresholds were examined. Multiple sequence alignments at the following thresholds of 0.05, 0.1, 0.15, 0.20, 0.5, 0.8, 0.9, 0.95, 0.975, and 0.99 were examined by eye. The resulting threshold of $K = 0.9$ was chosen. The multiple sequence alignments of sequences with $\frac{L_O}{L_E}$ values of 0.9 showed that the ratios differed from 1 only because of sequence divergence and not because of truncations. This also meant that aligned sequences were 90% similar in length to each other. The threshold of 0.9 was applied to orthologue gene sets, and only gene sets containing genes above this threshold were retained for subsequent analyses.

Predicting positive selection

To predict the selective forces at individual codon (amino acid) sites, the "Sitewise Likelihood Ratio" (SLR) method from Massingham & Goldman [120] was used with default parameters. SLR is described in detail in **Chapter 2.**

Simulations using Evolver

The program, Evolver, from the PAML package [146] was used to estimate the accuracy of the SLR predictions. As SLR predicts the deviation from neutrality of each site, I used Evolver to generate neutrally evolving sequences for each of the 2,869 orthologous gene sets. These 2,869 neutrally evolved sets contained the same lengths and numbers of sequences as their respective real data set, *i.e.* five sequences representing the five different *Saccharomyces* species per set, and with alignment

gaps retained. For each simulation, the following were provided as input to Evolver: the species tree (with branch lengths provided), ω (set to 1) and the transition/transversion ratio of each set. This randomized simulation of neutral sequence evolution was repeated ten times for each of the 2,869 gene sets. In summary, a total of 143,500 simulated sequences were generated.

Functional annotation of genes

Genes were functionally annotated using the three ontologies of the Gene Ontology (GO) database [128]: "Molecular function", "Biological process" and "Cellular component" categories. There are many categories in the GO database. To increase the statistical power of my analysis [149], I used a more general classification of terms which meant fewer classification categories. This was available in the form of GO-slim terms. These GO-slim terms are the parent nodes of GO terms. The hypergeometric distribution function (described in **Chapter 2**) was used to test if there were significant over- or under-representations of yeast genes in the 79 GO-slim categories (24 GO-slim categories from cellular component, 22 GO-slim categories under molecular function and 33 GO-slim categories under biological processes,

I also used two online specialised tools, TmHMM and SignalP, to further annotate the protein sequences of the genes of interest. TmHMM (version 2.0) uses hidden Markov models to predict transmembranes helices in proteins from their amino acid sequence. This application is hosted at http://www.cbs.dtu.dk/services/TMHMM-2.0/ [150]. SignalP (version 3.0) uses several artificial neural networks and hidden Markov models to predict signal peptide cleavage sites which indicate the presence of a signal

peptide (or non-signal peptide) in the amino acid sequence. The application is hosted at http://www.cbs.dtu.dk/services/SignalP/ [151].

Finding sequences in the PDB

BLASTP [80] was used to search the Protein Data Bank (PDB) [152] for tertiary structures of either *S. cerevisiae* sequences or their homologues. The options used were the SEG program to filter out low-complexity regions and only alignments with an *E*-value less than 1×10^{-5} were retained for analyses.

Relative solvent accessibility of positively selected residues

The relative solvent accessibility (RSA) of each amino acid residue was obtained by dividing the accessibility scores by the amino acid specific maximal accessibility values [153]. To predict the ω^+ residues' accessibility scores, the DSSP program (dictionary of protein secondary structure) [154] was used. DSSP uses structural information from the Protein Data Bank (PDB) to calculate the RSA scores for each amino acid residue. These RSA scores were then partitioned into three states as defined previously by Rost and Sander [153]: buried (<9% relative accessibility), intermediate (9%–35% relative accessibility), and exposed (≥36% relative accessibility). I expected that positively selected sites would be over-represented among exposed amino acid sites and thus have higher RSA values.

Results and Discussion

Homing in on a final dataset – unreported work that led to this chapter

All analyses in this chapter were repeated several times. In order to obtain the maximal number of orthologous gene sets with minimal information loss, I repeated the analyses with different numbers of *Saccharomyces* species; from nine to seven to five species from across the *Saccharomycetes* clade. Analyses were repeated again on the five least divergent *Saccharomyces* species with low complexity regions removed and then again when alignment gaps were removed (see below). Many results of these initial studies will not be reported here to improve the clarity of the chapter. However, as I spent significant effort on obtaining an optimal resulting gene set of five *sensu stricto* species, I would like to briefly mention why the original, more extensive set of yeast genomes was not used.

The original nine *Saccharomyces* species data set

Initially I carried out investigations using nine *Saccharomyces* species. I used four divergent species – *S. kluyveri, S. castellii, S. servazzi* and *S. exiguus*, in addition to five closely related species from the *sensu stricto* group: *S. cerevisiae, S. bayanus, S. mikatae, S. paradoxus* and *S. kudriavzevii*.

From nine to seven *Saccharomyces* species

Both *S. servazzi* and *S. exiguous* genomes contained only partially assembled contigs and I attempted to predict their genes using GeneWise [106] (data not shown). However, due to the low statistical coverage (0.2 – 0.3 fold) of these *S. servazzii* and *S. exiguus* genome assemblies and to their large evolutionary distances from *S. cerevisiae*, only low numbers of full-length transcript predictions were obtained:

2,128 (31%) *S. exiguus* and 1,823 (27%) *S. servazzii* gene models were obtained (percentages are based on SGD's *S. cerevisiae* gene count of 6,703). Thus it was decided that these two species should be excluded from further experiments.

The final set of five *Saccharomyces sensu stricto* species

I then repeated the project with the remaining seven yeast species. Starting from a recently revised number of 5,538 *S. cerevisiae* genes [20], these seven species' gene predictions yielded only 1,014 orthologous gene sets containing single genes found in all seven species, with all sequence lengths being at least 80% of the average gene length for that set. The median dS values given in **Table 6** provide an estimate of their evolutionary distances. However, beyond certain distances, saturation occurs and the numbers of substitutions which occurred cannot be reliably inferred. The distance where saturation occurs is debatable. Thus in this project, a distance of two was adopted; distances beyond two were considered saturated. *S. kluyveri* and *S. castellii* were too long branched (evolutionarily more distant from *S. cerevisiae*) and consequently, I opted to generate a less diverged set of genes based only on the five *sensu stricto* species.

Multiple alignments of *Saccharomyces stricto sensu* yeast genes

Closely related species show higher sequence similarity compared to highly diverged species. The greater similarity of the four *stricto sensu* genomes to that of *S. cerevisiae* allowed me to obtain many more orthologues present across all five species. By discarding *S. kluyveri* and *S. castellii*, I obtained 2,870 non-redundant orthologous gene sets, over twice the number achieved with seven species. All subsequent analyses were performed with these 2,870 gene sets.

	S. bayanus	*S. castellii*	*S. kluyveri*	*S. kudriavzevii*	*S. mikatae*	*S. paradoxus*
S. bayanus	-	-	-	-	-	-
S. castellii	7.53	-	-	-	-	-
S. kluyveri	8.4	8.93	-	-	-	-
S. kudriavzevii	0.82	7.63	8.44	-	-	-
S. mikatae	1.05	7.91	8.54	0.85	-	-
S. paradoxus	0.97	7.96	8.67	0.77	0.6	-
S. cerevisiae	1.1	8.17	8.52	0.89	0.74	0.36

Table 6. Matrix of median K_S values. The K_S values of *S. castellii* and *S. kluyveri* are in the two columns indicated by the black arrows. These values are large (>7.00) as both species are evolutionarily more distant from the other species.

Accounting for gaps and low complexity regions

All the alignment gaps in each gene set and compositionally-biased regions were discarded. I noted that SLR occasionally predicted positive sites for alignment columns containing gaps (**Figure 15**) despite their reduced information content. Hence I removed 50,603 alignment columns containing gaps, thereby reducing the number of sites examined to 1,140,116. Regions which were compositionally-biased were also removed from the *S. cerevisiae* sequence as the subsequent search against the Protein Data Bank (PDB) retrieved spurious alignments to these regions (see below).

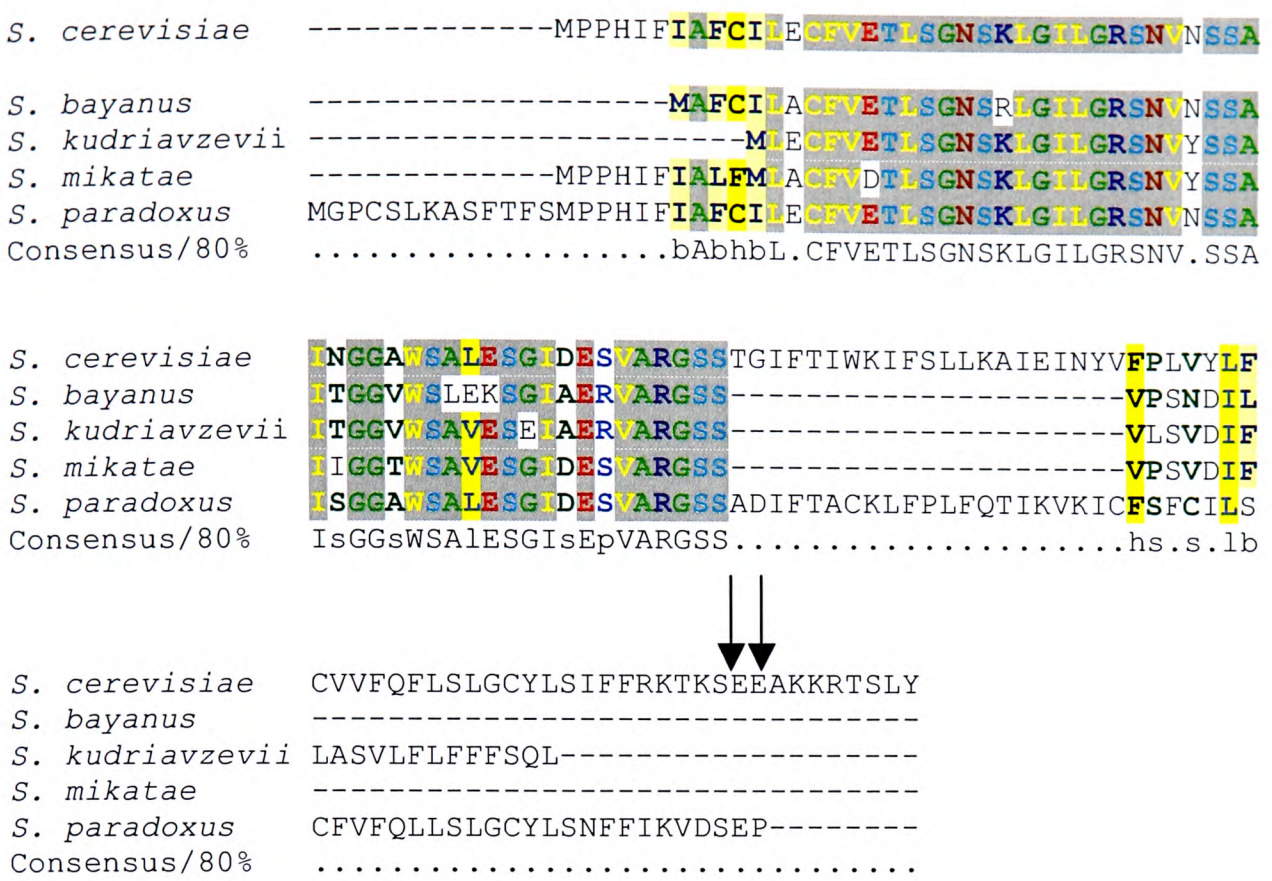


Figure 15. Amino acid alignment of the *S. cerevisiae* YBR124W protein and its orthologues in four other species. Predicted ω^+ sites are in positions 122 and 124 (indicated by the black arrows). These sites are within alignment columns with gaps and these columns are discarded.

SLR predictions

Upper-bound of SLR's false positive rate

The inference by SLR of positive selection at a single site is accompanied by an estimation of p_{SLR}, the probability that the site has evolved neutrally. As Massingham and Goldman note [120], distinguishing positive selection from neutral evolution represents the most challenging case and so p_{SLR} must be considered an upper-bound on the true probability of a false positive prediction.

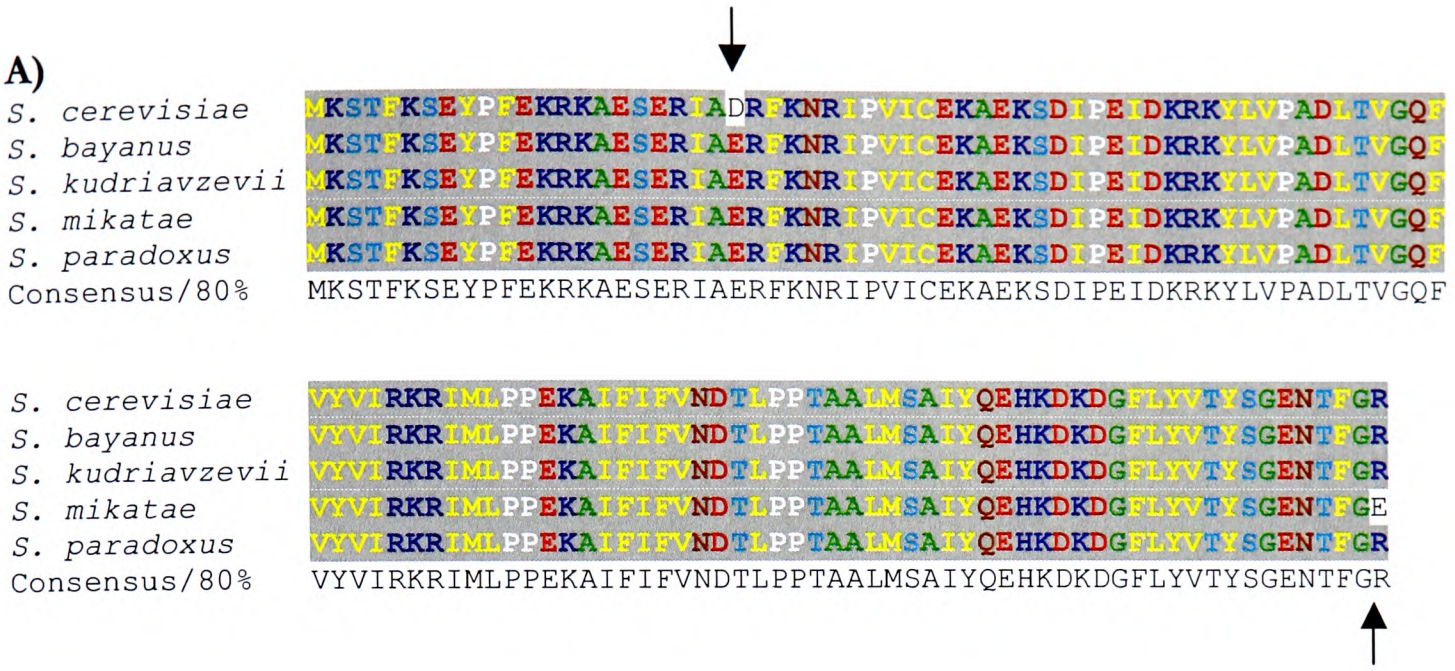
Each ungapped aligned position for each of the 2,870 orthologous gene sets was classified using SLR as having been subject to purifying ($\omega < 1$), neutral ($\omega = 1$) and positive ($\omega > 1$) selection. Such assignments were reported at three probability (*p*-values) thresholds: $p_{SLR} < 0.05$, $p_{SLR} < 0.01$ and $p_{SLR} < 0.001$. However, as a

conservative approach, I only accepted sites whose probabilities are less than 0.01 ($p_{SLR} < 0.01$).

Case study of when SLR fails

SLR failed to predict over the *YBL078C* gene set. The multiple alignments for these orthologues are shown in **Figure 16** below, generated by the program CLUSTALW [110] and coloured with the tool CHROMA [155]. It is likely that in this case, due to the small number (two) of non-synonymous substitutions, there was insufficient information for SLR to predict sites that had experienced positive selection. Hence, only 2,869 genes were available for subsequent analyses.

A)



B)

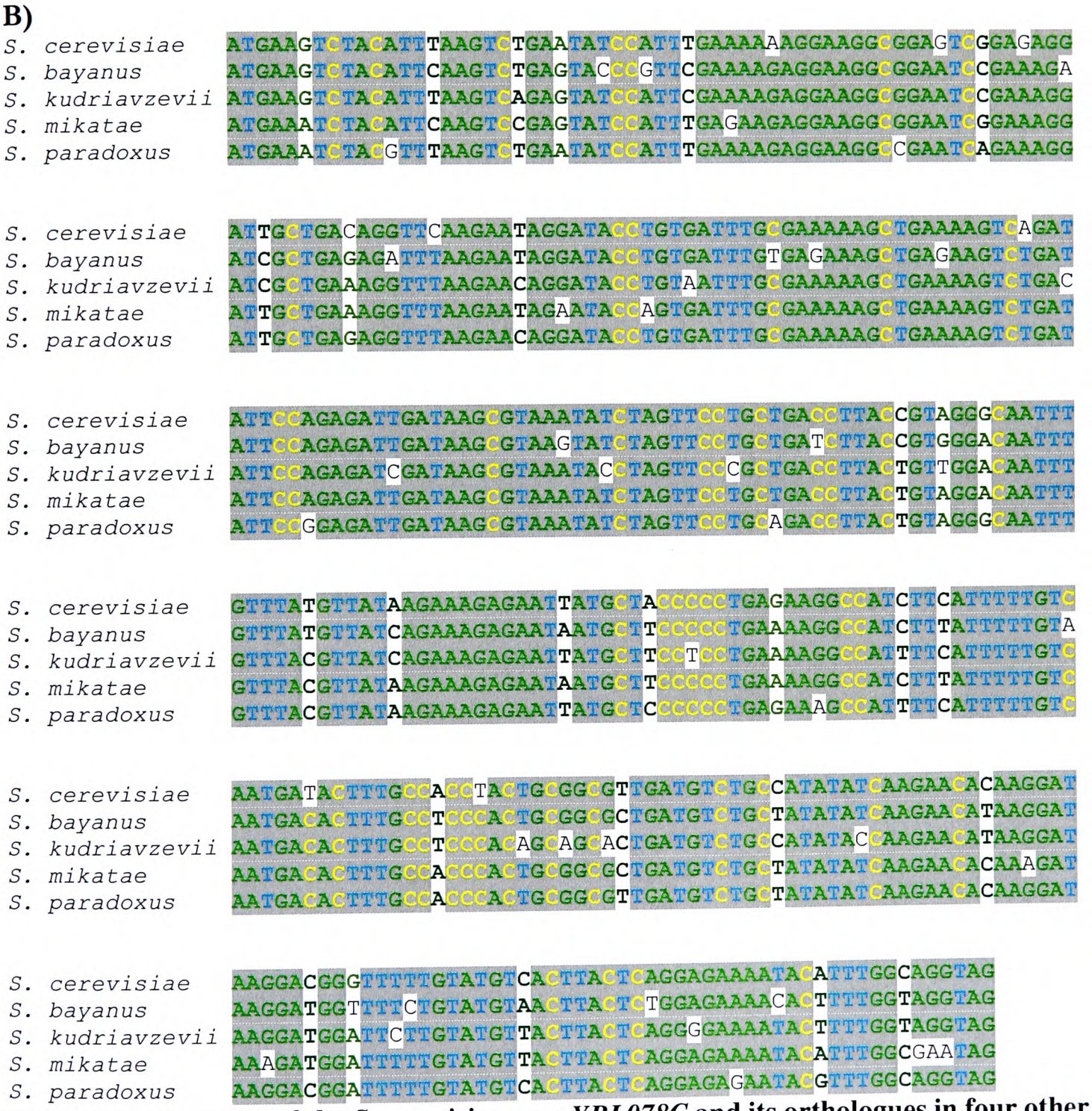


Figure 16. Alignment of the *S. cerevisiae* gene *YBL078C* and its orthologues in four other species. A) The proteins are highly conserved and two amino acid substitutions are indicated by the black arrows. Orthologue pairs' percentage identities from the pairwise alignments from CLUSTALW ranged between 99 and 100%. B) The 354bp of coding DNA for the corresponding proteins were also highly conserved (pairwise identities score ranged between 87 and 93%).

Summary of SLR predictions on 2,869 gene sets

The 2,869 gene sets from the five *sensu stricto* species provided a total of 1,140,116 aligned sites, excluding compositionally-biased regions and gaps in the alignment. Of these sites, SLR predicted 1,542 ω^+ sites (p_{SLR}-value < 0.05) from a possible 1,140,116 (*i.e.* 0.14% of sites were predicted as ω^+ sites). This decreased to 0.03% of sites for $p_{SLR} < 0.01$ (379 out of a possible 1,140,116 sites) and 0.004% of sites for $p_{SLR} < 0.001$ (49 out of 1,140,116 sites). The number of genes containing these predicted positive sites also decreased exponentially as the p_{SLR}-value threshold became more stringent. Numbers of predicted purifying, positive and neutral sites at the three *p*-values are shown in **Table 7**.

p_{SLR} threshold	Number of ω^+ sites	Number of neutral sites	Number of purifying sites	Number of all sites (purifying, neutral, positive) predicted with significance	Number of genes with ω^+ sites
-	39,431	6	1,100,679	1,140,116	2,653
0.05	1,542	0	798,955	800,497	854
0.01	379	0	472,446	472,825	300
0.001	49	0	52,813	52,862	42

Table 7. Summary of numbers of site types predicted by SLR for three different p_{SLR} thresholds.

Establishing the proportion of neutral sequences wrongly predicted as ω^+ sites

I used the Evolver program to simulate sequences matching the observed *Saccharomyces* (ungapped) genes in length, codon frequencies and transition/transversion rates. I then applied SLR to these simulated sequences and was able to confirm the accuracy of p_{SLR} estimation.

The simulation of neutral sequence evolution was repeated 10 times and produced a total of 1,140,116 sites. On average, at $p_{SLR} < 1\%$, 10,625 sites (*i.e.* 0.93%) from each simulation were found to be wrongly predicted as having been subject to positive selection (**Table 8**). This demonstrates that the p_{SLR} value accurately represents the probability of a false prediction for a site that has evolved neutrally. It is not, however, the probability of a false prediction for any site that has been evolving non-neutrally.

p_{SLR} threshold	Total number of predicted ω^+ sites (averaged over 10 runs)	Total number of sites	Percentage of predicted ω^+ sites among all sites
0.05	38,513.7	1,140,116	3.38%
0.01	10,625	1,140,116	0.93%
0.001	1,648	1,140,116	0.14%

Table 8. Numbers of SLR-predicted positive sites for simulated, neutrally-evolved sequence at different p_{SLR} thresholds.

False positive predictions of SLR revealed by simulations

It is particularly striking that for my experimental data set, the number of predicted ω^+ sites ($p_{SLR} < 0.01$) was 379 of 1,140,116 (*i.e.* 0.08%) ungapped alignment positions among 2,869 orthologous gene sets (**Table 7**). This number of ω^+ sites is approximately 28-fold lower than the number (10,625) (**Table 8**) of neutral sites wrongly predicted as ω^+ sites found from my simulations. This discrepancy arises because most sites will have been subjected to strong purifying selection rather than having evolved neutrally.

Proportions of neutral or ω^+ sites among yeast genes

Only a small minority of sites would have evolved neutrally and thus only these have the possibility of being mistaken by SLR as having evolved by positive selection. However, I can take advantage of these findings to estimate the upper bound for the fraction (η) of all 1,140,116 *Saccharomyces* alignment positions, lying outside of compositionally-biased regions, that have evolved neutrally. Under the worse case scenario that all predictions are false positives, the upper bound on η is 379/10,625, or approximately 3.57%. (379 is the number of ω^+ sites predicted by SLR from the 2,869 *stricto sensu* gene sets (**Table 7**) and 10,625 is the total number of predicted ω^+ sites from simulated data (**Table 8**), both at $p_{SLR} < 0.01$). Below I show that 85 of the 379 predicted ω^+ sites are enriched in 27 genes. Assuming that all of these are true positive ω^+ site predictions, this reduces the upper bound on η to 2.8%. Thus, greater than 97.2% of *Saccharomyces* alignment sites outside of compositionally biased sequences appear to have been subject to selection since the last common ancestor of the five *Saccharomyces* species.

Testing for the enrichment of ω^+ sites among *Saccharomyces* genes

As shown above, SLR predictions of positive selection contain both true and false positive assignments. Multiple testing corrections were not applied because, a *priori*, it was unclear what proportions of all sites had the opportunity of being falsely predicted as being subject to positive selection. Furthermore, I found that the length of a gene was positively correlated with its number of positive sites predicted by SLR ($r^2 = 0.53, p < 2.2\text{x } 10^{-16}$, **Figure 17**). Hence this meant that longer genes would contain more ω^+ sites simply because their numbers of neutral or adaptive (or both) sites are greater compared to shorter genes.

How many of these sites were true ω^+ sites? A large number of genes (300) (**Table 7**) contain only 1 or 2 ω^+ sites. If a gene contained one ω^+ site, what probability measure would be appropriate to estimate whether this was a true or false ω^+ site? Given the uncertainty of the false positive prediction rate, I conservatively selected all genes that were significantly enriched with ω^+ sites. More specifically, I tested if the number of positive sites in a gene was significantly elevated compared with that expected by chance. For this I used a null hypothesis that assumed false positive predictions to be uniformly scattered among all sites, and tested for this by applying the hypergeometric distribution and a probability (p_{OA}) upper threshold of 0.01. p_{OA} is the probability that the null hypothesis is false and the subscript "OA" is refers to the Over-Abundance of ω^+ sites in the genes examined.

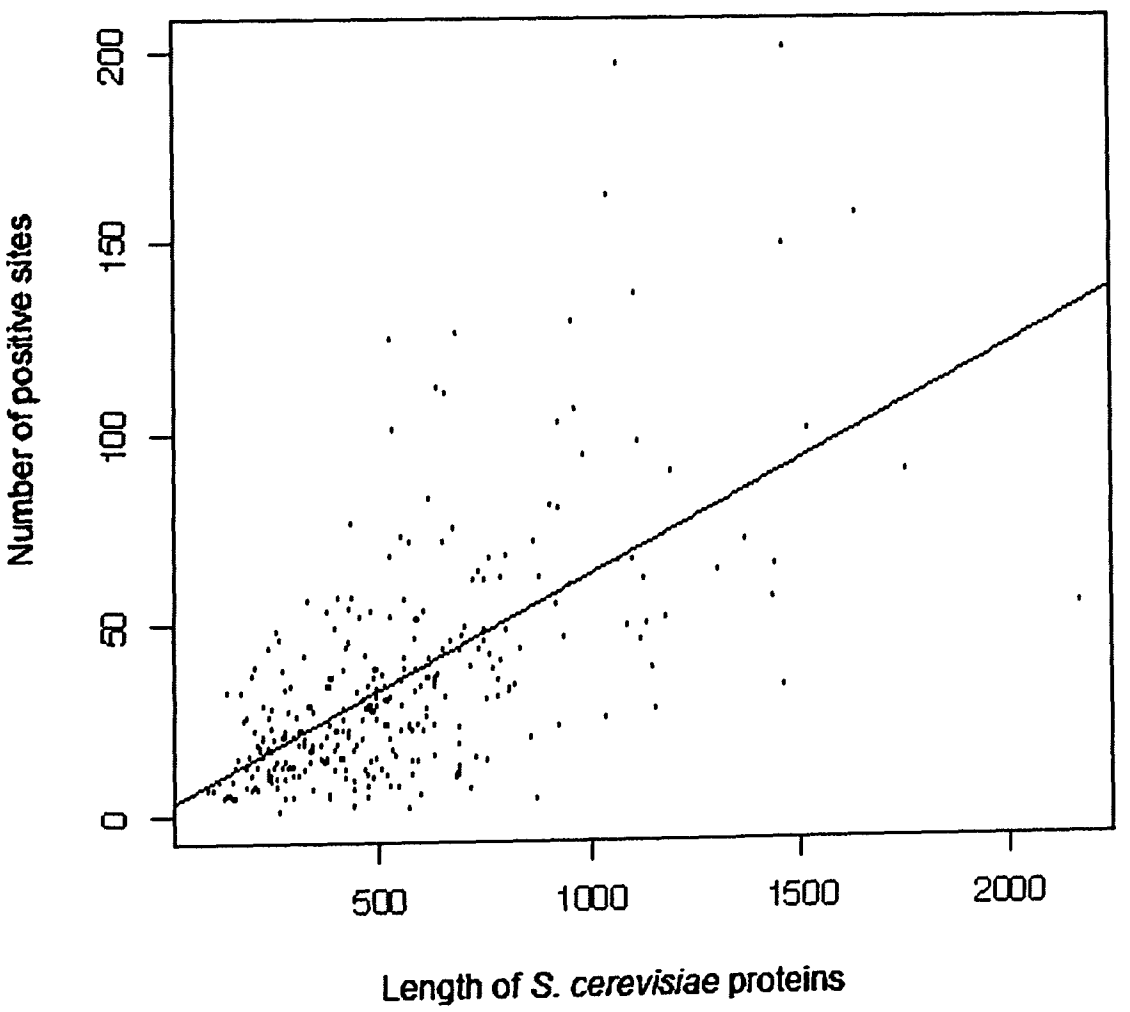


Figure 17. Significant positive correlation of ω^+ sites at $p_{SLR} < 0.01$ with the amino acid length of all genes ($r^2 = 0.61, p < 2.2 \times 10^{-16}$).

27 *Saccharomyces* genes were enriched with a total of 85 ω^+ sites at the conservative p_{SLR} and p_{OA} thresholds of 0.01. (The gene counts and numbers of ω^+ sites at the different thresholds of p_{SLR} and p_{OA} are tabulated in **Table 9**). 20 of these (mainly short) genes contained 2 or 3 ω^+ sites, but 7 genes possessed between 4 to 9 ω^+ sites (**Table 10**).

S. cerevisiae used to represent each orthologue set with ω^+ sites

As the *S. cerevisiae* genome is the best annotated among the *Saccharomyces* species, I use its genes from each of the 27 gene sets for the subsequent analyses. In **Table 11**, I provide a detailed summary of the 27 *S. cerevisiae* genes, their annotations found in SGD including the numbers of ω^+ sites found in each orthologous gene set and their protein lengths.

a) Gene count			
p_{SLR} / p_{OA}	0.05	0.01	0.001
0.05	90	45	41
0.01	38	27	11
0.001	17	9	1
b) ω^+ sites count			
p_{SLR} / p_{OA}	0.05	0.01	0.001
0.05	446	118	42
0.01	268	85	12
0.001	168	40	2

Table 9. a) Numbers of genes and b) the numbers of ω^+ sites predicted by SLR at p_{SLR} thresholds of 0.05, 0.01 and 0.001 each with an over-abundance of sites (p_{OA} thresholds of 0.05, 0.01 and 0.001). The numbers highlighted are at the threshold of 0.01, which I use to report the results in this thesis. 27 genes were significantly enriched and contained 85 ω^+ sites at p_{OA} and p_{SLR} threshold of 0.01.

Number of ω^+ sites	2	3	4	5	9
Number of genes	11	9	3	3	1

Table 10. Distribution of the number of ω^+ sites in the 27 genes defined at $p_{SLR} < 0.01$ and $p_{OA} < 0.01$.

Standard name	Aliases	Description from SGD; *Additional descriptions from NCBI searches*	Number of predicted ω^+ sites	Number of predicted ω^+ sites expected by chance	Length of protein
YBR005W	RCR1	Protein of the ER membrane involved in **cell wall** chitin deposition; may function in the endosomal-vacuolar trafficking pathway, helping to determine whether plasma membrane proteins are degraded or routed to the plasma membrane; *no homologue found outside of the Ascomycota phylum*	2	0.07	209
YBR052C	RFS1	Protein of unknown function; member of a flavodoxin-like fold protein family that includes Pst2p and Ycp4p; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm in a punctate pattern; *contains a flavodoxin_1 domain (Pfam); homologues found among in Escherichia coli and in other eukaryotes e.g. Oryza sativa (rice).*	2	0.07	206
YBR184W		Putative protein of unknown function; YBR184W is not an essential gene; *no homologue outside of the Ascomycota phylum*	3	0.17	505
YCL014W	BUD3|YCL012W	Protein involved in **bud**-site selection and required for axial budding pattern; localizes with septins to bud neck in mitosis and may constitute an axial landmark for next round of **budding**; *no homologue outside the Saccharomycetales order*	5	0.53	1589
YCR067C	SED4	Integral endoplasmic reticulum membrane protein, functions as a positive regulator of Sar1p probably through inhibition of GTPase activation by Sec23p; binds Sec16p, participates in vesicle formation, similar to Sec12p; *no homologue outside the Saccharomycetales order*	9	0.32	976
YDL146W	LDB17	Protein of unknown function; GFP-fusion protein localizes to the cell periphery, cytoplasm, **bud**, and **bud** neck; null mutant shows a reduced affinity for the alcian blue dye suggesting a decreased net negative charge of the cell surface; *homologue found in eukaryotes, e.g. D. melanogaster, Apis mellifera, Anopheles gambiae, Mus musculus and Homo sapiens. The human orthologue of this gene is the SH3 adaptor protein Spin90 gene*	2	0.16	487
YDR192C	NUP42|UIP1|RIP1	Subunit of the nuclear pore complex (NPC) that localizes exclusively to the cytoplasmic side; involved in RNA export, most likely at a terminal step; interacts with Gle1p; *matches other nucleoporins in other species in the Eukaryota lineage*	3	0.14	411
YDR282C		Putative protein of unknown function; *homologues found amongst most*	2	0.14	413

		lineages of "cellular organisms" e.g. chlamydias and Eukaryota. Protein may be involved in sporulation as it has a conserved domain "DUF155" defined as an "Uncharacterised ACR, YagE family COG1723". Another protein (YDL001W/RMD1) containing the same domain is required for sporulation and located in the cytoplasm.			
YDR465C	RMT2	Arginine methyltransferase; ribosomal protein L12 is a substrate; *homologues found in the lineage of "cellular organisms"*	2	0.13	406
YGR079W		Putative protein of unknown function; YGR079W is not an essential gene; *no homologues outside the Saccharomycetales order*	2	0.10	300
YHL044W		Putative integral membrane protein, member of DUP240 gene family; green fluorescent protein (GFP)-fusion protein localizes to the plasma membrane in a punctate pattern; *no homologue outside of the Saccharomyces genus*	2	0.08	226
YHR143W	DSE2	Daughter cell-specific secreted protein with similarity to glucanases, degrades **cell wall** from the daughter side causing daughter to separate from mother; expression is repressed by cAMP; *no homologue outside of the Saccharomyces genus*	3	0.10	311
YIL123W	SIM1	Protein of the SUN family (Sim1p, Uth1p, Nca3p, Sun4p) that may participate in DNA replication, promoter contains SCB regulation box at -300 bp indicating that expression may be cell cycle-regulated; *no homologue outside of the Ascomycota phylum*	5	0.14	430
YIR003W	AIM21	Protein of unknown function; may interact with ribosomes, based on co-purification experiments; green fluorescent protein (GFP)-fusion protein colocalizes with Sac1p to the actin cytoskeleton; interacts with the SH3 domain of Abp1p; *no homologue outside the Saccharomycetales order*	3	0.21	636
YJR004C	SAG1|AG(ALPHA)1	Alpha-agglutinin of alpha-cells, binds to Aga1p during agglutination, N-terminal half is homologous to the immunoglobulin superfamily and contains binding site for alpha-agglutinin, C-terminal half is highly glycosylated and contains GPI anchor; *no homologue outside the Saccharomycetales order*	5	0.21	625

YJR089W	BIR1	Essential chromosomal passenger protein involved in coordinating cell cycle events for proper chromosome segregation; C-terminal region binds Sli15p, and the middle region, upon phosphorylation, localizes Cbf2p to the spindle at anaphase; *no homologue outside of the Ascomycota phylum*	3	0.31	933
YKL023W		Putative protein of unknown function; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm; *no homologue outside of the Saccharomyces genus*	3	0.09	269
YLL021W	SPA2|PEA1|FUS6	Component of the polarisome, which functions in actin cytoskeletal organization during polarized growth; acts as a scaffold for Mkk1p and Mpk1p cell wall integrity signalling components; potential Cdc28p substrate; *no homologue outside of the Dikarya subkingdom in the Fungi lineage*	4	0.47	1410
YLR313C	SPH1	Protein involved in shmoo formation and bipolar bud site selection; homologous to Spa2p, localizes to sites of polarized growth in a cell cycle dependent- and Spa2p-dependent manner, interacts with MAPKKs Mkk1p, Mkk2p, and Ste7p; *no homologue outside of the Dikarya subkingdom in the Fungi lineage*	3	0.16	494
YLR330W	CHS5|CAL3	Protein involved in export from the Golgi to plasma membrane; involved in chitin biosynthesis through its role in Chs3p localization; interacts with Arf1p, Bch1p, Fmp50p, Bud7p, and Chs6p; *homologues found in the lineage of "cellular organisms"*	4	0.19	583
YLR390W-A	CCW14|SSR1|YLR391W-A	Covalently linked **cell wall** glycoprotein, present in the inner layer of the **cell wall**; *no homologues found outside the Saccharomycetales order*	3	0.07	219
YMR047C	NUP116|NSP116	Subunit of the nuclear pore complex (NPC) that is localized to both sides of the pore; contains a repetitive GLFG motif that interacts with mRNA export factor Mex67p and with karyopherin Kap95p; homologous to Nup100p; *no homologues found outside the Eukaryota superkingdom*	4	0.36	1095
YNL046W		Putative protein of unknown function; expression depends on Swi5p; GFP-fusion protein localizes to the endoplasmic reticulum; deletion confers sensitivity to 4-(N-(S-glutathionylacetyl)amino) phenylarsenoxide (GSAO); *no homologue outside of the Saccharomyces genus*	2	0.06	171
YNL195C		Putative protein of unknown function; shares a promoter with	2	0.08	242

		YNL194C; the authentic, non-tagged protein is detected in highly purified mitochondria in high-throughput studies; *no homologue outside the Saccharomycetales order*			
YNL260C		Putative protein of unknown function with similarity to a human protein overexpressed in oral cancers; localizes to the nucleus and cytoplasm; YNL260C is an essential gene; *May be involved in defence mechanisms. Contains two conserved domains – DUF1715 and COG3587. DUF1715 is a eukaryotic domain of unknown function and COG3587 is a restriction endonuclease involved in defence mechanisms. No homologues found outside the Eukaryota superkingdom*	2	0.05	162
YOL040C	RPS15|RPS21	Protein component of the small (40S) ribosomal subunit; has similarity to *E. coli* S19 and rat S15 ribosomal proteins; *homologues found in the lineage of "cellular organisms"*	2	0.05	142
YPL163C	SVS1	**Cell wall** and vacuolar protein, required for wild-type resistance to vanadate; *no homologue outside of the Saccharomyces genus*	3	0.08	255

Table 11. Functional descriptions of the proteins coded by the *S. cerevisiae* genes in the 27 gene sets. These are the descriptions of the proteins as taken from the *Saccharomyces* Genome Database (SGD) (2007 version). The first column contains the standard name as used in SGD and the names in the second column are aliases the genes are otherwise known by. The words in italics are additional information obtained from PSI-BLAST searches against the NCBI protein database. Words in bold correspond to the GO-slim categories in which these 27 genes are significantly enriched. The number of predicted ω^+ sites and lengths of the proteins are given in the corresponding columns. The predicted ω^+ sites expected by chance for each gene are calculated by taking the number of predicted ω^+ sites at $p_{SLR} < 0.01$ (379) and dividing it by the total number of sites (1,140,116) and multiplying it by the protein length.

Functional enrichments for GO-slim categories

The cellular functions of these 27 genes were examined using three Gene Ontology (GO) classifications: molecular function, cellular component and biological process. I reasoned that if all of these constitute false positive predictions, then their genes would exhibit no over-representations in GO-slim functional categories. However, at a false discovery rate of 5%, GO-slim annotations relating to cell budding, pseudohyphal growth and cytokinesis were found to be significantly enriched ($p_{GO} < 0.05$) (**Table 12A and Table 12B**). Based on the assumption that positive selection is more likely to act on some cellular functions than on others [62, 132-134], adaptive evolution, it would appear, has acted most on genes which regulate cellular growth by responding to environmental signals (see **Discussion**).

A) Cellular Component

GO-ID	GO term	*k*	*n*	*D*	*N*	p_{GO}-value	Genes
5933	Cellular bud	4	27	69	2790	3.91×10^{-3}	*BUD3*, *LDB17*, *SPA2*, *SPH1*
5618	Cell wall	5	27	45	2790	5.39×10^{-5}	*DSE2*, *SIM1*, *SAG1*, *CCW14*, *SVS1*

B) Biological Process

GO-ID	GO term	*k*	*n*	*D*	*N*	p_{GO}-value	Genes
7114	cell budding	3	27	37	2740	5.31×10^{-3}	*BUD3*, *SPA2*, *SPH1*
7124	pseudohyphal growth	3	27	33	2740	3.82×10^{-3}	*SPA2*, *SPH1*, *DSE2*
910	cytokinesis	5	27	48	2740	8.07×10^{-5}	*BUD3*, *DSE2*, *SPA2*, *SPH1*, *CHS5*

C) Molecular Function

GO-ID	GO term	*k*	*n*	*D*	*N*	p_{GO}-value	Genes
3674	molecular function unknown	17	27	968	2781	2.52×10^{-3}	*RCR1*, *RFS1*, *YBR184W*, *BUD3*, *SED4*, *LDB17*, *YDR282C*, *YGR079W*, *YHL044W*, *SIM1*, *YIR003W*, *BIR1*, *YKL023W*, *CHS5*, *YNL046W*, *YNL195C*, *YNL260C*, *SVS1*, *AIM21*

Table 12. Results of the GO-slim categories enriched in genes that contain an unusually high number of predicted ω^+ sites at a false discovery rate of 5%. These are categorised under the three main ontologies: A) Cellular Component, B) Biological Process and C) Molecular Function. From the definitions of the hypergeometric function in Chapter 2, *k* is the number of genes contributing to the particular trait sampled from a sample size of *n*, *D* is the proportion of the *k*-type item in the population *N*. *p*-values shown are corrected to three significant figures.

SGD Gene name	Position inside the cell	Coordinates predicted by TmHMM 2.0		ω^+ predicted by SLR	SignalP site predictions
		Start	End		
YBR005W	outside	1	39	31, 33	signal anchor (p = 0.999)
YBR005W	TMhelix	40	62		
YBR005W	inside	63	213		
YBR052C	outside	1	210	47, 48	non-secretory protein
YBR184W	outside	1	523	305, 381, 423	non-secretory protein
YCL014W	outside	1	1636	1026, 1050, 1129, 21, 843	non-secretory protein
YCR067C	outside	1	1065	463, 512, 542, 548, 562, 563, 564, 567, 980	non-secretory protein
YDL146W	outside	1	491	400, 441	non-secretory protein
YDR192C	outside	1	430	291, 303, 373	non-secretory protein
YDR282C	outside	1	414	217, 414	non-secretory protein
YDR465C	outside	1	412	48, 57	non-secretory protein
YGR079W	outside	1	9	137, 138	1-27, signal peptide (p = 0.889), cleavage site possibly between 27 and 28 (p = 0.629)
YGR079W	TMhelix	10	32		
YGR079W	inside	33	370		
YHL044W	outside	1	43	94, 104	signal anchor (p = 0.941)
YHL044W	TMhelix	44	66		
YHL044W	inside	67	72		
YHL044W	TMhelix	73	95		
YHL044W	outside	96	235		
YHR143W	outside	1	325	119, 182, 218	1-22, signal peptide (p = 0.96), cleavage site possibly between 21 and 22 (p = 0.754)
YIL123W	outside	1	475	127, 128, 129, 145, 188	1-19, signal peptide (p = 0.982), cleavage site possibly between 19 and 20 (p = 0.874)
YIR003W	outside	1	679	151, 233, 595	non-secretory protein
YJR004C	outside	1	629	357, 402, 434 521, 585	1-19, signal peptide (p = 0.959), cleavage site possibly between 19 and 20 (p = 0.950)
YJR004C	TMhelix	630	649		
YJR004C	inside	650	650		
YJR089W	outside	1	954	606, 611, 710	non-secretory protein
YKL023W	outside	1	277	40, 46, 54	non-secretory protein
YLL021W	outside	1	1466	860, 863, 978, 1174	non-secretory protein
YLR313C	outside	1	661	285, 324, 406	non-secretory protein
YLR330W	outside	1	671	470, 478, 522, 583	non-secretory protein

YLR390W-A	outside	1	238	146, 149, 150	1-22, signal peptide (p = 1.00), cleavage site possibly between 22 and 23 (p= 0.977)
YMR047C	outside	1	1113	621, 678, 876, 910	non-secretory protein
YNL046W	inside	1	99	28, 48	non-secretory protein
YNL046W	TMhelix	100	122		
YNL046W	outside	123	148		
YNL046W	TMhelix	149	171		
YNL046W	inside	172	172		
YNL195C	outside	1	261	103, 239	non-secretory protein
YNL260C	outside	1	198	55, 158	non-secretory protein
YOL040C	inside	1	142	7, 139	non-secretory protein
YPL163C	outside	1	260	134, 180, 210	1-19, signal peptide (p = 0.997), cleavage site possibly between 19 and 20 (p = 0.920)

Table 13. Transmembranes helices and signal protein cleavage sites in all 27 proteins by TmHMM and SignalP programs. TmHMM predicts if the protein is located on the inside (cytoplasm) or outside (exoplasm) of the cell, or if it has a transmembrane region (indicated by TMhelix in the second column of the table). SignalP predicts if there is a signal protein cleavage site. These secretory signal proteins are secreted outside of the cell and directs the post-translational transport of a protein to the endoplasmic reticulum in eukaryotes [151] [156].

Annotating genes with unknown function

Under the "molecular function" GO ontology, only one category – "molecular function unknown" was over-represented at $p_{GO} < 0.05$ (**Table 12C**). There are 19 genes found to be enriched within this category. Out of these 19 genes, nine encode proteins of unknown function or hypothetical predicted proteins. They are: *YBR052C*, *YBR184W*, *YDL146W*, *YDR282C*, *YGR079W*, *YIR003W*, *YKL023W*, *YLR330W*, *YNL046W* and *YNL195C*. Here, I have attempted to add to the functional description of these nine genes by using the PSI-BLAST interface provided at both SGD and NCBI. The PSI-BLAST searches against the UniRef90 protein dataset of UniProt (http://www.pir.uniprot.org/) [103] or the NCBI's GenBank non-redundant protein database respectively. I have also used online tools TmHMM and SignalP to predict if

these 27 genes code for transmembrane proteins (and if they are located inside/outside the cell) and/or secretory proteins respectively. Previous studies have shown that secreted proteins tend to show elevated ω values [22, 157]. **Table 13** summarises the results of the predictions of these two tools.

YBR052C – contains a close paralogue in *S. cerevisiae* named PST2/YDR032C. Based on the yeast 2-hybrid data of Uetz *et al.* [158], YDR032C binds to its paralogue RFS1/YBR052C. Pardo *et al.* [10] found that YDR032C is secreted by protoplasts and the authors speculate that it may be involved in cell wall construction. A protoplast is a plant, bacterial or fungal cell which has had its cell wall partially or totally removed by mechanical or enzymatic means. 67% of the YBR052C protein consists of a flavodoxin_1 domain (predicted by Pfam [159]) which is annotated in GO to have a molecular function of FMN (flavin mononucleotide) binding and oxidoreductase activity. This domain has been found in flavodoxin and nitric-oxide synthase. Also, YBR052C "localizes to the cytoplasm in a punctate pattern" (SGD website http://db.yeastgenome.org/cgi-bin/locus.pl?locus=YBR052C), and is predicted to be a non-secretory protein (**Table 13**).

YBR184W – only PSI-BLAST hit was to itself. No homologue found. YBR184W is predicted to be a non-secretory protein (**Table 13**).

YIR003W – has homologues within fungi only. YIR003W is predicted to be a non-secretory protein (**Table 13**).

YKL023W – only hit from PSI-BLAST was to itself. YKL023W is predicted to be a non-secretory protein (**Table 13**).

YDL146W – has homologues to hypothetical proteins in other fungal species. It contains a Pfam domain, "DUF2013", which is described as "protein of unknown function". YDL146W is predicted to be non-secretory (**Table 13**). In the first round of using PSI-BLAST against the GenBank protein database, YDL146W had hits to other metazoans: *Xenopus (Silurana) tropicalis* (frog) (*E*-value of 2 x 10^{-4}), *Ciona intestinalis* (*E*-value of 0.007), *Gallus gallus* (chicken) (*E*-value of 0.004). Subsequent rounds of PSI-BLAST detected homologies to human and mouse sequences. The human orthologue is an SH3 adaptor protein called Spin90 and its alignment to YDL146W is shown in **Figure 18**.

Spin90 is also called DIP and its domain structure, alignment to YDL146W and other orthologues are shown in **Figure 19**. Spin90 has been shown to regulate the organisation of the actin cytoskeleton through interaction with other complexes [160, 161]. As it also contains an SH3 domain that has been well characterized in many signalling pathways, especially the Ras/MAPK pathway [162], Spin90 has been suggested to also have a role as a signalling molecule [160]. YDL146W is an orthologue of Spin90, hence it may have a similar role in yeast but this would need to be verified experimentally.

```
gi|17433253|sp|Q9NZQ3|SPN90_HUMAN Gene info SH3 adapter protein SPIN90 (NCK-
interacting protein with SH3 domain) (SH3 protein interacting with Nck, 90
kDa) (VacA-interacting protein, 54 kDa) (VIP54) (AF3p21) (Diaphanous
protein-interacting protein) (Dia-interacting protein 1) (DIP-1) Length=722

 Score =  192 bits (488),  Expect = 4e-47, Method: Composition-based stats.
 Identities = 47/302 (15%), Positives = 104/302 (34%), Gaps = 40/302 (13%)

Query  139  YTQFAYLSKYGKERTVASKHQYNSNNSSTGTSLDSLDRSL---TDIDLGIIDEMK-----  190
            +     L  Y +    AS              SLD+   S    + + + +  +M+
Sbjct  438  FESVLALVAYYQMEHRASLRLLLLKCFGAMCSLDAAIISTLVSSVLPVELARDMQTDTQD  497

Query  191  --QI--STVLMDLLFQIMKYCKCVIANLQIVDDFFVYYMMESMRS---DTMDDMFNNAEF  243
              ++  S +++ ++F +         A+ + +   F  +++  +           +   +
Sbjct  498  HQKLCYSALILAMVFSMG--EAVPYAHYEHLGTPFAQFLLNIVEDGLPLDTTEQLPD---  552

Query  244  KLLLALNEQYMMFAKEYDIENKVYKYLINGSVSRCFTELLLLKFNRASDPPLQIMM----  299
              L       +         +N +   L   +  + F+E LLL  NR  DP
Sbjct  553  --LCVNLLLALNLHLPAADQNVIMAALSKHANVKIFSEKLLLLLNRGDDPVRIFKHEPQP  610

Query  300  ----CKIIYLILTPRGDYSPMNFFYTNDLRVLIDVLIRELQNISEDEEVLRNTLLRVLIP  355
                K +  +             FY  D+  LID+ +R + ++S  ++ LR   L ++
Sbjct  611  PHSVLKFLQDVFGSPA---TAAIFYHTDMMALIDITVRHIADLSPGDK-LRMEYLSLMHA  666

Query  356  LLKNTQLSKTHYRKDDLNKLLNYLSTLDNICVDSPALHEHQVTVALSRKCLQQIPWLETP  415
            +++ T   +  +R  DL  +L  +    N    SP     ++ V    +  ++   L
Sbjct  667  IVRTTPYLQHRHRLPDLQAILRRIL---NEEETSPQCQMDRMIVR---EMCKEFLVLGEA  720

Query  416  ST  417
              +
Sbjct  721  PS  722
```

Figure 18. Alignment of the *S. cerevisiae* YDL146W against the human protein SPIN90 after the third iteration of running PSI-BLAST against the GenBank protein database.

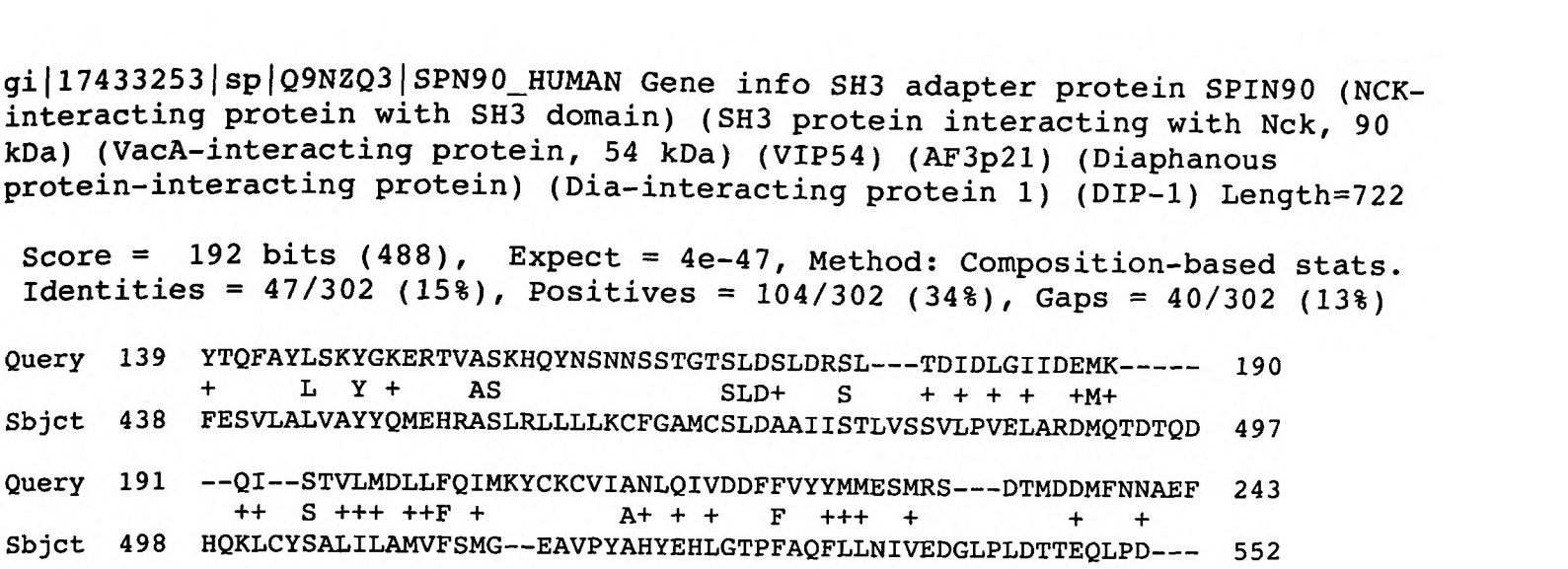


Figure 19. Schematic diagram showing the DIP leucine rich region, its domains and its alignment with orthologues in other species, for instance, the *S. cerevisiae* YDL146W. Sequences were aligned with CLUSTALW. The GenBank accession numbers for the sequences are available from [163]. Figure is also taken from [163].

YDR282C – contains a conserved domain "DUF155" which is an "Uncharacterized ACR, YagE family COG1723" as predicted by the Pfam. A cytoplasmic sporulation protein, YDL001W/RMD1 (rippling muscle disease 1), from *S. cerevisiae* also contains this domain, as does the membrane protein from *Schizosaccharomyces pombe*, Sad1-interacting factor 2, which is part of the nuclear pore complex. The domain is also found in hypothetical proteins of the chicken, fly, fish, mouse and in other fungal genomes. The YDR282C protein is predicted to be a non-secretory protein found in the cytoplasm (**Table 13**).

YGR079W – has no homologues outside of fungi. It is predicted to be both a secreted protein and also located on the outside of the cell (**Table 13**).

YLR330W – has homologues to other chitin biosynthesis proteins in the other fungal genomes. YLR330W contains three conserved domains (Fibronectin 3, BRCT and MDN1 domains) and is predicted to be non-secretory (**Table 13**).

YNL046W – only hit from PSI-BLAST to both the GenBank protein database or to the UniProt database was to itself. YNL046W is predicted to be non-secretory and contains transmembrane helices (**Table 13**).

YNL195C – No homologues found outside of fungi. It is predicted to be a non-secretory protein (**Table 13**).

Tertiary structure of the proteins of the 27 genes

Matches to Protein Data Bank (PDB) entries

I attempted to obtain either the tertiary structure of the 27 *S. cerevisiae* genes or their homologues from the PDB. To investigate the connection between positively selected and functional sites, the ω^+ sites were mapped onto 3-D protein structures of homologues to obtain their spatial position. Sites that are far apart in the primary sequence can be clustered in the 3-D structure. Previous studies of pathogens such as the human immunodeficiency virus (HIV), and *Neisseria meningitides* (commonly known as meningococcus), have shown that predicted sites of positive selection fall within functional regions and changes to these regions are biologically significant [164, 165].

However, searching through the PDB returned homologues for only six of the 27 gene products. The relative paucity of 3D structures of homologues for many genes is because they have yet to be elucidated. Yet in this instance, a more pertinent reason may be that the sequences of these 27 proteins are evolving rapidly. These sequences become so divergent that similarities between them are difficult to detect, hence known homologues to these proteins are few. Of the six matches to entries in the PDB, four of the genes had hits to various PDB homologues, yet only one gene, *YBR052C*, had a PDB alignment which encompassed the ω^+ sites predicted by the SLR program (**Figure 20**). The other genes had PDB matches to regions that did not contain ω^+ sites. Spurious results were obtained for two out of the six hits. One of the spurious hits was to the RNA sequence of a virus, and the remaining hit was to a region of low-complexity. The hit to the RNA sequence in a protein database shows

that data in public repositories are prone to errors and results obtained need to be verified manually.

```
>WrbA_Ecoli_3B6M Flavoprotein E.coli WrbA

          Length = 203


 Score =  121 bits (303), Expect = 2e-32
 Identities = 76/203 (37%), Positives = 110/203 (54%), Gaps = 18/203 (8%)

Query: 1   MPKVAILIYSVDDIIATLAENEKKGI-EIAGGEAEIFQVPDVSYKTEYATEEGK-EAAKV 58
           M KV +L YS+    I T+A     +G  ++ G E  + +VP+       +    GK + A V
Sbjct: 1   MAKVLVLYYSMYGHIETMARAVAEGASKVDGAEVVVKRVPETMPPQLFEKAGGKTQTAPV 60

Query: 59  AKTNADFSYKILTRETLVEYDYYLFGIPTKFGNFPAEWKSFWDSNTGGLWAKGSLHGKIA 118
           A           T + L +YD  +FG PT+FGN   + + ++F D  TGGLWA G+L+GK+A
Sbjct: 61  A-----------TPQELADYDAIIFGTPTRFGNMSGQMRTFLDQ-TGGLWASGALYGKLA 108

Query: 119 GLFVSGAISGKGDTEMCIMNAMSTLVHHGVIYVPLGYKNAYKELTDVEDVNGSCAWGAGC 178
            +F S    G    E  I +  +TL HHG++ VP+GY  A +EL DV  V G   +GA
Sbjct: 109 SVFSSTGTGG--GQEQTITSTWTTLAHHGMVIVPIGY--AAQELFDVSQVRGGTPYGATT 164

Query: 179 VSGIDGGRPPSLSELRVHQLQGK 201
           ++G DG R PS  EL + + QG+
Sbjct: 165 IAGGDGSRQPSQEELSIARYQGE 187
```

Figure 20. BLASTP alignment of YBR052C with its PDB homolog, 3B6M – chain A. The "query" sequence is YBR052C and the "sbjct" (subject) sequence is the A chain of the Trp repressor binding protein, WrbA [166], which has the PDB identifier "3B6M"). The 2 positive sites located at positions 47 and 48 of YBR052C are highlighted in yellow, as are the corresponding sites of 48 and 49 of the WrbA protein.

Spatial distribution of ω^+ sites

To observe the spatial orientation of the 2 positively selected sites within the YBR052C protein, the corresponding sites at position 57 and 58 have been displayed on the three-dimensional structure of its homologue, WrbA protein (**Figure 21**). The two sites are shown as spheres and labelled. They are located on the outside of the protein and within a flexible region of crystallographic disorder. Disordered sequences are proteins or local regions that do not fold into a stable 3-D structure

[167]. Such regions of disorder are not easily resolved in crystal structures [168] and it was only recently that the 3-D protein structure of this region has become available.

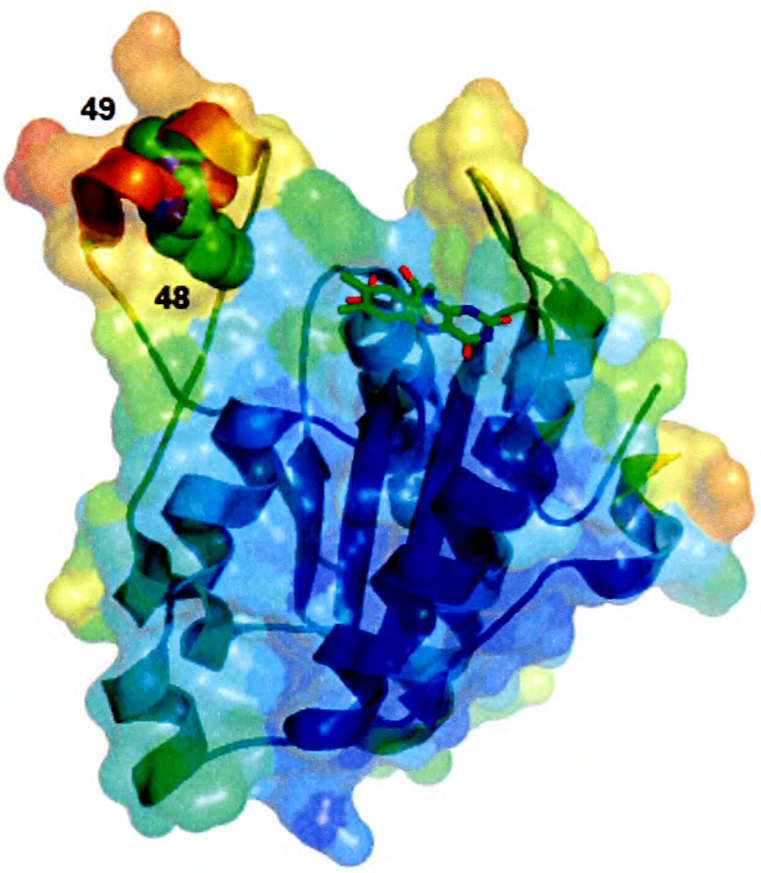


Figure 21. A cartoon representation of WrbA from *E. coli* bound to the FMN cofactor [166]. This is the PDB homolog of YBR052C. WrbA is a Trp binding receptor protein and is shown with the labelled sites 48 and 49 which correspond to the sites 47 and 48 in the protein encoded by the *YBR052C* gene. The surface of the WrbA molecule has been coloured by the temperature factor (B-factor) of the residue with cold areas shown in blue and warmer areas in orange. Here, the ω^+ sites are located in the warmest region coloured in orange, indicating that the region is very flexible and in constant motion within the crystal.

Residue solvent accessibility (RSA) of *YBR052C/RFS1*

A residue's solvent accessibility is the degree of interaction between the residue and its surrounding solvent molecules [169]. The residue solvent accessibility (RSA) values for the two ω^+ sites in the protein of the *YBR052C* gene provide further

evidence that the sites are within exposed regions. Site 48, which is situated at the hinge of the flexible region (see **Figure 21**) has high solubility (relative solubility of > 35%; RSA value of 119) and site 49 has intermediate solubility (relative solubility between 9 and 35)%; RSA value is 26.4). The range of solubility is defined by Rost and Sander [153]. The remainder of the sites, their RSA values and solubility indicators are detailed in **Appendix A**.

Discussion

The availability of multiple complete fungal genome sequences provides a powerful model system for comparative genomics analysis. Currently, 18 hemiascomycetes, eight euascomycetes and four basidiomycetes [170] genomes have been sequenced which span different fungal groups, the analyses of which provide an insight into the genome evolution in yeasts.

Enrichment of GO categories among genes showing ω^+ sites

For this evolutionary study, I have chosen for analyses the monophyletic hemiascomycetes group of *Saccharomyces sensu stricto* yeasts because they exhibit strong phenotypic similarities. My analyses have detected 27 genes which contain sites showing significant evidence of positive selection ($p_{SLR} < 0.01$ and $p_{OA} < 0.01$). These genes appear to have evolved adaptively and certain sites have gained mutations which seem to benefit the fitness of the organism. These genes are significantly enriched in cellular components such as "cell bud" and "cell wall" which are linked to the biological processes listed in **Table 12B**, for instance, "cell

budding", "pseudohyphal growth" and "cytokinesis". Below, I discuss possible ways by which these genes may have aided the organism to adapt to its environment.

The yeast cell copes in times of stress through pseudohyphal growth

During times of stress, pseudohyphal growth of the yeast cells occurs for the survival of the organism. Pseudohyphal growth is a form of polarized growth. *S. cerevisiae* cells reproduce by producing daughter cells through a process called budding or polarized cell growth (and directional cell division). **Figure 22** shows the three different phases in *S. cerevisiae* in which polarized cell growth occurs – vegetative growth, pseudohyphal growth and mating. In the wild, vegetative growth occurs in the rare event when all nutrients required for growth are available [12, 171]. The *S. cerevisiae* cells are oval or round in shape with defined bud patterns – the bud first undergoes atopic growth (growth at its tip) and then isotropic growth (growth throughout the bud). However, when nutrients are lacking (in times of stress), polarized growth occurs. The cells elongate and daughter cells bud from one end of the cell to form chains to spread widely in search of nutrients. Polarized growth also occurs in the mating of haploid cells.

Importance of the cell wall and cell bud during pseudohyphal growth

Polarized growth is an asymmetric process as the daughter cells are much smaller than their mothers and during this budding process, both the cell wall and plasma membrane are synthesized [17]. During budding, the manipulation of the yeast actin cytoskeleton, located in the cell cortex, is polarized towards the site for growth and positions the polarized cell in response to the coordinated control of different signals [9, 12]. Actin-binding proteins such as those encoded by the *SPA2* gene (from **Table**

12A and **Table 12B**) are important for cytokinesis and regulate the actin cytoskeleton during pseudohyphal growth [12]. This is essential for the correct division of chromosomes into mother and daughter cells as budding occur at the poles of the mature mother cell.

The yeast cell wall is known to adapt itself to the changing environment via its cell wall proteins located on the cell surface and cell wall organization [15]. The cell wall also contributes to the polarity of the cell by maintaining the shape of the cell which is essential for bud formation leading onto cell division and morphogenesis [12, 13].

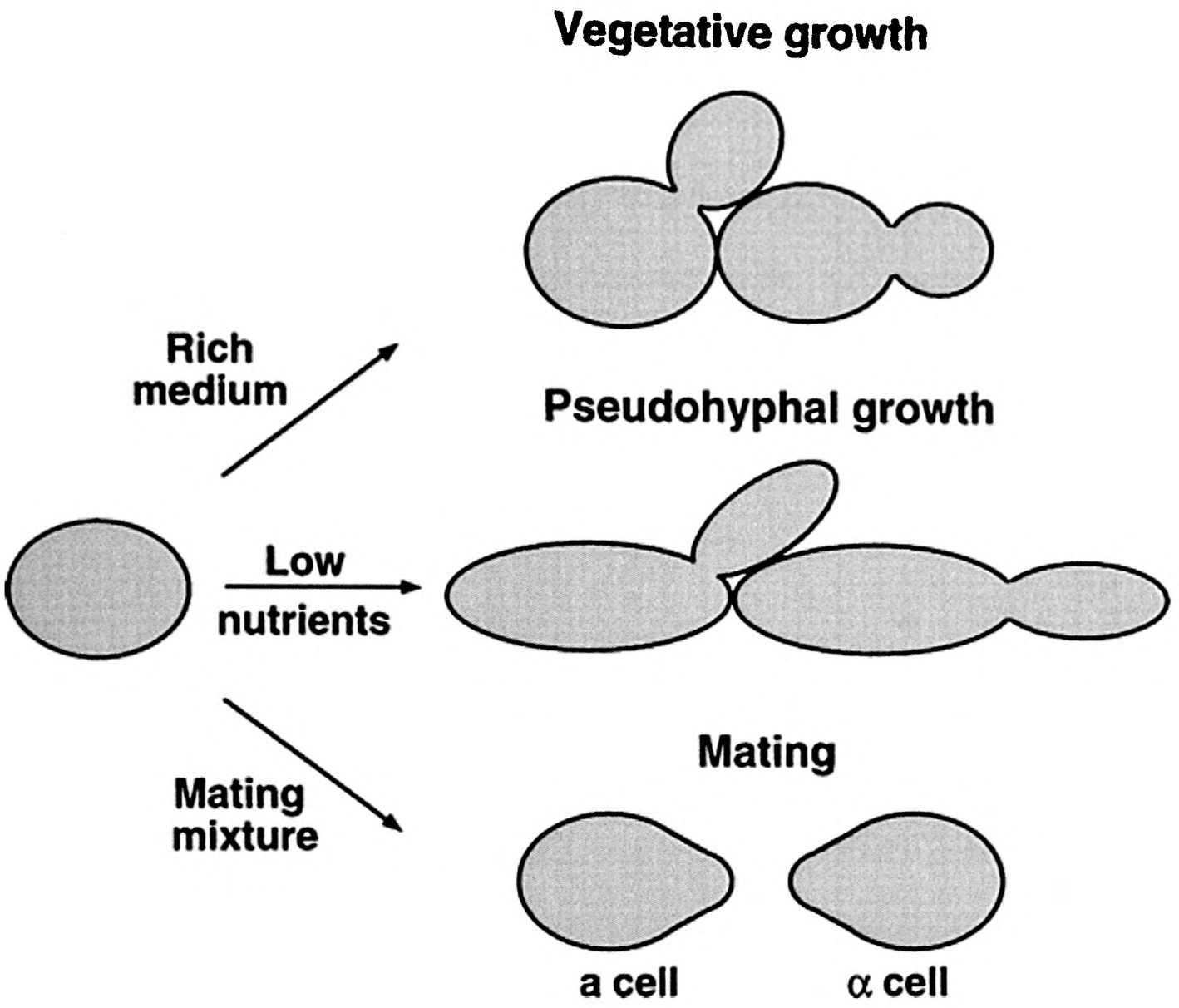


Figure 22. Polarized growth in *S. cerevisiae* and the three different phases in which they occur. In the presence of ample nutrients, *S. cerevisiae* cells are round or oval and have defined bud patterns. However, in the absence of specific nutrients, for example, nitrogen, the cells are elongated and budding into pseudohyphae occurs from the distal end. In the presence of pheromones of opposite types, haploid cells produce an extension towards their mating partners. This diagram is reproduced from Madden and Snyder [12].

Proteins of rapidly evolving genes rarely have homology with other proteins

I also found that genes whose functions are unknown are significantly enriched among the set of 27 genes ($p_{GO} < 0.05$) (**Table 12C**). I account for this observation by first noting that gene sequences that have been susceptible to adaptive evolution more frequently diverge to such an extent that their homology with genes from other, more distantly-related, species is obscured [172]. As noted elsewhere, ascomycetes-specific

genes, in general, tend to be more evolutionary divergent than other genes [173]. Moreover, rapidly-evolving genes are more rarely investigated by the research community simply because they lack demonstrable homology to genes in more distantly-related species such as metazoa and other fungi. A lower scientific interest paid to rapidly-evolving genes thus directly results in the unusually high number of genes of unknown function among those I find to have evolved by positive selection.

Functional annotation of *YBR052C/RFS1*

Of these genes with unknown function, I would like to highlight the *YBR052C/RFS1* gene. It was the only gene (out of the 27 genes) for which a homologue encompassing the ω^+ sites predicted was found within the PDB. Little is known about this *S. cerevisiae* gene: there are no GO annotations for its protein. As mentioned above, the description from the SGD states that it is a "member of a flavodoxin-like fold protein family" (**Table 11**). The PDB protein whose aligned sequence encompassed the ω^+ sites predicted in *YBR052C* was the Trp repressor protein, WrbA found in *E. coli*. Two ω^+ sites at positions 47 and 48 (indicated by the boxed area in **Figure 23**) were predicted in YBR052C. These sites were mapped to the exterior of WrbA within a short flexible region that was disordered in its crystallographic structure (**Figure 21B**). As these sites are under positive selection, it strongly suggests that they have a functional role [174]. It remains to be investigated the role of the YBR052C protein and that of its two ω^+ sites.

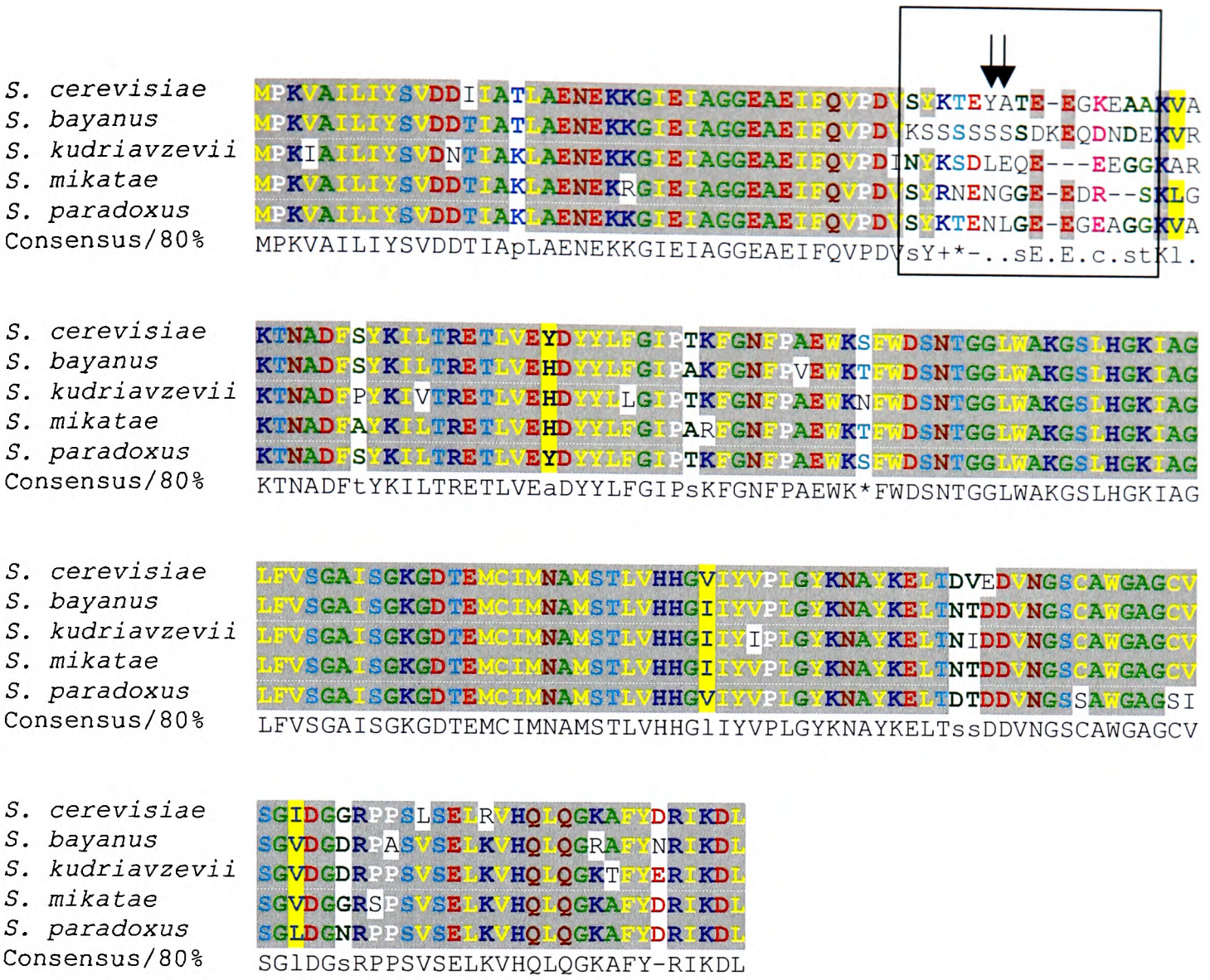


Figure 23. Amino acid alignment of the YBR052C protein in all five species. The corresponding residues of the PDB homologue, WrbA to YBR052C are sites 47 and 48 (indicated by the black arrows). The five homologues appear to have an area of high variability, indicated by the boxed region.

The 27 gene sets are not codon biased

Genes in the 27 gene sets appear to show positive selection because of adaptive evolution. However, one may argue that it is because of codon usage bias that elevates the ω rate, leading to false predictions of positive selection. More specifically, as mentioned before, ω is the ratio of K_A/K_S. If these genes in the 27 gene sets were to have been codon biased, then perhaps their ω ratio would be > 1 simply because K_S was suppressed due to codon usage bias as nucleotide substitution rate would be low (degeneracy of the amino acids). However, in a separate experiment similar to that of **Chapter 4**, the 27 gene sets were tested for codon usage bias. In **Chapter 4**, I discuss in detail the codon usage in yeast and methods used to calculate the codon usage bias

of the genes. The CAI (codon adaptation index) was used to provide an indication of the degree of codon bias (see **Methods and Materials** section in **Chapter 4**). CAI ≤ 0.25 indicates reduced or no codon bias and CAI > 0.25 indicates codon bias. 89% (24/27) of the gene sets were not affected by codon bias (CAI ≤ 0.25). Of the remaining three gene sets, two were found to be borderline cases between being affected and not being affected by codon bias (CAI ≈ 0.25).

In summary, I have shown that these 27 genes sets are substrates of adaptive evolution. As more experimental works are performed on yeast genomes and more protein structures are being made available, analyses such as the above may shed light on understanding how selection has helped the simple eukaryotic yeast to better adapt to its environment.

CHAPTER 4: Constraint on *Saccharomyces* genes

Summary

In this chapter, I show that purifying selection has acted on the genes of the genus *Saccharomyces* both at synonymous and non-synonymous sites. Highly expressed genes are known to exhibit codon bias that suppresses K_S rates. However, the majority of yeast genes are not codon-biased and hence I attempt to determine additional forces which might have constrained the synonymous site substitution rates. Furthermore, a recent study has suggested that K_A rates are under constraint to maintain the protein folding integrity of highly expressed genes. I show that constraints on non-synonymous sites are not limited to highly expressed genes but show that they are also seen among genes with low expression levels.

Introduction

Analyses on constraints on *Saccharomyces* gene sequences started first with the question of whether evolutionary forces other than translational selection on codon usage have acted at synonymous sites. Translational selection is partly responsible for the unequal usage of synonymous codons (codon usage bias) in highly expressed protein coding genes and is correlated with both the tRNA pools and genome size [175, 176]. This is mainly due to highly expressed genes using a subset of codons that correspond to their respective tRNA pools to allow improved translational efficiency.

Selection at both synonymous and non-synonymous sites reduces the number of misfolded proteins that are toxic to the cell [143]. This may confer greater fitness for survival and encourage translational efficiency. Additional constraints on non-synonymous sites also contribute to the evolution of the protein. For instance, highly expressed housekeeping genes which are crucial for a cell's survival encode proteins that are usually under strong constraint. These proteins appear to evolve slowly not only due to selection for translational accuracy but also due to increased constraints at non-synonymous sites to ensure that proteins correctly fold.

Codon usage bias of genes

Numerous analyses have shown that there is a preferential use of certain codons to encode amino acids and this preference varies widely between and within species [177-182]. This biased use of codons for the same amino acid reflects a balance between mutational biases, as seen in the distinctive G+C nucleotide composition within neutral sequences of the genome, and selection seeking to improve translational efficiency [178, 179, 182-186].

In fast-growing organisms like *Saccharomyces cerevisiae* and *Escherichia coli*, *it* appears that natural selection for efficient translation is largely responsible for biased codon usage [187, 188]. Optimal codons may help to achieve faster translation of an mRNA into its protein and greater accuracy. The composition of the genomic tRNA pool may thus be a reflection of these optimal codons. As a result, highly expressed genes are subject to stronger translational selective pressure than lowly expressed ones [185, 189, 190]. In contrast, codon preferences in other organisms, for instance

Homo sapiens (humans), are not mainly due to selection [191], but appear to be a consequence of the mutational biases, such as G+C content or isochore composition seen in the genome [175]. An intermediate level of codon usage bias is seen among organisms such as the *Drosophila* species, (including *Drosophila melanogaster* (fruit fly)) [192] and *Bacillus subtilis* [193] where both translational selection and mutational bias play a part.

Codon usage bias due to translational selection may reflect the evolutionary forces which act on highly expressed genes to adapt their codon use to the tRNA pool in the host genome [175, 176]. Therefore, it may be predicted, in general, that genes with a high codon bias are highly expressed and have a low K_S value as a result of constraints on synonymous rate substitutions such as those at the third codon position. Therefore, in yeast, a high codon bias may indicate a strong selective pressure and hence a slower K_S rate of synonymous substitutions [187, 188, 190, 194].

Within *S. cerevisiae*, factors currently known to correlate with the degree of codon usage bias include gene expression level, gene length, relative recombination rate, intergenic spacing, G+C content in the third position and the rate of synonymous substitutions [179] [195]. Gene expression levels for cells growing fermentatively have a positive and significant correlation [195].A phenomenon termed transcription-associated mutation (TAM) is seen in yeast where highly expressed genes suffer higher mutation rates, thereby creating a positive correlation between codon usage bias and the rate of synonymous substitutions [11, 188, 196]. Intergenic spacing also negatively correlates with codon usage bias within yeast although this relationship is

not linear. The use of preferred codons increases as the gene spacing decreases to a point where the preferred codon usage decreases as the spacing continues to decrease [179]. Kilman *et al.* [179] suggested that the correlations of recombination rate, gene length and intergenic spacing are due to the Hill-Robertson effect where selection is less effective among linked targets. However, the existence of the Hill-Robertson effect in yeast has been examined and questioned (see [197] for a discussion).

Many methods to measure the non-uniform use of synonymous codons (codon bias) have been proposed and evaluated (reviewed by Comeron [198]). The Codon Adaptation Index (CAI) and Nc (number of effective codons used in a gene) [186] were the two most robust methods [198] as they were independent of the lengths of coding sequences. Both Nc and CAI both showed significant correlations between species' codon usage biases, with a stronger correlation detected when CAI was used as a measure between closely related species [192]. Thus for this project, CAI was used as a measure of codon bias and other methods will not be further discussed. (Please refer to the following **Methods and Materials** section for a detailed discussion on how CAI values for each species were calculated). In summary, I found the codon usage bias of *S. cerevisiae*, *S. bayanus*, *S. mikatae* and *S. paradoxus* to be highly similar. Hence, as *S. cerevisiae* is a very well studied genome and the genes are well annotated, I used the genes from this species to represent the other three when performing additional analyses (for example, enrichment of functional categories amongst gene sets) (see below).

Methods and Materials

Gene selection and group distinction

For this study, I initially investigated orthologues present in seven species of the *Saccharomyces* genus that coded for proteins at least 100 amino acid in length. This length criterion was the same as that used by Kellis *et al.* (WashU dataset) [20]. The seven species – *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. kluyveri* and *S. kudriavzevii* yielded only 1,014 orthologous full length genes compared to the possible 6,703 annotated full length genes (see **Chapter 3**). In order to increase the gene count, I made use of orthologues from only four of the *sensu stricto* species – *S. cerevisiae*, *S paradoxus*, *S. mikatae* and *S. bayanus* which brought the gene count up to 4,017 orthologous and full length genes. Subdivision of the genes into four groups A – D was based on CAI values (above 0.25: groups A and D; below 0.25: groups B and C) and either ranked K_S or ranked K_A values. Groups A and B were distinguished by ranked K_S, or ranked K_A values < 0.1, while groups C and D were distinguished by K_S, or ranked K_A values > 0.1 (**Figure 24**).

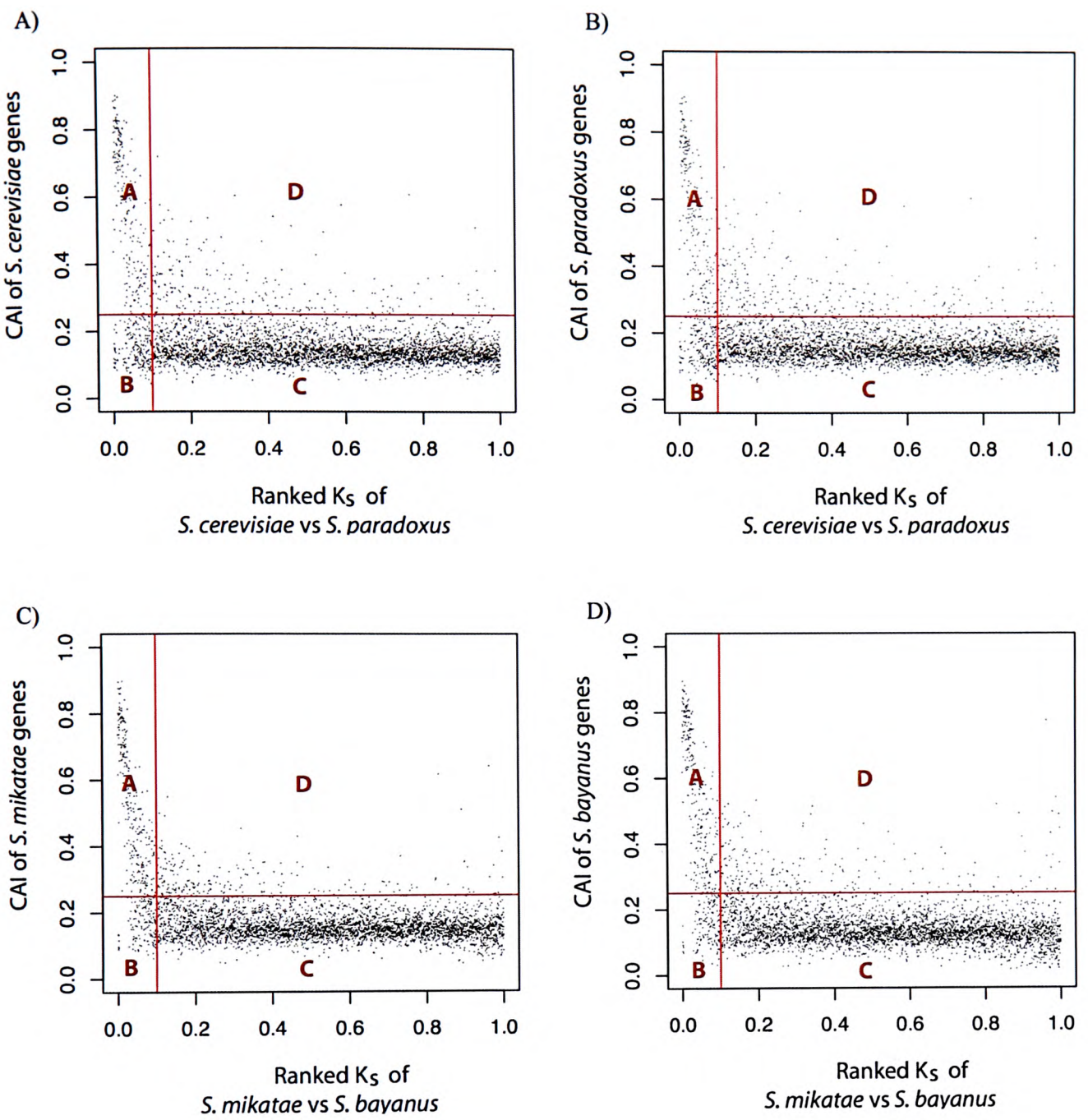


Figure 24. Ranked K_S values of each plotted against the CAI values for each of the 4 *stricto sensu* species – *S. cerevisiae*, *S. mikatae*, *S. paradoxus* and *S. bayanus*.

Calculating mRNA abundance and translational activity

Data for mRNA and protein abundances for each *S. cerevisiae* ORF were obtained from the published works of Beyer *et al.* [199] who studied the post-transcriptional expression regulation of *S. cerevisiae* on a genomic scale. Beyer *et al.* also provided the data from which I calculated translational activity. Translational activity, defined as the product of the mRNA abundance and translation rate, provides an indication of the number of proteins synthesized per unit time. The translation rate was calculated

as the product of the ribosome density on each mRNA and the ribosome occupancy, which in turn indicated the fraction of mRNA bound to the ribosomes. These data provided an understanding of how selection on the cell controls protein levels, both transcriptionally and post-transcriptionally.

Estimating the codon usage bias of the genes

CAI was used to estimate the degree of synonymous codon usage bias. CAI measures the deviation from a reference set of preferred codons, and is created from highly expressed genes in the organism [194]. This is because codon usage bias is found mostly in highly expressed genes as a consequence of translational selection (as discussed above). CAI is derived from the normalized codon preference statistic. This statistic provides the likelihood of finding a codon in a highly expressed gene as compared to the likelihood of finding the same codon in a random sequence with the same amino acid composition. The normalization of each amino acid removes variation in the amino acid composition between different genes and allowed comparison both within and between species [194]. Thus, a sequence with a high CAI value, close to 1.0, indicates that there is a higher probability of finding the same codon usage in highly expressed genes of the same organism compared to finding it in a random sequence. A low CAI value, much less than 1.0 (but more than 0.0), indicates the use of less preferred codons.

An improved method to derive the dominating codon bias was devised recently [177]. This had the main advantage that the user did not need to seed the process by first defining a reference set of highly expressed genes. Instead, the method by Carbone *et al.* considered all genes of an organism and selected a reference set which was

representative of the family of codons that appeared with the highest frequency in most genes (and within the genome). This reference set would score highly on the CAI scale (values close to 1). In this instance, disregarding whether the bias was due to translational selection or not, CAI values were used as a measure to detect the most dominant codon bias in the genome. The added advantage of this method was that an index of the codon usage bias for newly sequenced genomes (for example, *S. paradoxus*) may be easily obtained, even when orthologues of highly expressed genes were not easily identifiable. Hence, I used the method of Carbone *et al.* to derive the codon usage bias of all genes for each of the four species independently – *S. cerevisiae*, *S. bayanus*, *S. mikatae* and *S. paradoxus*.

Results and Discussion

Analysing the constraints on the genes of the *Saccharomyces* genus revealed purifying selection acting at both synonymous and non-synonymous sites. Below, I detail the experiments undertaken and highlight results found at each type of sites (synonymous/non-synonymous).

Constraints at synonymous sites

I was interested to see if other forces, apart from translational selection, resulted in the low synonymous substitution rates in yeast protein coding genes. Hence, the K_S values calculated from two independent pairs of closely-related species, *Saccharomyces mikatae/S. paradoxus* and *S. cerevisiae/S. bayanus* were ranked (normalized) between 0 and 1 for comparison and analyses. These two sets of ranked

K_S values were tested for any significant correlations to test if synonymous substitution rates in one pair of species were predictive of those in the other pair. The two substitution rates were indeed found to be highly correlated with each other. In addition, there was a set of genes which had low (ranked) K_S in both sets of species (**Figure 25**). These genes appeared to show suppressed substitution rates in both pairs of species indicating that they were highly constrained at their synonymous sites.

Majority of *S. cerevisiae* genes are not codon biased

I next tested whether genes with suppressed substitution rates are subject to increased codon usage bias. Firstly, for each species, as shown in **Figure 24**, I defined four groups of genes –A, B, C and D to distinguish between genes with high/low codon bias and high/low ranked K_S values. After which I obtained orthologues present in all species for each of the four groups. Groups B and C were constrained at their synonymous sites and had low codon usage biases (low CAI values, less than 0.25), whereas Group A and D of genes were under considerably less constraint at the same sites and had high codon usage biases (high CAI values, above 0.25). The CAI value of 0.25 was chosen experimentally to distinguish genes with low and high codon usage bias. For instance, in *S. cerevisiae*, 16.3% of the genes had high codon usage biases (CAI > 0.25) and 83.7% were genes with low codon usage biases (CAI ≤ 0.25). Thus the majority of genes in *S. cerevisiae* were found to be not codon usage biased.

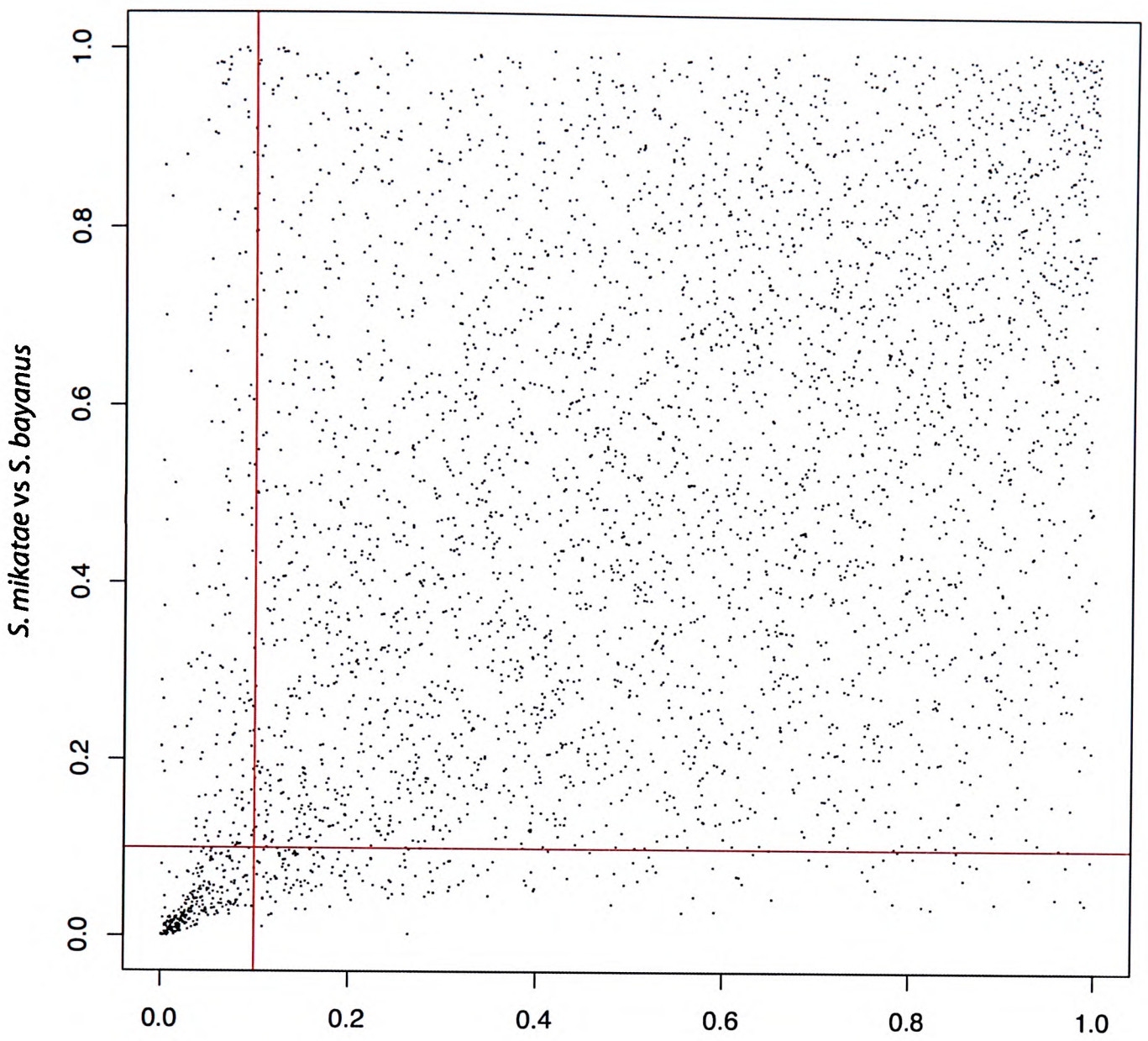


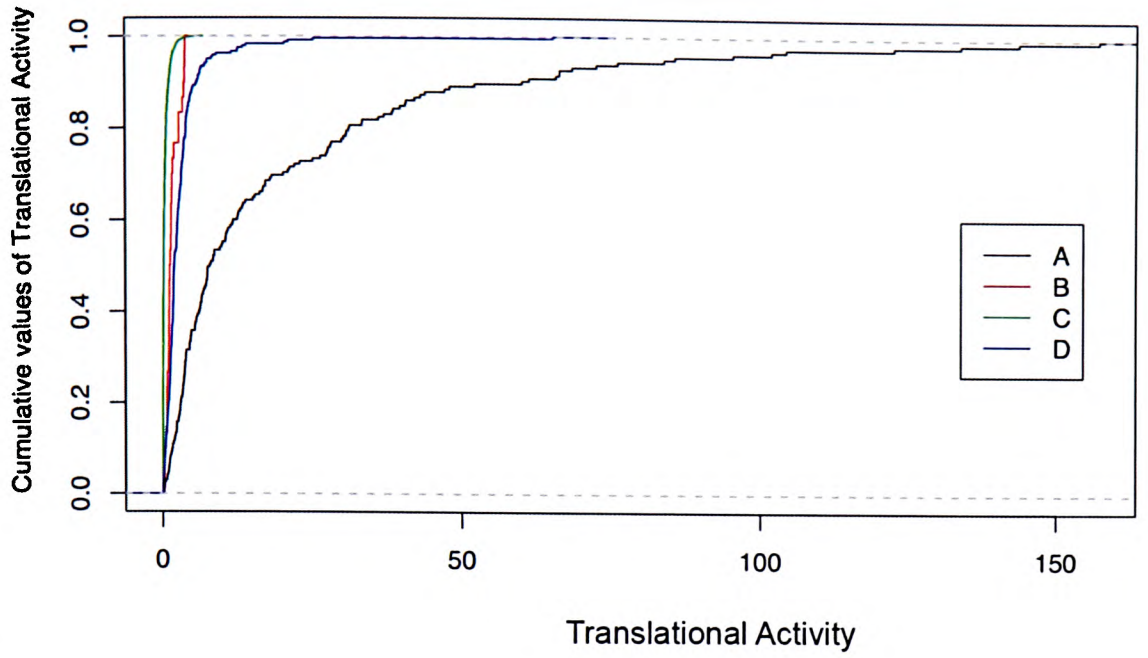
Figure 25. Ranked K_S values calculated between the orthologues from *stricto sensu* species – *S. mikatae vs S. bayanus* (y-axis) and *S. cerevisiae vs S. paradoxus* (x-axis). This scatter plot shows a high concentration of genes with low K_S values at the bottom 10%. Ranked K_S values of all the genes between the two pairs of species are highly correlated ($r^2 = 0.44$, *p*-value $< 2.2 \times 10^{-16}$).

Group A *S. cerevisiae* genes, which were highly constrained at synonymous sites due to translational selection (observed as codon usage bias) were, as expected, highly expressed (**Figure 26**). Group B *S. cerevisiae* genes, however, although highly constrained at synonymous sites, exhibited low expression levels (**Figure 26**). These genes in group B were few in number (37 genes). Group B genes were those with

ranked $K_S < 0.1$ for both pairs of species (*i.e. S. cerevisiae vs S. paradoxus* and *S. mikatae vs S. bayanus*) and with CAI values ≤ 0.25. Nevertheless, I investigated whether their low K_S values might reflect a biological process that was independent of translational selection. For this, I used GO-slim terms which represent a broad overview of Gene Ontology (GO) ontologies [128] without fine-grained and detailed terms. I asked whether these 37 Group B *S. cerevisiae* genes were significantly enriched in one, or more, of these GO-slim terms. However, no such enrichments were observed. Perhaps if the number of group B genes were larger, repeating the above GO-slim bias analysis might yield a stronger signal which may provide clues to other potential selective mechanisms.

However, genomes of these four *Saccharomyces* species have already been sequenced to a high coverage to obtain the maximum number of genes, and it is unlikely that any further significant addition to the genome gene counts will be established.

A)



B)

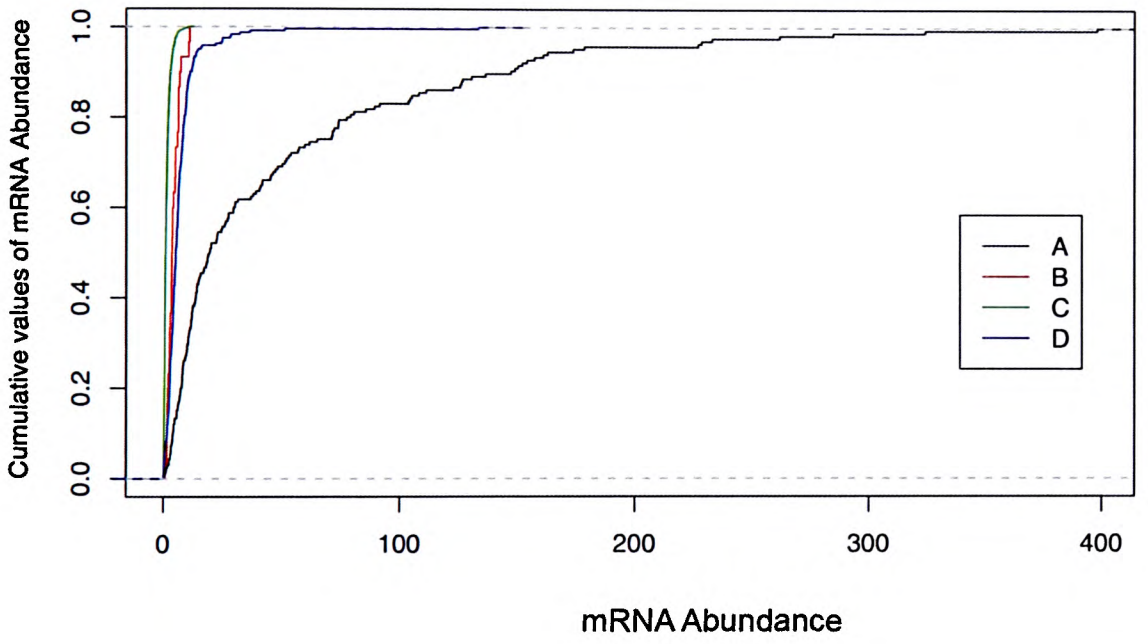


Figure 26. A) Translational activity and B) mRNA abundance of the four groups (A, B, C and D) of *S. cerevisiae* genes. Genes were examined for constraints at synonymous sites. Group A genes which show high codon usage bias show, as expected, the highest expression levels.

Constraints at non-synonymous sites

Following on from the above experiment, I wanted to test if genes which were constrained at their synonymous sites were also constrained at their non-synonymous sites. This was because selection could constrain both protein sequence (K_A) and

synonymous site evolution (K_S) to reduce the number of error-induced proteins, resulting in greater fitness for survival and translational efficiency [143]. As discussed above, translational accuracy increases the pressures for preferred codons to reduce the mistranslation of proteins [176], thus constraints would be observed at synonymous sites. Likewise, pressure for protein sequences to fold properly despite mistranslation would constrain non-synonymous site substitutions.

A similar study was published at the time of this work by Drummond *et al.* [200]. They used *S. kudriavzevii* and *Kluyveromyces waltii*, in addition to the 4 *Saccharomyces* species I used, to study the link between proteins and their expression levels. They found that in both duplicated and non-duplicated genes, expression level strongly influenced protein evolutionary rate through the number of translation events; constraining the protein sequence directly rather than through translational efficiency selection. Structural or functional differences between proteins (for instance, between paralogues in a pair) with differing expressions levels may bias this K_A-expression relationship. In addition, the higher expressed protein of a paralogue pair was found to evolve disproportionately slower.

From their studies, Drummond *et al.* suggested the reason why highly expressed proteins evolved slowly was because of constraint on both synonymous *and* non-synonymous sites. They coined the term "translational robustness" whereby proteins would fold properly despite translation missense errors. They reasoned that if translational selection was the reason for constrained synonymous site evolution, it would result in the protein sequence being constrained as a side effect of selection on the mRNA sequence. Drummond *et al.* also argue the case for translational robustness

which reflects the direct cost of misfolded proteins and is independent of protein function. Thus, constraints on both K_A and K_S reflect two independent modes of selection which are not explainable by a translational preference for either codons or amino acids.

Other possible constraints at non-synonymous sites

Next I tested if, in addition to the work of Drummond *et al.*, there were genes that were *lowly* expressed which had constrained protein evolution. Such genes would be highly constrained in their amino acid sequences and yet their synonymous sites would be tolerant of substitutions that do not lead to amino acid changes. Indeed, I find proteasome and DNA repair genes to be significantly over-represented in this set. Below, I discuss the results that led to these findings and also discuss one such gene as an example.

Highly expressed genes are constrained at non-synonymous sites

By correlating the ranked K_A values for two pairs of species (*S. mikatae*/*S. paradoxus* and *S. cerevisiae*/*S. bayanus*), I observed that constraint on the amino acid sequence was preserved between the two pairs of species (**Figure 27**). In addition, I find two areas where genes were concentrated in both pairs of species: the top and bottom 10% of the ranked K_A values (boxed in green and red, respectively in **Figure 27**). I further analyzed the bottom 10% of the ranked K_A values as I was interested in genes from both pairs of species that were highly constrained at non-synonymous sites. Because highly expressed genes constrained at their synonymous sites tend to have high codon

usage bias [187, 188, 190, 194], it can be assumed that genes with low codon usage bias would have low expression levels.

Thus, I categorized the genes from each of the four species into four groups again – A, B, C and D using ranked K_A and CAI values as parameters (**Figure 28**). From the high CAI values (> 0.25), Groups A and D appeared to be constrained at their synonymous sites. Groups B and C were less constrained at the same sites (low CAI values, ≤ 0.25). Groups A and B were genes which were found in the bottom 10% of the ranked K_A, as shown in **Figure 27**). Using the hypergeometric distribution test, Group A *S. cerevisiae* genes were significantly over-represented in their Gene Ontology (GO) [128] annotations for ribosomal genes (**Table 14**). Ribosomal proteins are thus greatly constrained at their non-synonymous sites, as well as at their synonymous sites. These genes are also well known to be highly expressed which accounts for their high codon usage biases [185].

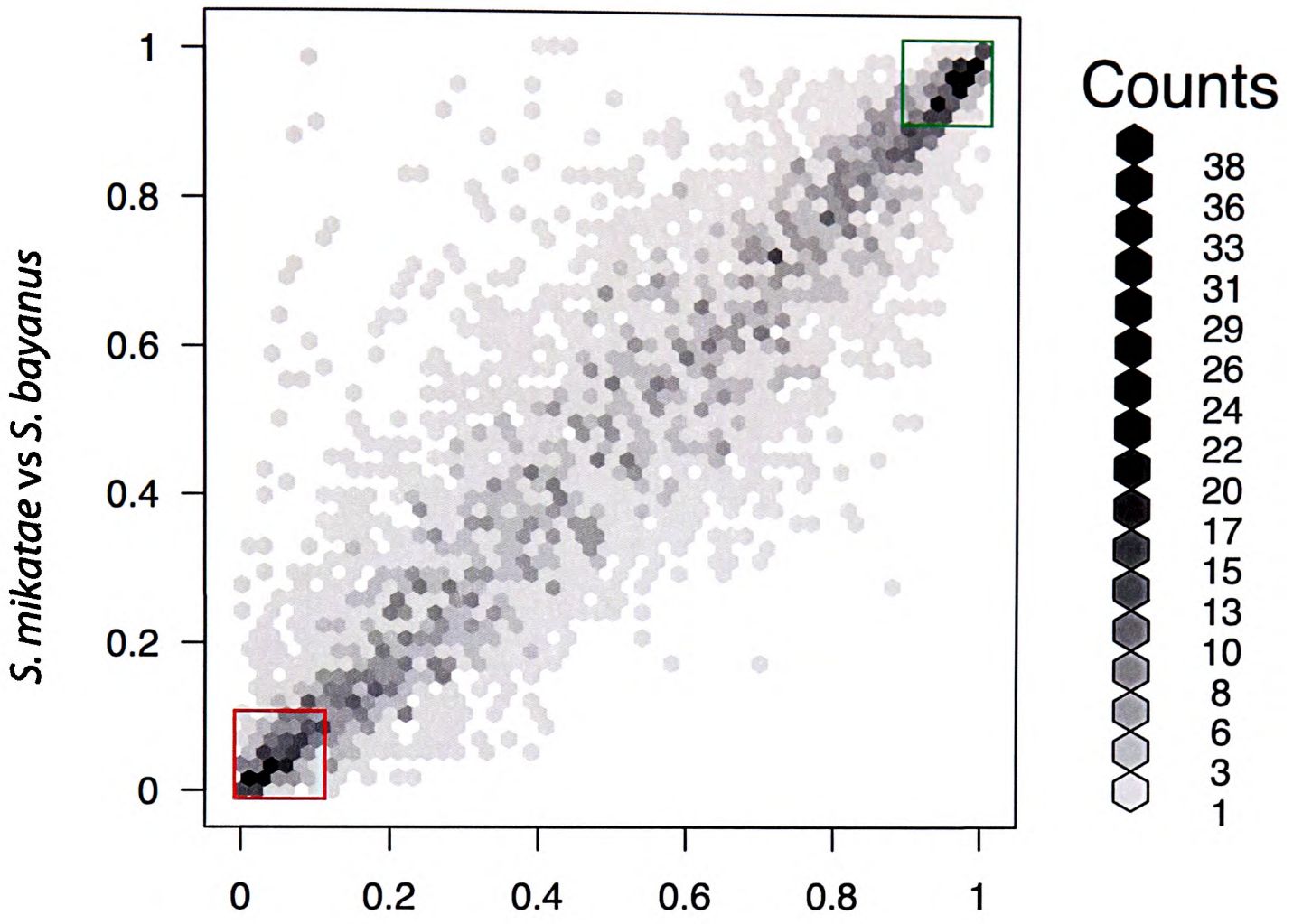


Figure 27. Ranked K_A values between orthologues from *stricto sensu* species – *S. mikatae vs S. bayanus* (y-axis) and from *S. cerevisiae vs S. paradoxus* (x-axis). This density plot shows a concentration of orthologues at both the top and bottom 10% (boxed in green and red and showing low and high constraint, respectively, on amino acid sequences). Ranked K_A values of the genes between the two pairs of species are highly correlated ($r^2 = 0.82$, *p*-value $< 2.2 \times 10^{-16}$).

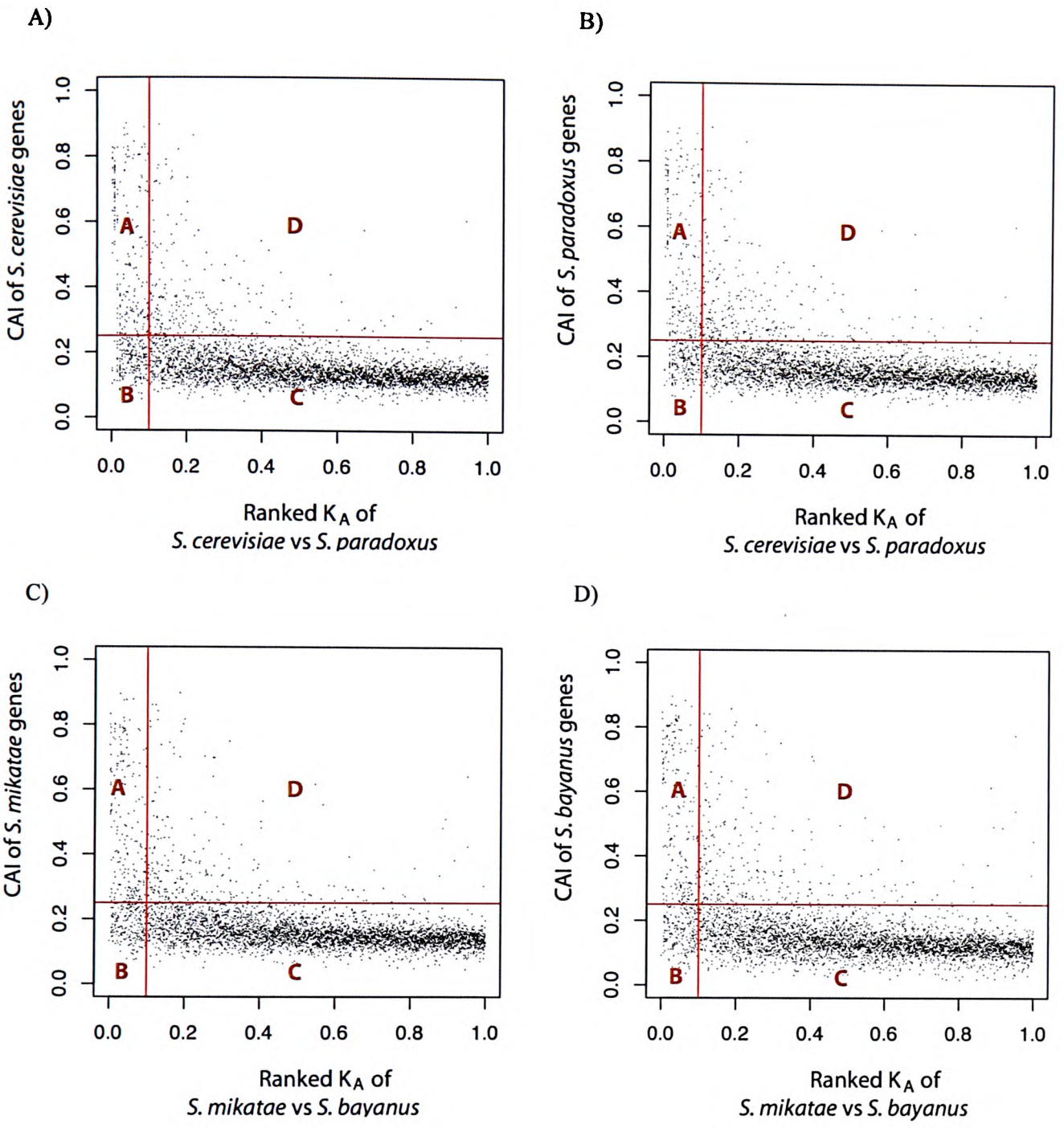


Figure 28. Ranked K_A values plotted against CAI values for four *stricto sensu* species – *S. cerevisiae*, *S. mikatae*, *S. paradoxus* and *S. bayanus*.

A) Cellular Component $p < 0.0005$		
GOid	**GO term**	**Definition**
GO:0005842	cytosolic large ribosomal subunit (sensu Eukaryota)	The large subunit of a eukaryotic cytosolic ribosome; has a sedimentation coefficient of 60S. As in, but not restricted to, the eukaryotes (Eukaryota, ncbi_taxonomy_id:2759).
GO:0005843	cytosolic small ribosomal subunit (sensu Eukaryota)	The small subunit of a eukaryotic cytosolic ribosome; has a sedimentation coefficient of 40S. As in, but not restricted to, the eukaryotes (Eukaryota, ncbi_taxonomy_id:2759).
GO:0005840	Ribosome	An intracellular organelle, about 200 A in diameter, consisting of RNA and protein. It is the site of protein biosynthesis resulting from translation of messenger RNA (mRNA). It consists of two subunits, one large and one small, each containing only protein and RNA. Both the ribosome and its subunits are characterized by their sedimentation coefficients, expressed in Svedberg units (symbol: S). Hence, the prokaryotic ribosome (70S) comprises a large (50S) subunit and a small (30S) subunit, while the eukaryotic ribosome (80S) comprises a large (60S) subunit and a small (40S) subunit. Two sites on the ribosomal large subunit are involved in translation, namely the aminoacyl site (A site) and peptidyl site (P site). Ribosomes from prokaryotes, eukaryotes, mitochondria, and chloroplasts have characteristically distinct ribosomal proteins.
GO:0016281	eukaryotic translation initiation factor 4F complex	The eukaryotic translation initiation factor 4F complex is composed of eIF4E, eIF4A and eIF4G; it is involved in the recognition of the mRNA cap, ATP-dependent unwinding of the 5'-terminal secondary structure and recruitment of the mRNA to the ribosome.
GO:0000788	nuclear nucleosome	A complex comprised of DNA wound around a multisubunit core and associated proteins, which forms the primary packing unit of DNA in the nucleus into higher order structures.
GO:0005754	proton-transporting ATP synthase, catalytic core (sensu Eukaryota)	The hexamer that possesses the catalytic activity of the mitochondrial hydrogen-transporting ATP synthase. As in, but not restricted to, the eukaryotes (Eukaryota, ncbi_taxonomy_id:2759).
GO:0005850	eukaryotic translation initiation factor 2 complex	Complex of three heterogeneous polypeptide chains, that form a ternary complex with initiator methionyl-tRNA and GTP. This ternary complex binds to free 40S subunit, which subsequently binds the 5' end of mRNA.

B) Molecular function p < 0.0005		
GOid	**Go term**	**Definition**
GO:0003735	structural constituent of ribosome	The action of a molecule that contributes to the structural integrity of the ribosome.
GO:0003743	translation initiation factor activity	Functions in the initiation of ribosome-mediated translation of mRNA into a polypeptide.
GO:0004396	hexokinase activity	Catalysis of the reaction: ATP + D-hexose = ADP + D-hexose 6-phosphate.
GO:0051082	unfolded protein binding	Interacting selectively with an unfolded protein.

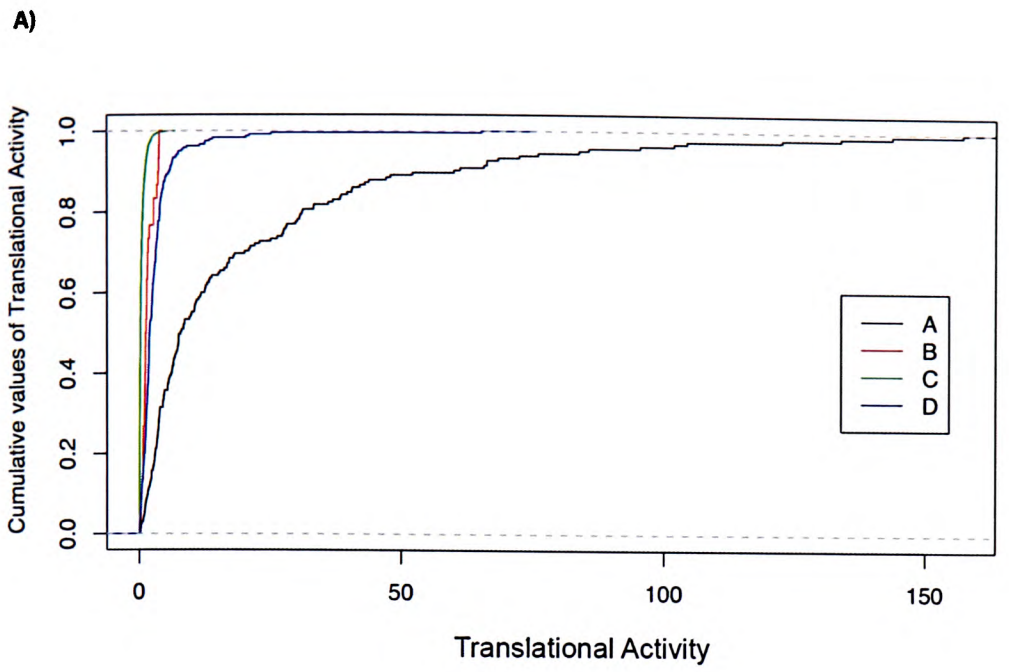
C) Biological process p < 0.0005		
GOid	**name**	**Definition**
GO:0043037	translation	A ribosome-mediated process in which the information in messenger RNA (mRNA) is used to specify the sequence of amino acids in a polypeptide chain.
GO:0006096	glycolysis	The breakdown of a monosaccharide (generally glucose) into simpler components, including pyruvate.
GO:0006414	translational elongation	The successive addition of amino acid residues to a nascent polypeptide chain during protein biosynthesis.
GO:0006413	translational initiation	The process preceding formation of the peptide bond between the first two amino acids of a protein. This includes the formation of a complex of the ribosome, mRNA, and an initiation complex that contains the first aminoacyl-tRNA.
GO:0006696	ergosterol biosynthesis	The formation from simpler components of ergosterol, (22E)-ergosta-5,7,22-trien-3-beta-ol, a sterol found in ergot, yeast and moulds.
GO:0006094	gluconeogenesis	The formation of glucose from noncarbohydrate precursors, such as pyruvate, amino acids and glycerol.
GO:0006333	chromatin assembly or disassembly	The formation or destruction of chromatin structures.
GO:0006098	pentose-phosphate shunt	The process by which glucose is oxidized, coupled to NADPH synthesis. Glucose 6-P is oxidized with the formation of carbon dioxide (CO2), ribulose 5-phosphate and reduced NADP; ribulose 5-P then enters a series of reactions interconverting sugar phosphates. The pentose phosphate pathway is a major source of reducing equivalents for biosynthesis reactions and is also important

		for the conversion of hexoses to pentoses.
GO:0006450	regulation of translational fidelity	Any process that modulates the frequency, rate or extent of activities to ensure translational fidelity.
GO:0051083	cotranslational protein folding	The process of assisting in the correct noncovalent assembly of the ribosome-bound nascent chains of a multidomain protein whilst other parts of the protein are still being translated.

Table 14. GO analyses of the A group of genes in *S. cerevisiae*, obtained from the ranked K_A between *S. cerevisiae* and *S. paradoxus vs* CAI values of *S. cerevisiae* genes. The three main ontology categories in GO: A) cellular component, B) molecular function and C) biological process; the GO categories within each of the three ontologies which were shown using the hypergeometric distribution to be over-represented in the group A set of genes, Bonferroni corrected *p*-values < 0.0005.

Some lowly expressed genes show constraints at non-synonymous sites

Not all genes constrained at their non-synonymous sites had high expression levels. This was seen in group B genes which, although highly constrained at non-synonymous sites, had low CAI values. This meant that they were less affected by codon usage bias and may imply that they were not highly expressed (**Figure 29**). Using the hypergeometric distribution function test and the conservative Bonferroni multiple testing correction, group B *S. cerevisiae* genes were found to be significantly over-represented in transcription and protein catabolism functions when a high-level overview of GO was performed using only GO-slim categories (p <0.0005). When the more detailed GO analysis was carried out, proteasome and DNA repair genes were found to be over abundant (**Table 15**).



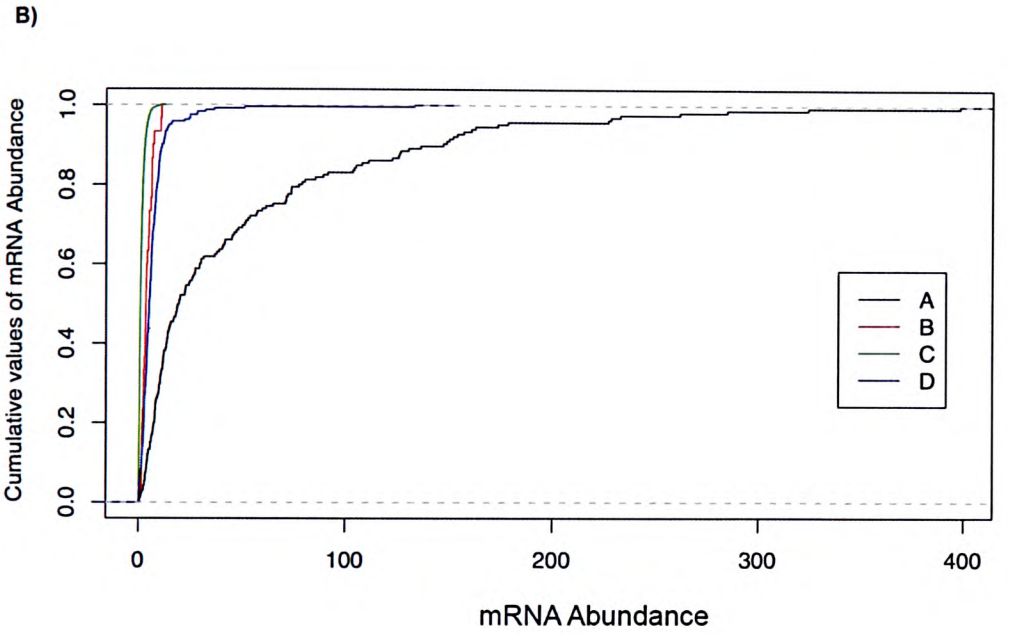


Figure 29. A) Translational activity and B) mRNA abundance of the four groups (A, B, C and D) of *S. cerevisiae* genes. Group B genes are not highly expressed and yet show constraints at non-synonymous sites.

A) Biological process

GO-ID	GO term	*r*	*k*	*p*	*N*	*p*-value	Genes
6511	ubiquitin-dependent protein catabolism	10	98	50	3219	7.07×10^{-6}	*YGL011C, YML092C, YMR314W, YOL038W*
717	nucleotide-excision repair, DNA duplex unwinding	3	98	1	3219	9.80×10^{-5}	*YDR460W, YER171W, YPL122C*
7070	negative regulation of transcription from RNA polymerase II promoter, mitotic	3	98	2	3219	2.40×10^{-4}	*YDR460W, YER171W, YPL122C*

B) Molecular function

GO-ID	GO term	*r*	*k*	*p*	*N*	*p*-value	Genes
3924	GTPase activity	11	98	30	3219	1.29×10^{-8}	*YHR107C, YFL038C, YER031C, YFL005W, YPR088C, YML121W, YBR164C, YNL090W, YGR163W, YLR229C, YML001W*
4175	endopeptidase activity	9	98	16	3219	1.64×10^{-8}	*YGL011C, YOR117W, YGL048C, YML092C, YMR314W, YOL038W, YDL007W, YFR004W, YBL041W*
3899	DNA-directed RNA polymerase activity	7	98	13	3219	8.97×10^{-7}	*YDR404C, YOR116C, YDR045C, YOL005C, YJL140W, YGL070C, YJR063W*
16251	general RNA polymerase II transcription factor activity	6	98	23	3219	1.56×10^{-4}	*YDR460W, YPR086W, YKL058W, YER171W, YPL122C, YER148W*

C) Cellular component

GO-ID	GO term	*r*	*k*	*p*	*N*	*p*-value	Genes
19773	proteasome core complex, alpha-subunit complex (sensu Eukaryota)	4	98	1	3219	3.51×10^{-6}	*YGL011C, YML092C, YMR314W, YOL038W*
5665	DNA-directed RNA polymerase II, core complex	4	98	1	3219	3.51×10^{-5}	*YDR404C, YOL005C, YJL140W, YGL070C*
5832	chaperonin-containing T-complex	3	98	1	3219	9.80×10^{-5}	*YIL142W, YDL143W, YJL008C*
112	nucleotide excision repair factor 3 complex	3	98	2	3219	2.40×10^{-5}	*YDR460W, YER171W, YPL122C*
8540	proteasome regulatory particle, base subcomplex (sensu Eukaryota)	3	98	3	3219	4.69×10^{-4}	*YOR117W, YGL048C, YDL007W*
5675	transcription factor TFIIH complex	3	98	3	3219	4.69×10^{-4}	*YDR460W, YER171W, YPL122C*

Table 15. GO analyses of the B group of genes in *S. cerevisiae*, obtained from the ranked K_A between *S. cerevisiae* and *S. paradoxus vs* CAI values of *S. cerevisiae* genes. The three main ontology categories in GO: A) biological process, B) molecular function and C) cellular component. Using the hypergeometric distribution and Bonferroni multiple testing correction, the GO categories listed in the tables are significantly over-represented in the group B set of genes (*p*-value < 0.0005). These genes have low expression levels and are highly constrained at their non-synonymous sites.

Importance of tertiary structure for protein function

A reason why these lowly expressed genes show constraints at non-synonymous sites would be to maintain the tertiary structure of their proteins. After a gene is expressed and translated into a protein sequence, the flexible, irregular polymer chain of the protein folds into a compact and specific structure that is required for its biological function. The thermodynamics of the underlying molecular interactions are important in stabilizing this folded conformation [201]. Thus, constraints on non-synonymous sites would guarantee the integrity of the protein sequence at the genic level, which in turn enables the formation of the required protein structure.

An example is the YDR460W/TFB3 protein (from **Table 15**) that is involved in both DNA repair and transcription [202]. This protein is part of the nine-subunit TFIIH transcription factor and is necessary for cell cycle progression. YDR460W contains a RING (Really Interesting New Gene) domain and mutations to this domain show it is important for the stability of the kinase moiety of TF11H and its core [202]. The RING domain has been also been suggested to be important in the architecture of large complexes and protein ubiquitination complexes [203].

In summary, although expression levels may be a factor determining the constraints on its synonymous sites to improve a cell's translational efficiency, yet for a certain subset of yeast proteins, the importance of the resulting protein and its tertiary structure is reflected in constraints on its non-synonymous sites, regardless of expression levels.

CHAPTER 5: Schistosome gene evolution

Summary

In this chapter, two evolutionarily related species of schistosomes whose genomes have a sequence identity of about 84% [46] were analysed: *S. japonicum* and *S. mansoni*. After obtaining 7,458 orthologues between the two pathogens, I partitioned the genes by stage-specificity and identified 2,112 stage-specific genes. These stage-specific genes exhibited significantly elevated dN/dS values compared to genes expressed in multiple stages of the pathogen life cycle. From these stage-specific genes, the top 10% fastest evolving gene sequences were analysed in detail. I identified genes coding for nuclear receptors, a potential coactivator of nuclear receptors and transcription factors which bound to the hormone receptor element of target genes in schistosomes. In addition, I discuss how these fast evolving genes may provide adaptations for growth, development and defence for these pathogens in their immunologically hostile host environment.

Introduction

After malaria, schistosomiasis is the most prevalent tropical disease in the world (source: National health service, UK [204]). Chemotherapy remains the most important course of intervention alongside anthelminthic drugs. However, reinfection decreases the efficacy of such chemical treatment and the antigenic vaccines developed so far are not sufficient to control infection rates [205]. Praziquantel, the

current drug of choice, is effectively used against all forms of schistosomes. However, drug resistance to certain strains of schistosomes is starting to emerge [39]. Although praziquantel was developed in the early 1970s and its pharmacokinetics in mammals studied extensively [42, 206], the exact molecular mechanism of this drug has yet to be elucidated. Likewise, much is known about the schistosome's life cycle (see **Chapter 1**) and its mode of infection. However, the molecular mechanisms by which it interacts with the host are still being determined.

Many studies are now pointing to the role of hormones and nuclear receptors (NRs) as a mode of interaction between parasite and host (for a review, see de Mendonça *et al.* [207]). Hormones control a variety of developmental and differentiation processes in metazoans. In schistosomes, they influence the development, survival, growth, migration and maturation of the parasite [207, 208]. These hormones are thought to originate from the host and/or the parasite. Despite evidence for hormonal influences on schistosome function via specific NRs, it is still unclear which schistosome NRs are involved [207-211].

Through the studies of schistosome NRs, the molecular signalling pathways between the schistosome and its hosts (vertebrate and snails) and also within the parasite itself are starting to emerge.

Nuclear receptors

NRs are increasingly viewed as very promising drug targets as they bind small molecules that are easily modified by drug design, and they regulate key metabolic pathways associated with major diseases (for example, cancer, diabetes and

osteoporosis) [207, 211, 212]. NRs belong to a superfamily of ligand-regulated, evolutionary related transcriptional factors. They play a critical role in metamorphosis (differentiation and development), metabolism, homeostasis and reproduction in metazoans [209, 213].

NRs induce transcription of other genes when a coactivator, or ligands such as steroids, retinoids, vitamin D, among others, are bound [208, 209, 213]. Here, the coactivators are proteins that bind to NRs already bound to the regulatory region of the target gene, to facilitate the transcription of the gene into mRNA. **Figure 30** shows how hormones modulate gene expression through specific NRs.

However, an increasing number of "orphaned" NRs are being discovered for which no ligands are yet known. The mechanism by which these orphan NRs activate (or inhibit) transcription remains unclear [207, 208, 210, 213].

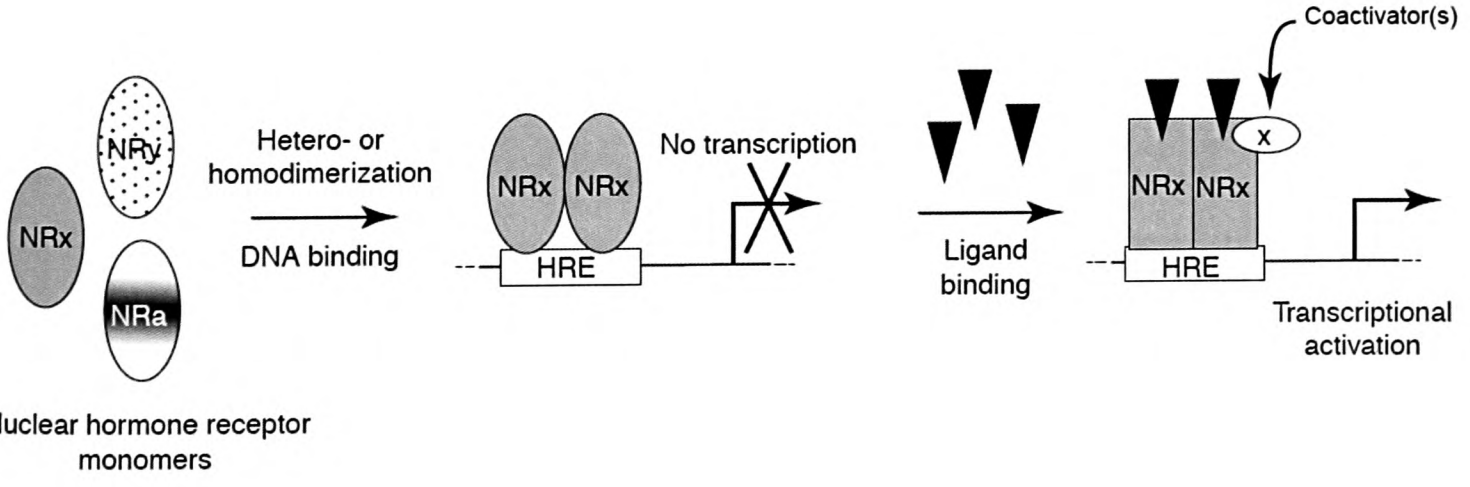


Figure 30. Hormones modulate gene expression through nuclear receptors by either repressing or activating transcription. The DNA binding domain of the NRs' complex is a motif called the "Hormone response element" located inside the promoter of the target gene. In ligand-dependent NRs, a conformational change occurs when ligands (shown as black triangles) bind and associate with coactivators. Figure adapted from [207].

Methods and Materials

Obtaining schistosome sequences

S. japonicum EST sequences

84,499 *S. japonicum* sequences were obtained from the expressed sequenced tag (EST) sequencing project at the Chinese National Human Genome Centre, Shanghai (China) (http://function.chgc.sh.cn/sj-proteome/index.htm). These had been labelled as "clean ESTs" because repetitive, mitochondrial, ambiguous and vector sequences had been removed [214]. These *S. japonicum* EST sequences were sequenced from transcriptomic cDNA libraries from the following six schistosome life cycle stages:

1. Adult (mixed sex) worms
2. Egg
3. Schistosomula
4. Miracidia
5. Adult female worms, and
6. Adult Male worms

Despite correspondence with the authors requesting a more complete categorisation of these ESTs, it remains unclear why mixed sex adult worms were not separated into "Adult female" and "Adult male" categories. As such, I postulate that the "Adult mixed sex worms" may exhibit a different phenotype as they were analysed and maintained separately as mentioned in the original paper published [214]. Hence I have maintained the "Adult mixed sex worms" as a separate category/stage for comparison with the other stages.

S. japonicum proteomic data

Proteomic data from the following nine stages and tissues were used in this project:

1. Cercariae
2. Adult males
3. Adult females
4. Mixed sex worms
5. Eggs
6. Miracidia
7. Schistosomula
8. Tegument (from hepatic schistosomula, adult female, adult male and mixed sex worms)
9. Eggshell (obtained after miracidia hatched from the eggs)

These proteomic sequences were also obtained from the Chinese group and totalled 5,702. The group identified protein fragments using mass spectrometry and identified putative ORFs using protein and peptide sequences from human, mouse and rabbit hosts and transcriptomic *S. japonicum* EST data [214]. These peptide sequences and their corresponding mRNA sequences were made available on the Chinese *S. japonicum* website (as above). I obtained these mRNA sequences and used them to build full-length transcripts of the *S. japonicum* cDNA (explained below).

S. mansoni sequences

S. mansoni sequences were obtained from the joint project between the Wellcome Trust Sanger Institute (UK) and The Institute for Genomic Research (now the J Craig Venter Institute) (USA). About 3.8 million pooled reads were generated from whole

genome shotgun sequencing with a 7-8 fold coverage. Version 3.0 of the genome assembly was published recently [46] and contained 50,367 contigs with an average size of about 7.6 kb. 18,831 scaffolds were built from these contigs with an average size of 20.4kb, the maximum size being more than 2Mb.

Gene predictions were made with the TWINSCAN [215] program. Gene structures were predicted using alignment-based methods, where predictions are made using a gene structure model, evolutionary conservation information between *S. mansoni* and *S. japonicum* and the DNA to be annotated.

At the start of this project, version 4.0 of the assembly was available and I recently updated my analyses with the latest data set of 13,185 cDNA sequences (dated 30/04/08; versioned "e") obtained from

ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/.

Clustering of the EST sequences

To enable a full-length transcript of the *S. japonicum* cDNA to be built, both the proteomic data and EST sequences were used. As mentioned above, the mRNA sequences corresponding to the protein sequences obtained through mass spectrometry were made available. The combined *S. japonicum* mRNA and EST sequences from all the different life cycle stages were clustered based on their synonymous rate substitution (dS) values as described below. All *S. japonicum* ESTs and mRNA sequences were aligned to their translated *S. mansoni* cDNA counterparts using the alignment model "protein2dna" in the Exonerate [105] program. The dS values were then calculated using the CODEML [146] program between *S. japonicum*

sequences and its matching *S. mansoni* cDNA sequence. In cases where many *S. japonicum* outparologous sequences matched one *S. mansoni* cDNA, only one *S. japonicum* EST/mRNA was chosen. The *S. japonicum* EST/mRNA with the lowest dS value was chosen. It is proposed that using dS as a distinguishing criterion allows true orthologues to be selected, in preference to out-paralogues since these would have elevated dS values. *S. japonicum* inparalogues were conservatively discarded as this method would not have been able to choose only one sequence from among them correctly.

Finally, for each *S. mansoni* cDNA sequence, all matching *S. japonicum* EST/mRNA sequences were merged into a consensus sequence. The dN, dS and dN/dS values for the final paired *S. japonicum* consensus sequence and *S. mansoni* cDNA were calculated once more using the CODEML tool.

Obtaining stage specific data

The *S. japonicum* sequences (either EST or mRNA) used to build the consensus sequence against its *S. mansoni* candidate orthologue were previously categorised into the stages they were sequenced from by the Chinese group. Hence, I assigned the *S. mansoni*/*S. japonicum* orthologue pair to the schistosome life cycle stage based on the *S. japonicum* expression data. For instance, if there were three *S. japonicum* EST/mRNA sequences from the stages "Egg", "Schistosomula" and "Adult female" that were used to build the consensus sequence, I would classify the corresponding *S. mansoni* sequence as being expressed in each of the "Egg", "Schistosomula" and "Adult female" stages. This *S. mansoni* sequence would thus be classified as being non-stage specific as it was expressed in more than one stage. If the *S. japonicum*

sequences were all derived from only one stage, the corresponding *S. mansoni* sequence would be defined as being stage-specific.

I discarded stage-specific data for the following stages and tissues: "Miracidia", "Tegument", "Eggshell" and "Cercariae" as they contained fewer than ten transcripts each. This left me with the following five stage-specific data sets: "Adult Female", "Adult Male", "Schistosomula", "Mixed sex" and "Egg".

Methods used to account for short sequences

Short sequences lead to potential inaccuracies in dN/dS values. From previous simulations performed by a post-doc in my lab (Andreas Heger, unpublished), it was found that CODEML estimates of dS and κ (transition/transversion ratio) for short sequences had large standard error ranges, and were thus generally more inaccurate. In addition, shorter sequences tended to have spuriously high dN/dS values. Thus, to ensure that differences seen in stage-specific dN/dS values were significant, I had to account for the length biases in rates caused by short sequences.

As CODEML predicts dN, dS and dN/dS values from aligned sites of paired sequences (**Figure 31**), I investigated whether stage-specific dN/dS values varied with their numbers of aligned codons. As expected, from these stage-specific distributions (**Figure 32**), higher dN/dS values were achieved for paired orthologous sequences of *S. japonicum* and *S. mansoni* that contained smaller numbers of aligned codons.

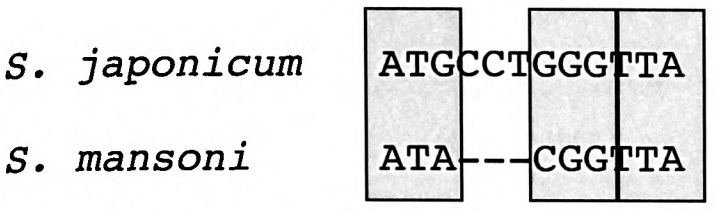


Figure 31. Sequence alignment between a *S. japonicum* and a *S. mansoni* sequence. The three grey-boxed areas depict the nine "aligned sites" or three "aligned codons" where gaps are absent from either sequences. CODEML predicts the dN, dS and dN/dS values of the sequences over all aligned codons. Hence, the longer the number of aligned codons, the more information CODEML has to accurately predict substitution rates.

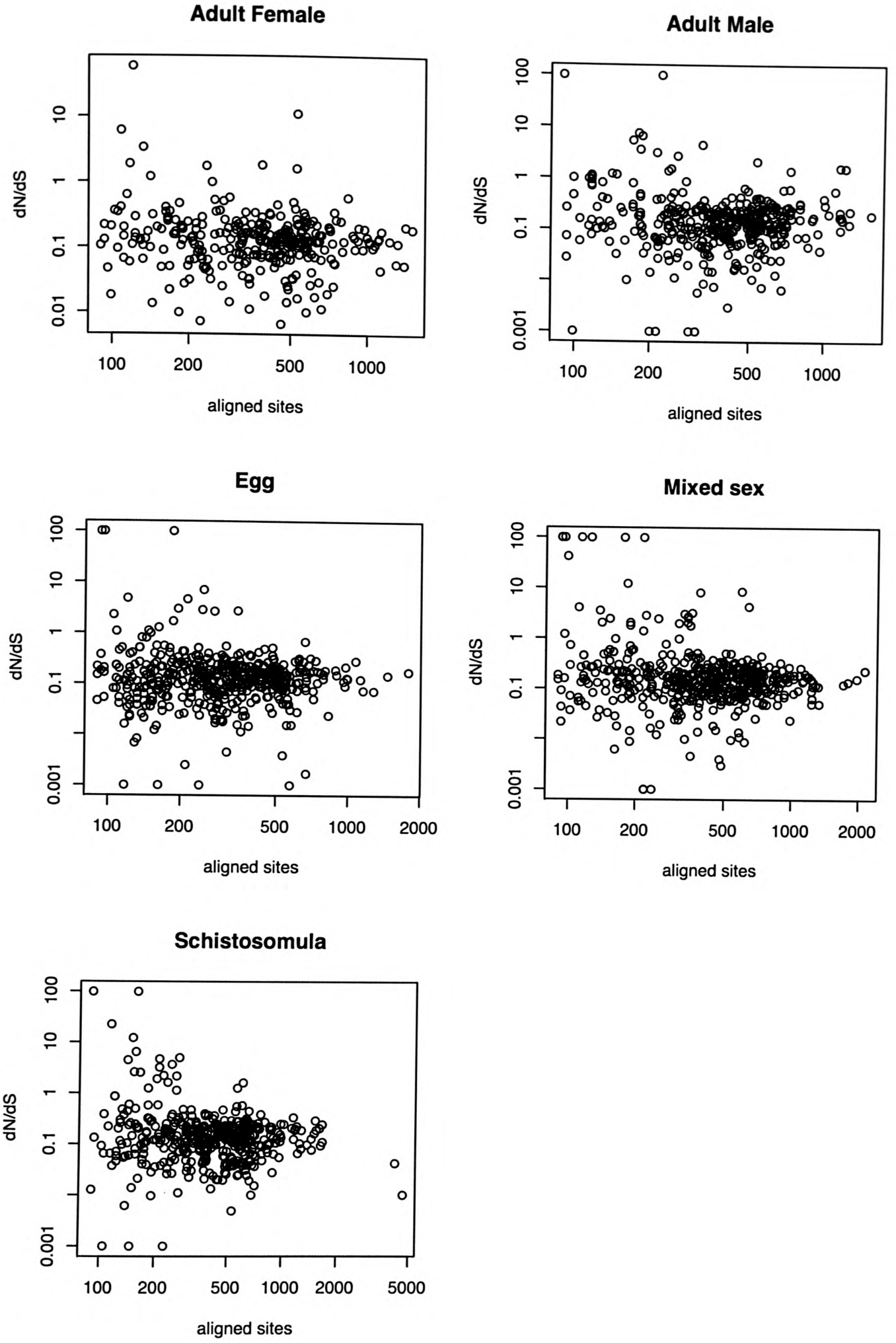


Figure 32. Log-Y plot (y-axis is on a log scale) of dN/dS values for stage-specific transcripts *vs* their numbers of aligned codons. Smaller numbers of aligned codons in the paired orthologous sequences of *S. japonicum* and *S. mansoni* tended to produce larger departures in dN/dS values.

Accounting for length biases by creating mosaic sequences

As discussed above, shorter sequences introduce greater variances in dN/dS estimations. Hence to address the issue of having short sequences, I randomly sampled from all aligned codons (between *S. mansoni* and *S. japonicum*) to built longer "mosaic" gene sequences. Before selecting the codons from each group of sequences, for example, stage-specific "Adult female" genes, I first shuffled all aligned pairs of codons so that every aligned pair had an equal chance of being chosen and then picked the first pair of codons without replacement. This method could be likened to a game of cards – each time before a card was dealt (without replacement), the deck of cards was shuffled.

Because a longer sequence length enabled a more accurate estimation of substitution rates by CODEML, I generated 1000 mosaic sequences each containing 1000 codons. For the notation used from this point on in the thesis, I use "c" to denote number of codons in a sequence and "s" to indicate the sample size. Thus, "c1000.s1000" denotes 1000 sequences each containing 1000 codons.

Accounting for length biases by applying a threshold

In order to examine the functional enrichment of GO categories among the fast evolving genes and to determine which stage was enriched within fast/slow evolving genes, I had to use the original sequences instead of sampled data. Here, I used another method to account for length biases which removed short sequences below a length threshold. The threshold was applied by limiting the minimum number of aligned codons allowed in a paired sequence. I chose to use one threshold for further analyses: namely 150 aligned pairs of codons. This was the threshold that retained

most sequences while retaining relatively high accuracy in dN/dS estimations (**Figure 32**).

These were the steps taken in choosing this threshold: paired sequences containing a minimum from at least 150 to 450 codons were counted and their median dN/dS values tabulated. **Table 16** summarises the different thresholds applied, the numbers of sequences remaining in each stage after the threshold was applied and the median dN/dS values of the resulting sequences. The gene count of each stage decreased approximately exponentially as higher numbers of aligned codons were required (**Figure 33**).

a) Stage-specific median dN/dS values

Thresholds	Adult female	Adult male	Egg	Mixed sex	Schistosomula
None (raw data)	0.127	0.136	0.130	0.149	0.137
>c150	0.123	0.149	0.131	0.142	0.132
>c170	0.126	0.148	0.123	0.141	0.129
>c190	0.124	0.168	0.137	0.141	0.126
>c200	0.124	0.166	0.153	0.146	0.132
>c250	0.131	0.176	0.147	0.137	0.129
>c300	0.129	0.183	0.127	0.144	0.135
>c350	0.128	0.183	0.135	0.138	0.142
>c450	0.192	0.190	0.160	0.159	0.117

b) Transcripts count in each stage

Thresholds	Adult female	Adult male	Egg	Mixed sex	Schistosomula
None (raw data)	332	391	472	487	430
>c150	140	163	115	219	197
>c170	100	127	81	186	166
>c190	69	82	53	158	133
>c200	57	75	46	136	117
>c250	28	26	19	59	52
>c300	20	18	11	36	33
>c350	11	12	6	22	24
>c450	3	1	2	5	12

Table 16. Summary of stage-specific median dN/dS values and transcript counts as the codon length threshold increases. Median dN/dS values are shown to 3 decimal places. A variety of thresholds were applied to paired sequences containing the minimum number of codons. For instance, at a threshold of 150 codons, only paired sequences with more than 150 aligned codons (450bp) were considered. The numbers of transcripts in each stage decreased approximately exponentially as the threshold was increased.

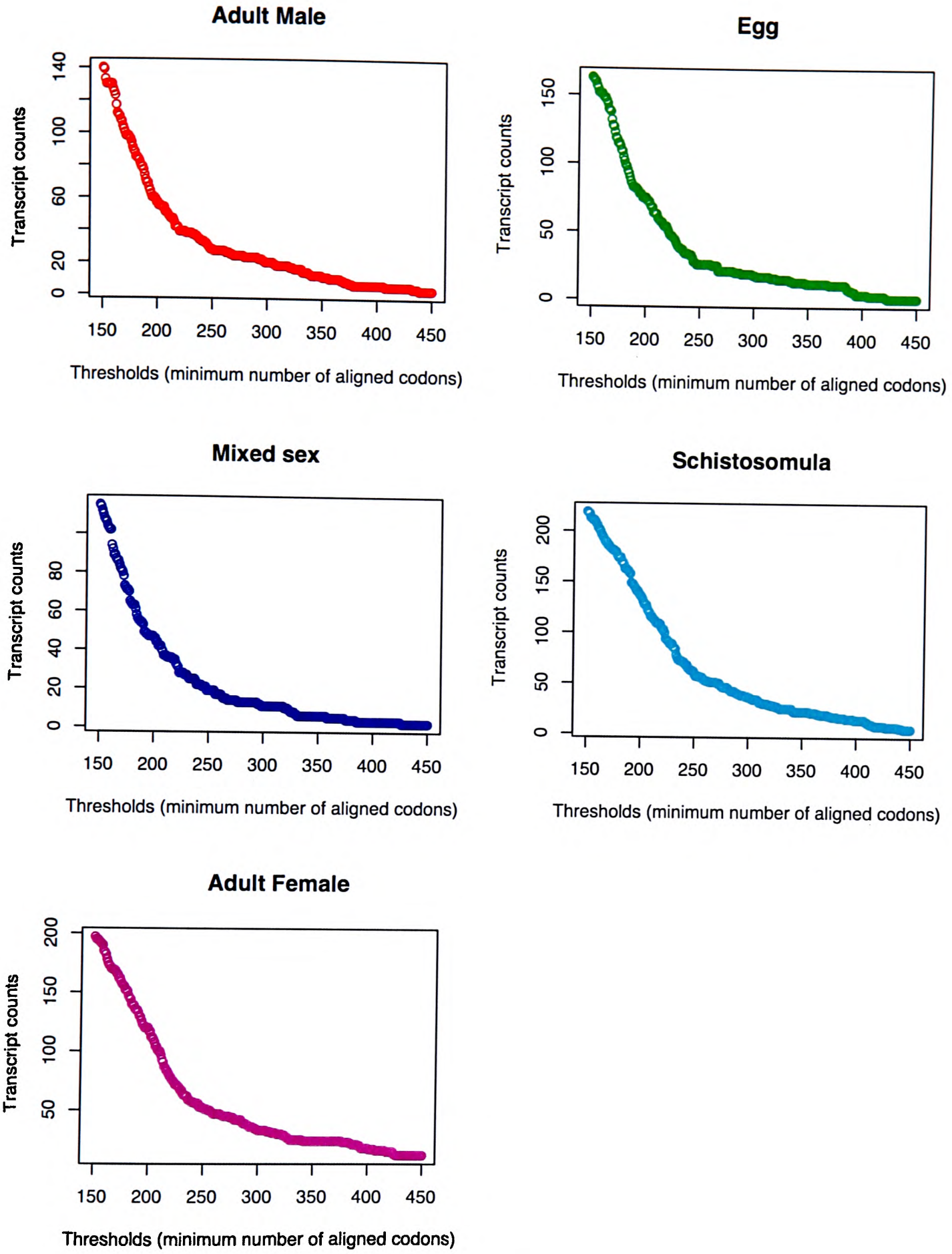


Figure 33. Plots of the transcript counts in each stage at different length thresholds. The number of gene transcript counts decreases in an exponential fashion as larger numbers of aligned codons are required in a sequence. The range of the thresholds implemented is 150 to 450 codons, incremented by 2 codons (*i.e.* 150, 152, 154, 156 etc).

Mapping of GO terms to GO-slim terms

InterPro [216] is an integrated database of the following protein signature databases: PROSITE, PRINTS, ProDom, Pfam, SMART, TIGERFAMs, PIRSF, SUPERFAMILY, Gene3D and PATHER [216]. It is an excellent resource for protein families, domains and functional sites. Hence all GO annotations I had for my schistosome genes were inferred from InterPro annotations that I obtained from annotating *S. mansoni* sequences. This was because the *S. mansoni* sequences I used from the TIGR/Sanger project had yet to be published and thus their corresponding GO terms had not yet been released into the public GO database.

The Gene Ontology (GO) categories can also be broadly classified by GO-slim terms. These GO-slim terms are the parent nodes of the GO terms. It is useful to apply GO-slim terms when a general classification of terms is required and fewer classification categories are desired. This increases the power of the statistical analysis by using the least number of tests possible [149].

In order to map GO to GO-slim terms, I used the "map2slim" PERL script produced by Chris Mungall available on the GO website (http://www.geneontology.org/GO.slims.shtml), and the ontology exchange format file ("OBO" file) used for the GO Annotation project (GOA) (downloaded 7th May 2008) [217] merged with the generic GO-slim (downloaded 29th April 2008) file. The reason I merged both the GOA and generic GO-slim file was because the GOA project team produced the OBO file containing the mapping of InterPro IDs to GO annotations [218].

Multiple testing correction

To account for multiple testing, both the Bonferroni and false discovery rate method were used. The conservative Bonferroni correction was used when there were small numbers of tests (≤10). In other cases, I controlled the rate of false discoveries where there were a large number of tests (>10) (described in **Chapter 2**).

Results

Results of clustering *S. japonicum* EST and mRNA sequences

Transcripts were built from both the 14,962 ESTs and 5,702 mRNA sequences from the Chinese *S. japonicum* project using 13,185 *S. mansoni* cDNA sequences as templates (see **Methods and Materials** above). A total of 7,458 *S. japonicum* transcripts were obtained. This was fewer than the published 8,420 sequences containing potential ORFs derived by the Chinese group [214]. Detailed comparisons of the clustering methods used by the Chinese group and myself were not possible. This was due to the Chinese group using a commercial package, CAT 3.5 (Pangea) [44] for which sufficient information on the algorithms employed was not available. Furthermore, in their calculations of dN, dS and dN/dS values between *S. mansoni* and *S. japonicum* alignments (average values of 0.165, 0.926, 0.149 respectively [214]), the Chinese group used a counting method by Nei and Gojobori (SNAP [219]), whereas I used the maximum likelihood method of CODEML [146]. The median dN, dS and dN/dS values for my 7,458 sequences were 0.115, 0.907 and 0.121 respectively (3 significant figures).

Do stage-specific sequences exhibit elevated dN/dS values?

In order to investigate whether stage-specific sequences show elevated dN/dS values, I first tested the null hypothesis that there was no significant difference between the evolutionary rates of stage-specific genes and non-stage-specific genes.

After this, I compared the evolutionary rate distributions among all five stages ("Adult Female", "Adult Male", "Egg", "Mixed sex", "Schistosomula") to determine which stage showed the most elevated dN/dS values. Finally, I examined in detail the fastest 10% evolving genes across all stages to determine their functionality and to consider possible reasons for their unusually rapid evolution. Below are the results of these analyses.

Comparing stage-specific and non-stage-specific genes

I separated all stage-specific and non-stage-specific sequences from my 7,458 sequences (see **Methods and Materials** in this chapter). **Table 17** shows the median dN/dS values and transcript counts from each of the five stages, and from the non-stage-specific group based on the original sequences. I then accounted for biases in sequence length by randomly sampling 1000 codons to generate 1000 mosaic sequences for each of the stages, and for the non-stage-specific group (see **Methods and Materials** section).

	Stage-specific					Non-stage-specific
	Adult Female	Adult Male	Egg	Mixed sex	Schistosomula	-
Median dN/dS	0.127	0.136	0.130	0.149	0.137	0.117
Transcripts count	332	391	472	487	430	5334

Table 17. Stage-specific and non-stage-specific median dN/dS values (3 significant figures) and gene counts of the original data.

I then considered whether there was a significant difference between the stage-specific and non-stage-specific sequences, having accounted for biases in sequence length. I found that stage-specific sequences were significantly elevated in dN/dS values (p-value < 2.2×10^{-16}) (**Figure 34**). **Figure 35** shows the distributions for each of the five stage-specific groups. Thus, transcripts that were only expressed in one stage had, on average, higher dN/dS values than those expressed in multiple stages.

The differences between each of the five stage-specific groups are discussed below.

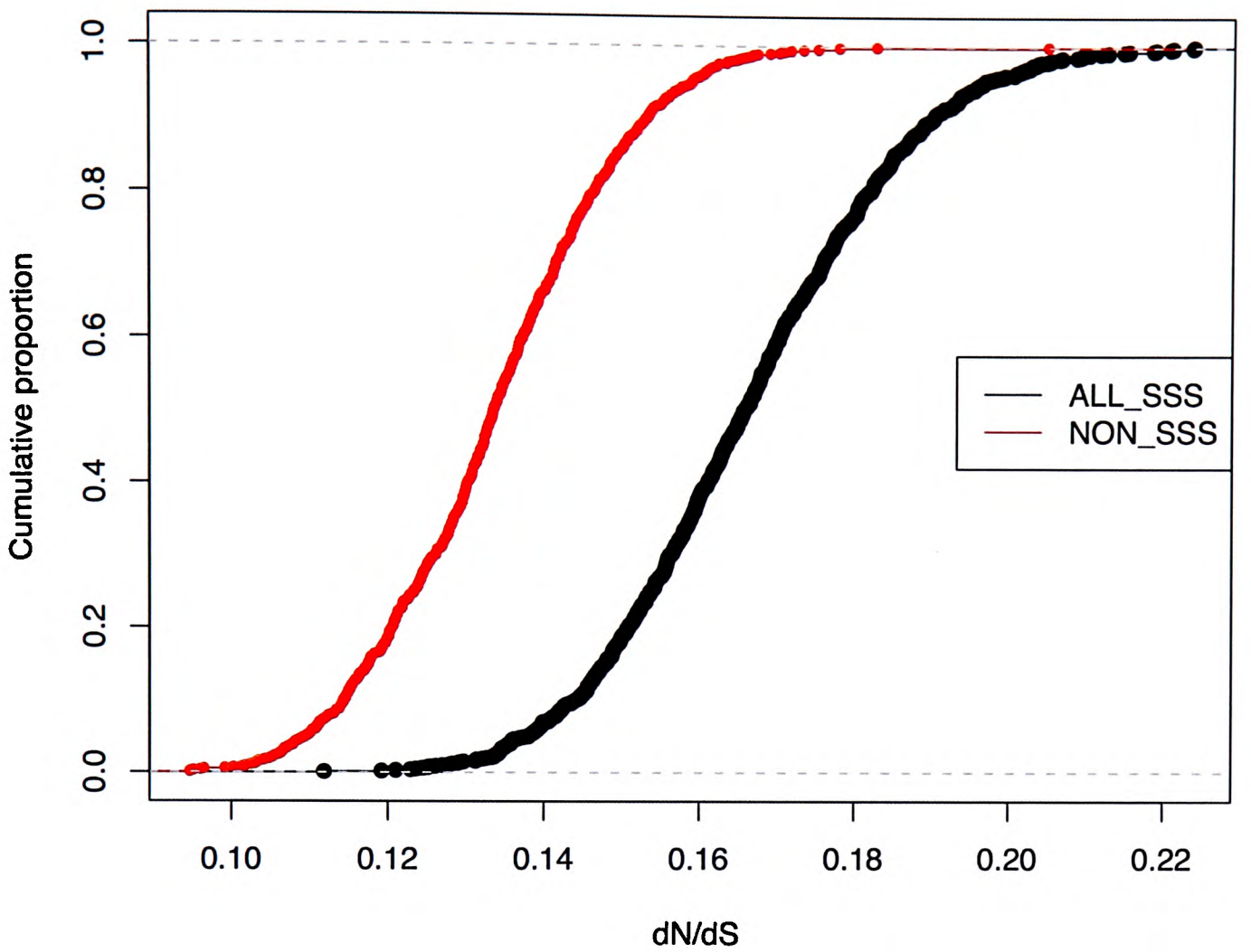


Figure 34. Stage-specific sequences tended to have elevated dN/dS values compared to non-stage-specific sequences. The cumulative frequency distribution for all stage specific genes (shown in black) was compared to that for all non-stage-specific genes (shown in red). Each distribution was composed of 1000 sequences consisting of 1000 codons each. "ALL_SSS" refers to "all stage-specific sequences" and "NON_SSS" represents "non-stage-specific sequences".

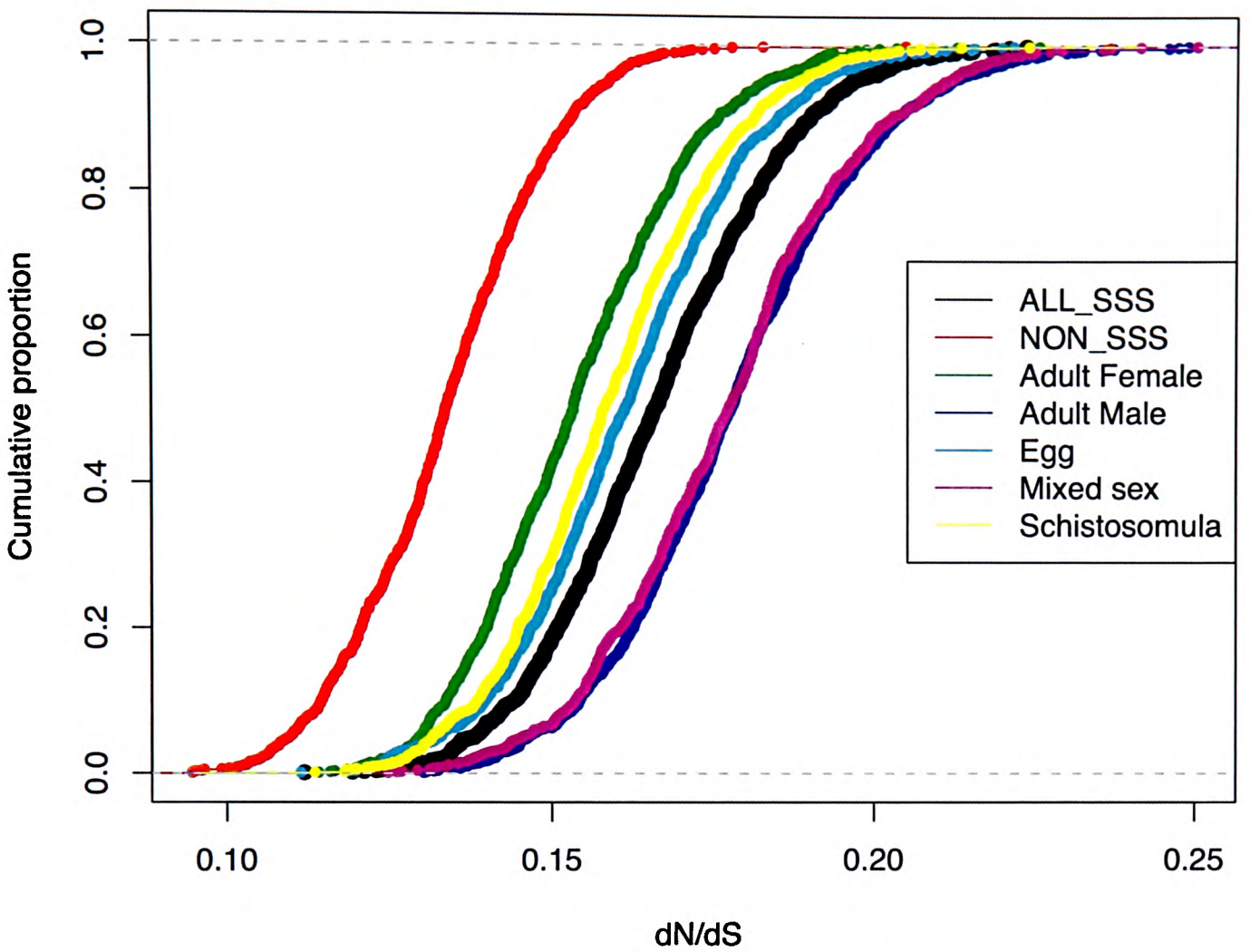


Figure 35. Stage-specific sequences tended to have elevated dN/dS values compared to non-stage-specific sequences. Cumulative frequency distributions of dN/dS values of individual stages are represented by different colours (Adult Female – green, Adult Male – blue, Egg – light blue, Mixed sex – purple and Schistosomula – yellow). The distribution of all stage specific genes (shown in black) was compared to that of all non-stage-specific genes (shown in red). Each distribution was composed of 1000 sequences consisting of 1000 codons each. "ALL_SSS" refers to "all stage-specific sequences" and "NON_SSS" represents "non-stage-specific sequences".

Next, I tested whether the dN/dS values of transcripts tended to become more suppressed or elevated as the numbers of stages they were expressed in increased. For this, I categorised all non-stage-specific genes into the numbers of stages in which they were expressed (**Table 18**).

Number of stages	All	1	2	3	4	5	6	7	8
Number of genes	7458	2112	1742	1372	1136	763	273	45	2

Table 18. Numbers of genes expressed in one or more stages. "All" indicates the total number of non-stage-specific genes, excluding the four stages that each had fewer than ten transcripts expressed in them (see Methods and Materials). The stage numbers indicate the number of stages in which a gene appears is found to be expressed.

Using randomly sampled sequences from each of these partitions in **Table 18**, I found that the more stages transcripts were expressed in, the lower the dN/dS value tended to be (**Figure 36**).

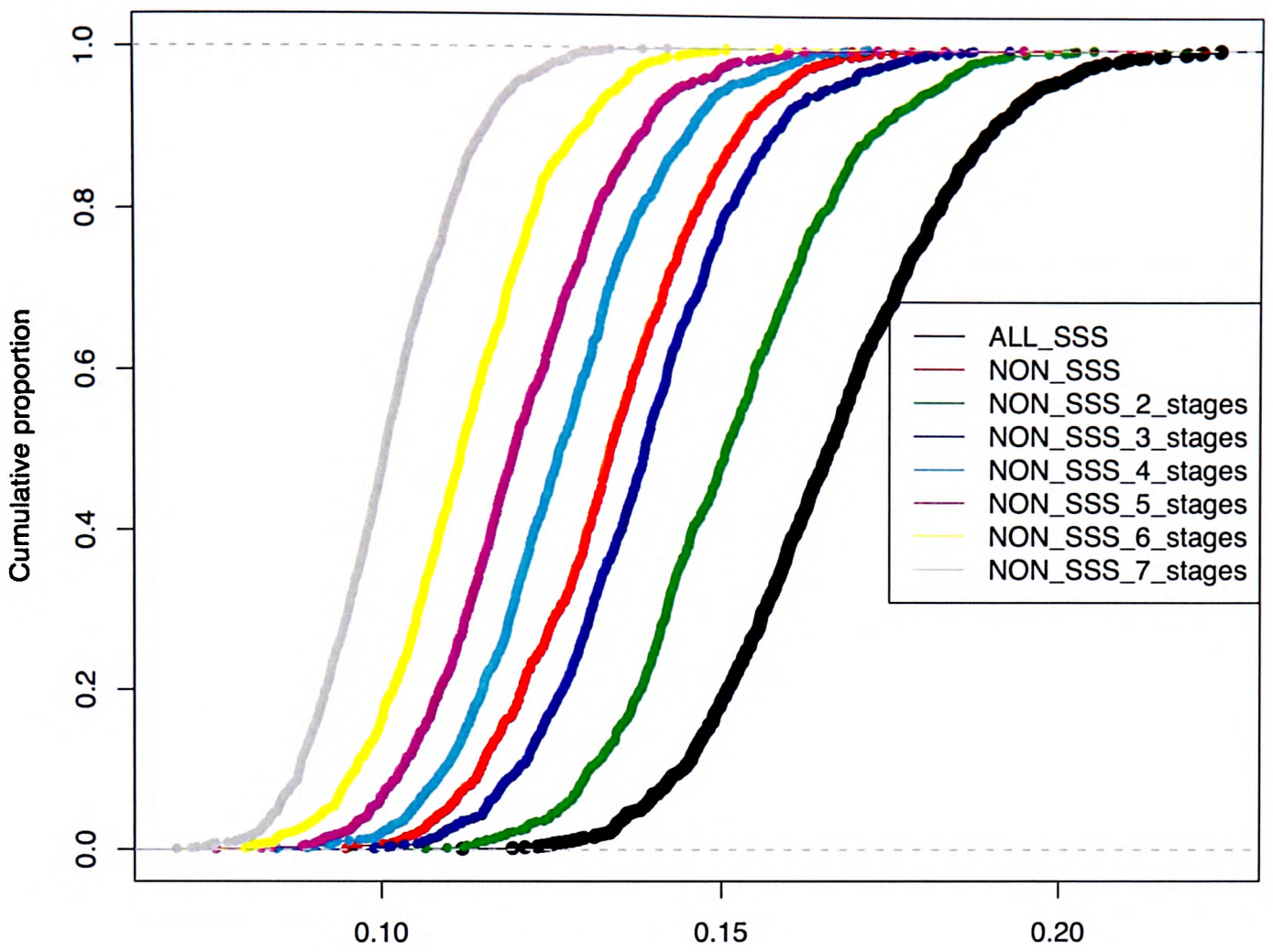


Figure 36. Cumulative frequency distributions of dN/dS values from nine stages: Adult Male, Adult Female, Egg, Schistosomula, Mixed sex, Cercariae, Miracidia, Tegument and Eggshell. Samples were 1000 mosaic sequences consisting of 1000 codons each. dN/dS values of genes tended to become more suppressed as they were expressed in increasing numbers of stages. Although there were nine stages, only seven stages from which non-stage-specific genes were sampled are shown. This is because there were no genes expressed in all nine stages and only two genes were expressed in eight stages, hence random sampling to obtain the dN/dS distributions from the eight and nine stages was not performed. "ALL_SSS" refers to "all stage-specific sequences" and "NON_SSS" represents "non-stage-specific sequences".

Differences between the stages

To test for significant differences between distributions of dN/dS values for the five stages, I used the Kruskal-Wallis test. Any differences detected by this test would be due to the median values differing significantly [130]. I also tested if the underlying dN/dS distributions were significantly different between any two stages using the Kolmogorov-Smirnov test. All these distributions were obtained using the random sampling approach (see **Methods and Materials**). The results of these tests are tabulated in **Table 19**. The Kolmogorov-Smirnov test showed significant differences ($p < 0.005$) among the distributions between all stages except for that of "Adult male" and "Mixed Sex". Using the Kruskal-Wallis test, there were significant differences ($p < 0.005$) detected between the median values for all stages except between the following pairs of stages: *1*. "Adult male" and "Mixed sex" and *2*. "Schistosomula" and "Egg".

a) Kruskal-Wallis test

	Adult female	Adult male	Egg	Mixed sex	Schistosomula
Adult female					
Adult male	*				
Egg	*	*			
Mixed sex	*	×	*		
Schistosomula	*	*	×	*	

b) Kolmogorov-Smirnov test

	Adult female	Adult male	Egg	Mixed sex	Schistosomula
Adult female					
Adult male	*				
Egg	*	*			
Mixed sex	*	×	*		
Schistosomula	*	*	*	*	

Table 19. Summary of Kruskal-Wallis test and of Kolmogorov-Smirnov test between pairs of stages. For simplicity, only the lower half of the matrix is used to show results. A * indicates a significant difference between the medians ($p < 0.005$) and a × indicates no significant difference ($p > 0.005$) for that pair of stage combinations.

The distributions of dN/dS values for the five stages are illustrated in **Figure 37** and **Figure 38**. The dN/dS distributions for "Adult male" and "Mixed sex" stage were not significantly different from one another. This may be an indication that the "Mixed sex" group consisted predominantly of adult male worms. The "Adult female" stage appears to be associated with more suppressed dN/dS values compared to other stages. **Figure 38** represents the same data but provides an indication of the standard error of median values (red horizontal bars between the green crosses).

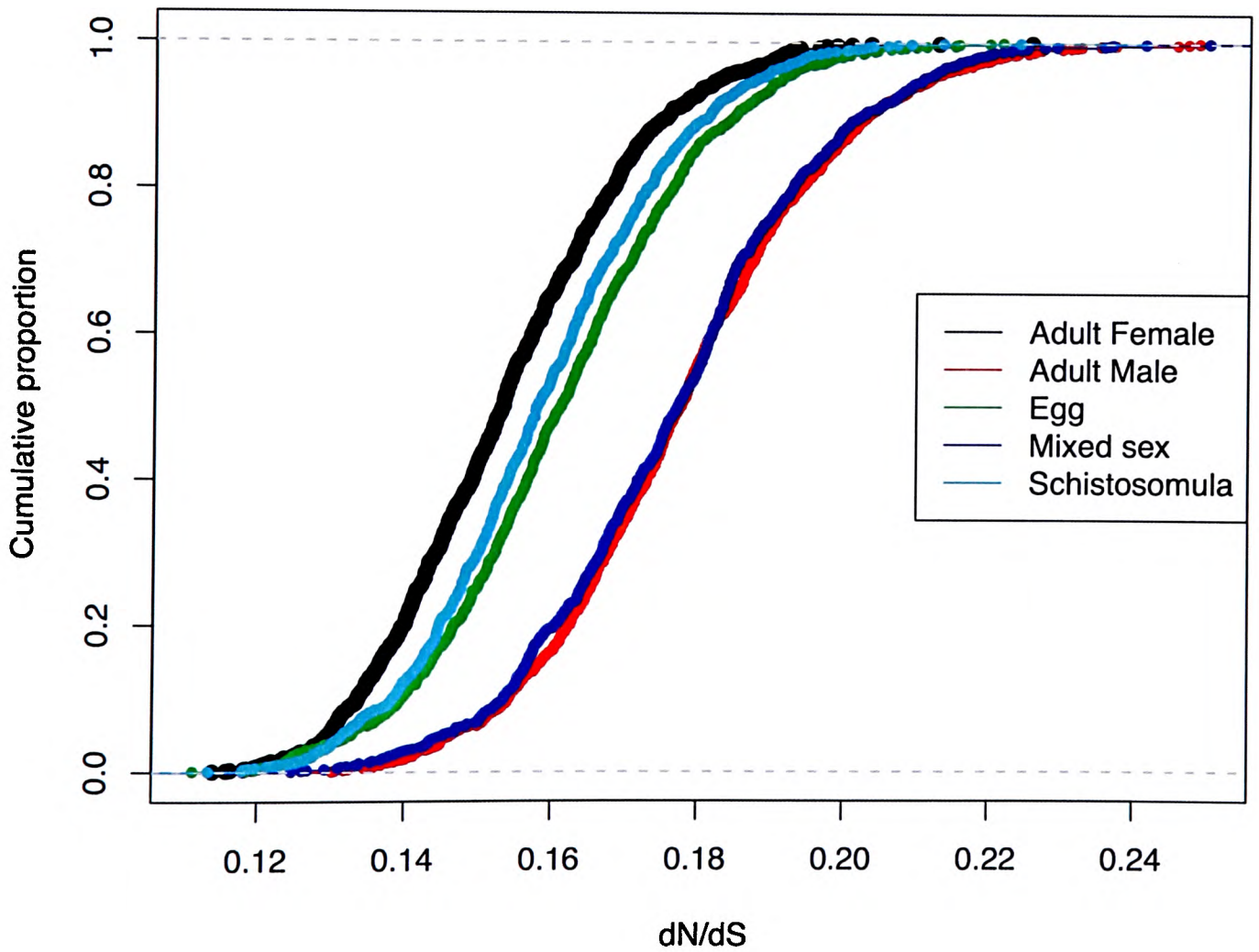


Figure 37. Cumulative frequency distributions of stage-specific dN/dS values. These are derived from 1000 randomly sampled mosaic sequences each containing 1000 codons. "Adult female" has a distinctly less elevated dN/dS value distribution compared to other stages. "Adult male" and "Mixed sex" have similar dN/dS distributions; this indicates that "Mixed sex" may consist mainly of adult male worms.

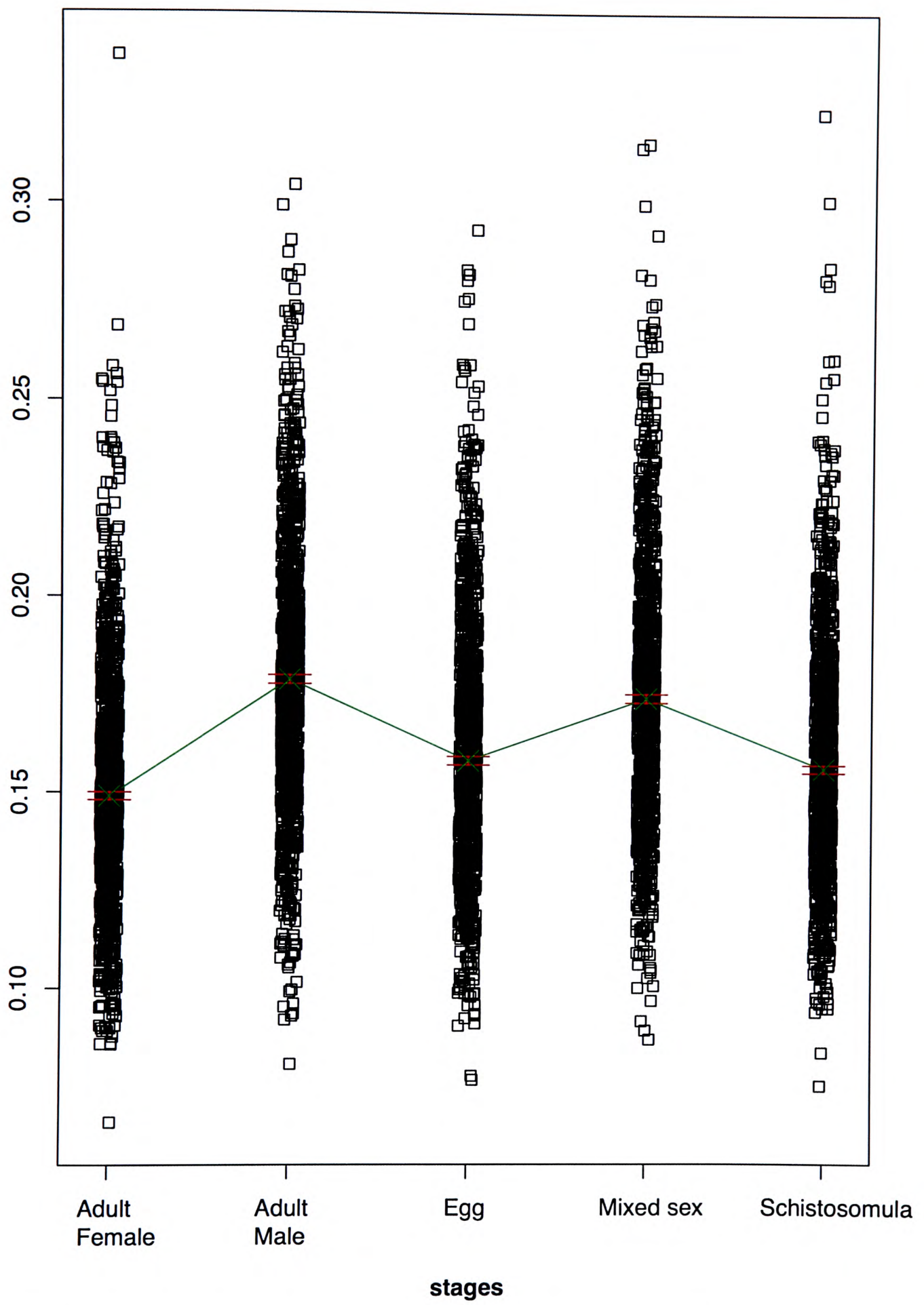


Figure 38. dN/dS values for each of the stages. Randomly sampled 1000 mosaic sequences of codon length 1000. dN/dS values are plotted and overlaid with an indication of the medians (green crosses between the red bars) and standard error of the medians (red bars).

Are fast evolving genes particularly expressed in certain stages?

From the above findings, I noted that stage-specific genes tended to be those with the most elevated dN/dS values. Thus I wanted to know if there was a particular stage that was significantly enriched in these fast evolving genes. For this, I had to use the original data to identify such genes. Thus, as previously discussed, I applied different thresholds (sequences with more than 150, 250 or 350 aligned codons) on my data to account for biases in length. Doing so, I found that the top 10% of the fastest evolvers were significantly ($p < 0.025$) enriched with genes from the "Adult male" stage when a threshold of at least 150 or 250 aligned codons was applied. The numbers of genes from the "Adult male" stage and the corresponding p-values are shown in **Table 20**. However, this enrichment of "Adult male" genes among fast evolving genes was lost when a minimum of 350 aligned codons was required and multiple testing correction (Bonferroni correction) was applied. This was likely due to the number of available genes decreasing approximately exponentially as more stringent thresholds were chosen (**Figure 33**).

Top 10% of the fastest evolvers	Stage	Counts in sample	Counts not in sample	p-value
>c150	Adult Male	27	136	0.002
>c250	Adult Male	8	18	>0.001
>c350	Adult Male	3	9	0.111

Table 20. "Adult Male" is significantly ($p < 0.01$) enriched with fast evolving sequences within the top 10% of fastest evolving genes. Thresholds of 150, 250 and 350 aligned codons were applied such that pairs of sequences containing fewer codons than the thresholds were discarded. When a threshold of 350 aligned codons was applied, "Adult Male" was no longer significant ($p > 0.01$).

Functional annotations for fast evolving genes

InterPro domain annotations of protein sequences

What are the functions of these fast evolving genes in *S. mansoni* and *S. japonicum*? Using the longer and more complete *S. mansoni* orthologue, I queried the integrated database InterPro to annotate the top 10% of the fastest evolving genes (84 genes) **Table 22**. Only domain annotations with *E*-value < $1x10^{-5}$ were accepted. Five *S. mansoni* sequences predicted by the TIGR/Sanger consortium had stop codons in their protein sequences and could not be annotated by InterPro. Of the remaining 84 genes, there were 25 genes that had no InterPro annotations; these were significantly enriched with fast evolving "Adult Male" genes ($p < 0.005$). This significant enrichment of genes lacking domain annotations among the fastest evolving genes ($p < 9.38 \times 10^{-8}$) could have arisen from genes evolving at such a fast rate that sequence similarities to existing domains in InterPro were not detectable.

Mapping of the InterPro resources to GO

As the *S. mansoni* project from which I obtained the sequences has not yet been published, Gene Ontology (GO) [128] entries for the corresponding sequences were not available. Fortunately, as part of the ongoing Gene Ontology Annotation project of complete genomes [217], mapping of InterPro resources to GO had been carried out. Thus I was able to annotate GO terms to orthologue pairs using InterPro annotations [217, 218].

Enrichment of GO and GO-slim categories among fast evolving genes

After obtaining the GO terms assigned for each gene, I tested if certain categories of the GO terms were significantly over-represented. After controlling for false discoveries at a 5% level, the following three terms were found to be significantly enriched among the fastest evolving genes:

- Biological process – transcription (FDR < 0.05)
- Molecular function – transcription factor activity (FDR < 0.05)
- Cellular location – cytosol (FDR < 0.05)

The above tests were repeated using GO-slim (broader categories compared to GO) and similar results were obtained (**Table 21**). I chose to ignore any GO-slim term that was populated by only single genes (in this instance, cytosol in the "Cellular location" category).

A) Biological Process

GO-ID	GO term	*r*	*k*	*p*	*N*	p_{GO}-value	Genes
6350	Transcription	5	19	34	263	8.10×10^{-2}	Smp_097730 Smp_144170 Smp_168810 Smp_173470 Smp_174320

B) Molecular Function

GO-ID	GO term	*r*	*k*	*p*	*N*	p_{GO}-value	Genes
3700	Transcription factor activity	4	22	18	336	2.30×10^{-2}	Smp_097730 Smp_174320 Smp_144170 Smp_168810

Table 21. GO-slim categories that are significantly enriched among the fastest evolving 10% of genes at a false discovery rate of 5%. From the definitions of the hypergeometric function in Chapter 2, *r* is the number of genes contributing to the particular trait sampled from a sample size of *k*, *p* is the proportion of the *r*-type item in the population *N*.

Detailed analyses of genes of interest

From the above analyses, five genes listed were singled out as genes of interest. These genes have been categorised in two GO-slim / GO categories that were significantly over-represented among the top 10% of fast evolving genes. Properties of the five genes are listed in **Table 22**. Below, I provide a detailed discussion of each of these five genes in turn.

Equivalent analyses using a previous build of *S. mansoni* genes (version "a" dated 06/02/07) yielded a sixth gene, Smp_180860 that was absent from the updated version (version "e" dated 30/04/08). Smp_180860 was one of the genes represented in the two GO-slim categories that were significantly enriched (**Table 21**). I have included an analysis on this gene because it has been touted as one of the largest NRs found [210] and may be a potential vaccine candidate.

Gene name	Number of aligned sites (bp)	Stage in which gene is expressed	Domains in Protein (prediction program)	dS	dN/dS	Length of *S. mansoni* protein sequence (bp)
Smp_097730	564	Adult female	MADS(HmmSmart), SRF-TF (HMMPfam), MADSDOMAIN (FPrintScan), MADS-BOX2 (ProfileScan)	0.515	0.341	1290
Smp_144170	555	Mixed sex	ZnF_C4 (HmmSmart), Nuclear_Rec_DBD2(ProfileScan), Nuclear_REC_DBD1 (ScanRegExp), Zf-c4 (HmmPfam), Hormone receptor (HmmPfam), Q9I8T3_AMBME_Q9I8T3(BlastPRoDom), STROIDFINGER (FPrintScan), Nuclear Hormone Receptor (HmmPanther)	0.517	0.334	1926
Smp_174320	666	Mixed sex	Lozenge 1 (HmmPanther), Runt-related (HmmPanther), Runt (Acute myeloid leukaemia 1 protein (AML 1) (HmmPanther, HmmPfam, FPrintScan), p53 and RUNT-type transcription factor, DNA-binding (InterPro)	1.090	0.286	2109
Smp_168810	621	Mixed sex	ETS_DOMAIN_1(ScanRegExp), ETS_DOMAIN_2(ScanRegExp), ETS(HmmSmart , HmmPfam, HmmPanther), ETSDOMAIN(FPrintScan), SAM_PNT(HmmPfam), GaBP Alpha (HmmPanther), Sterile Alpha Motif-type (InterPro)	0.595	0.281	2109
Smp_173470	732	Egg	Homeodomain-related (Gene3D), TRF2-INTERACTING TELOMERIC RAP1 PROTEIN (HmmPanther)	0.860	0.287	4524
Smp_180860	507	Adult female	STROIDFINGER (FPrintScan), VitaminDr (FPrintScan), Nuclear_Rec_DBD_1(ScanRegExp), Zinc finger, NHR/GATA-type (InterPro), Nuclear hormone receptor family member nhr-41(HmmPanther), Zf-C4(HmmPfam), Hormone Receptor(HmmPfam), Znf-C4(HmmSmart),Zinc finger, nuclear hormone receptor-type(InterPro)	0.6567	0.297	4485

Table 22. Properties of the six genes of interest. The domains predicted by InterPro and the individual databases within InterPro are shown in brackets under the "Domains in Proteins" column. Sequences relate to *S. mansoni* proteins.

Analyses of Smp_097730

Smp_097730 is most likely the schistosome orthologue of the serum response factor (SRF) seen in other organisms such as mouse, zebrafish, rat, chicken and *Drosophila* (*E*-value <1 x 10^{-25} for all in BLAST). Due to the lack of orthologous sequences to Smp_097730, which may be due to the sequence being more evolutionarily more diverged, the species tree does not agree with the phylogenetic tree shown in **Figure 39**.

The serum response factor contains the MADS (**M**CM1, **A**gamous, **D**eficiens, and **S**erum response factor) box domain. This conserved DNA-binding domain is required for DNA binding, dimerization and interaction with accessory factors [220] and is also found in other transcription factors. Across the eukaryotic kingdom, there are more than 100 MADS-domain sequences known and most of these play important roles in developmental processes. For instance, in *Drosophila*, the *SRF* is crucial to the formation and maintenance of the trachea [220, 221]. In mice and chicken, *SRF* is expressed in the early developmental stages of the heart and during myogenesis (muscle formation) [221] and is required for the activation of muscle-specific genes.

SRF has been shown to bind nuclear receptors, namely retinoic acid and retinoid X receptors (RXR) in mammalian and yeast two-hybrid systems [222]. The transactivation of target genes resulting from this interaction has been seen in cellular proliferation [223] and the control of metabolic pathways such as glucose metabolism [211], among others. In insulin production in the pancreas, SRF is required together with another transcription factor of the ETS family, to bind and regulate the response of the nuclear receptor, LXRB

to glucose [211].

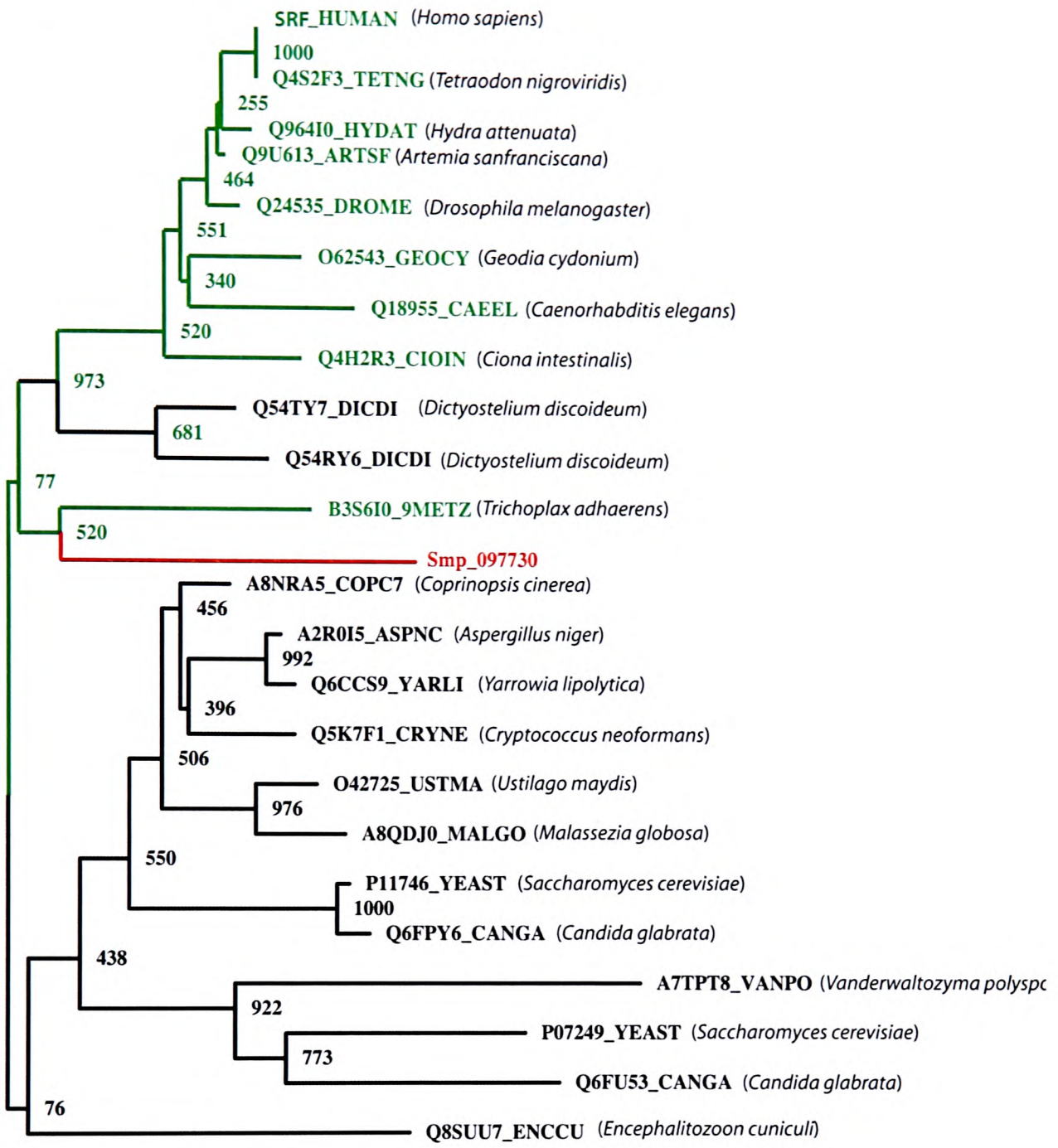


Figure 39. Phylogenetic tree of Smp_097730 (shown in red) built by neighbour joining. Bootstrap values (based on 1000 replicates) are shown on the internal nodes. This phylogenetic tree does not agree with the species tree which may be due to the lack of sequences available. Smp_097730 has orthologues in the metazoans (shown in green). Genes are shown with their UniProt identifications and with their common species name in parentheses. The scale bar represents the average number of amino acid substitutions per site (0.1). The phylogenetic tree was built with the aid of Dr. Luis Sánchez-Pulido.

Analyses of Smp_144170

Smp_144170 has been annotated as a Retinoic Acid Receptor (RAR)-like nuclear receptor by the TIGR/Sanger consortium. This sequence is the alternatively spliced form of the recently published nuclear receptor subfamily 1 (NR1) member for *S. mansoni* (*SmNR1*) [213]. This is a case of alternative splicing in the non-coding region as both variants code for an identical protein. Hence, as there are no publications that discuss Smp_144170, the recent experimental paper on *SmNR1* published by Wu *et al.* [213] provides some insight into its function.

Despite their divergent roles, NRs show a remarkable conservation in their amino acid sequences and different functional regions. The distinct functional and structural domains give rise to a common structural organization. Of these domains, the C and E regions are evolutionarily conserved and the A/B, D and F (if present) regions are more divergent [212]. The *S. mansoni* NR1 (*SmNR1*) gene structure consists of eight exons spanning 14kb with the splice donor and acceptor sites conforming to the GT-AG rule and contains A/B, C, D and E domains (**Figure 40**). SmNR1 is expressed throughout development but it may have a more significant role in egg development, secondary sporocysts in 30-day infected snails and 21-day schistosomula due to the higher mRNA expression observed for these stages [213]. From the phylogenetic tree generated by Wu *et al.* [213], SmNR1 appears to be an orthologue of other mammalian NRs, for instance, LXR, PXR, PPARs, RORs, Reverbs and RARs (**Figure 41**). I have now putatively identified its likely orthologue in *S. japonicum*. SmNR1 alone has been shown to enhance the transcription of mammalian cells by transactivation [213]. SmNR1 regulates transcription as a

heterodimer with the schistosome retinoid X receptor (RXR), another nuclear receptor. Together, SmNR1/SmRXR1 was also shown to interact with mammalian coactivators of transcription to drive transcription in a host mammalian cell [213]. Hence, it has been suggested that *S. mansoni* coactivators of transcription may have a similar interaction with SmNR1/SmRXR1 to drive transcription of target schistosome genes [213].

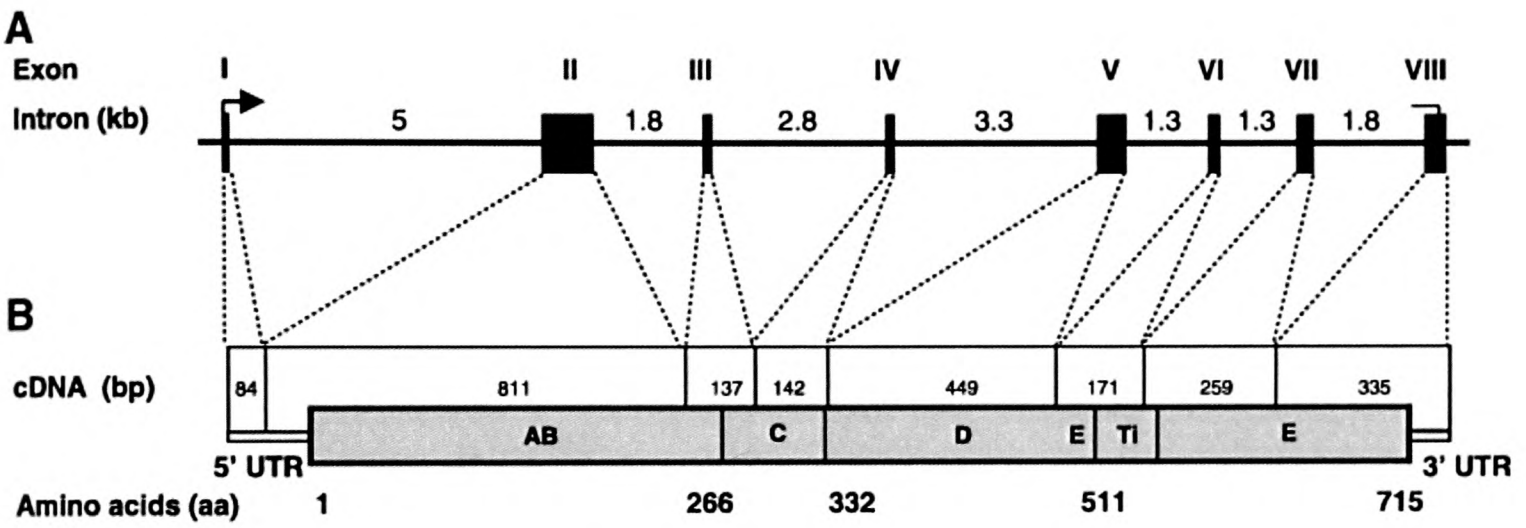


Figure 40. Gene structure of SmNR1. (A) Exons and intron sizes are shown in kilobases (kb). Roman numerals indicate exons. (B) The corresponding cDNA and the 716 amino acid protein sequence. Domains within the protein are: the N-terminal A/B domain, C domain which is the DNA-binding domain, D domain functions as the flexible hinge between C and E domains. The E domain is the ligand-binding domain that ligands and coactivators bind to, to drive transcription activity of target genes. Figure taken from [213].

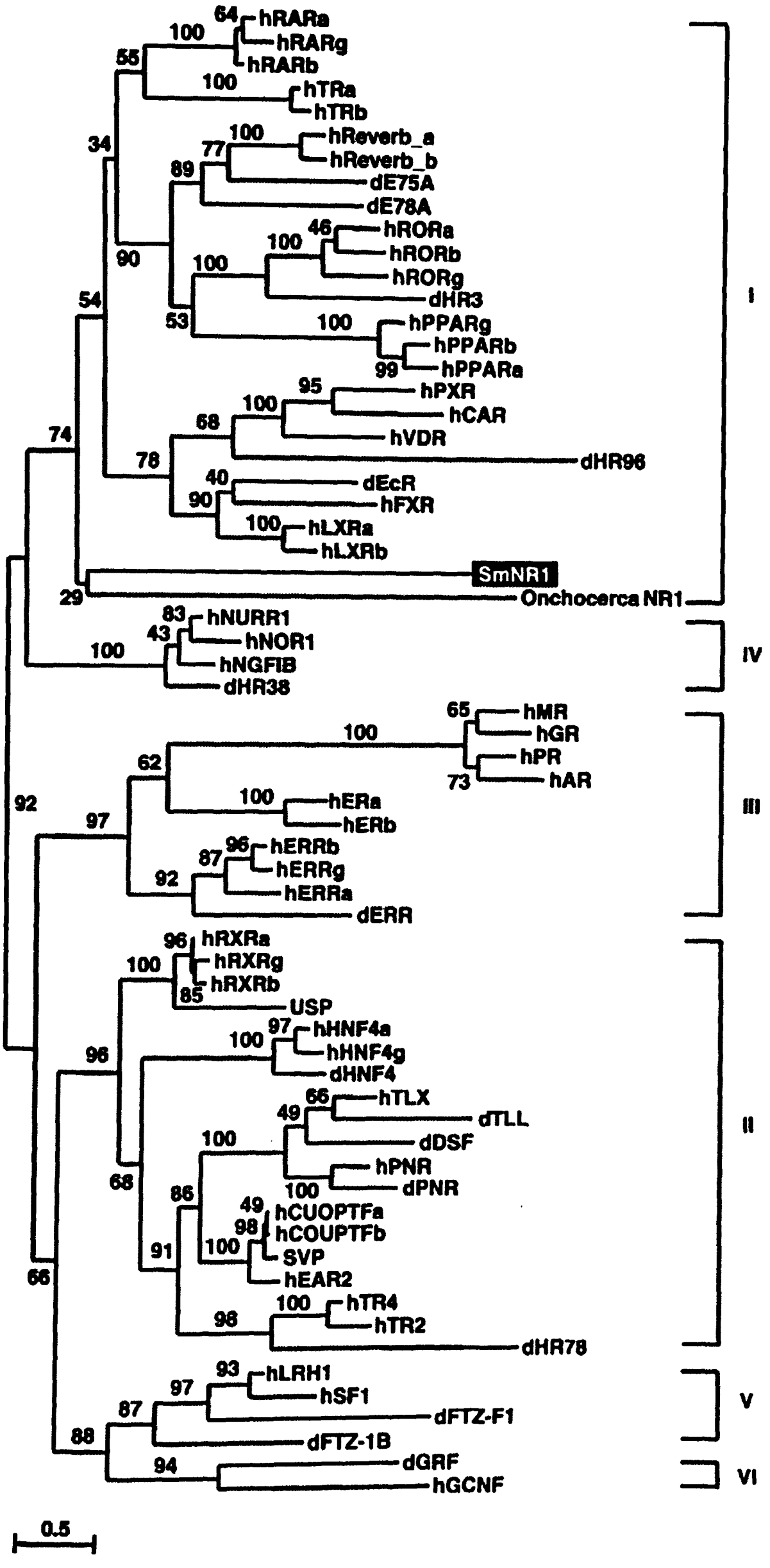

hRARa
hRARg
hRARb
hTRa
hTRb
hReverb_a
hReverb_b
dE75A
dE78A
hRORa
hRORb
hRORg
dHR3
hPPARg
hPPARb
hPPARa
hPXR
hCAR
hVDR
dHR96
dEcR
hFXR
hLXRa
hLXRb
SmNR1
Onchocerca NR1
hNURR1
hNOR1
hNGFIB
dHR38
hMR
hGR
hPR
hAR
hERa
hERb
hERRb
hERRg
hERRa
dERR
hRXRa
hRXRg
hRXRb
USP
hHNF4a
hHNF4g
dHNF4
hTLX
dTLL
dDSF
hPNR
dPNR
hCUOPTFa
hCOUPTFb
SVP
hEAR2
hTR4
hTR2
dHR78
hLRH1
hSF1
dFTZ-F1
dFTZ-1B
dGRF
hGCNF
I
IV
III
II
V
VI
0.5

Figure 41. Phylogenetic tree of SmNR1. This tree was generated by the maximum likelihood model and bootstrap values (based on 100 replicates) are shown on the internal nodes. SmNR1 is a divergent member of the NR subfamily 1. Taken from [213].

Analyses of Smp_174320

Smp_174320 was annotated by the TIGR/Sanger consortium as the orthologue of the *Drosophila* lozenge protein. A query against the non-redundant GenBank protein database at NCBI [104] shows that Smp_174320 has high sequence similarity to the *Drosophila* lozenge protein (*E*-value of 3×10^{-42}). This orthologous relationship can be seen in the phylogenetic tree which was built using conserved RUNT domains among all sequences in **Figure 42** as well as the p53 family that showed structural homology to the RUNT domain. The homology of the RUNT domain and the p53 family was obtained through "Pfam clans". Briefly, a Pfam clan is a collection of homologous Pfam entries and is useful for prediction of function and structure of large, divergent families [224]. Clans are built manually through primary literature, structure information and profile-profile comparisons.

Lozenge is a member of the RUNT family of transcription factors that regulate the expression of many different transcription factors which in turn may regulate target genes in different cells. Lozenge is involved in the prepatterning events leading to the development of the eye disc in *Drosophila* [225, 226].

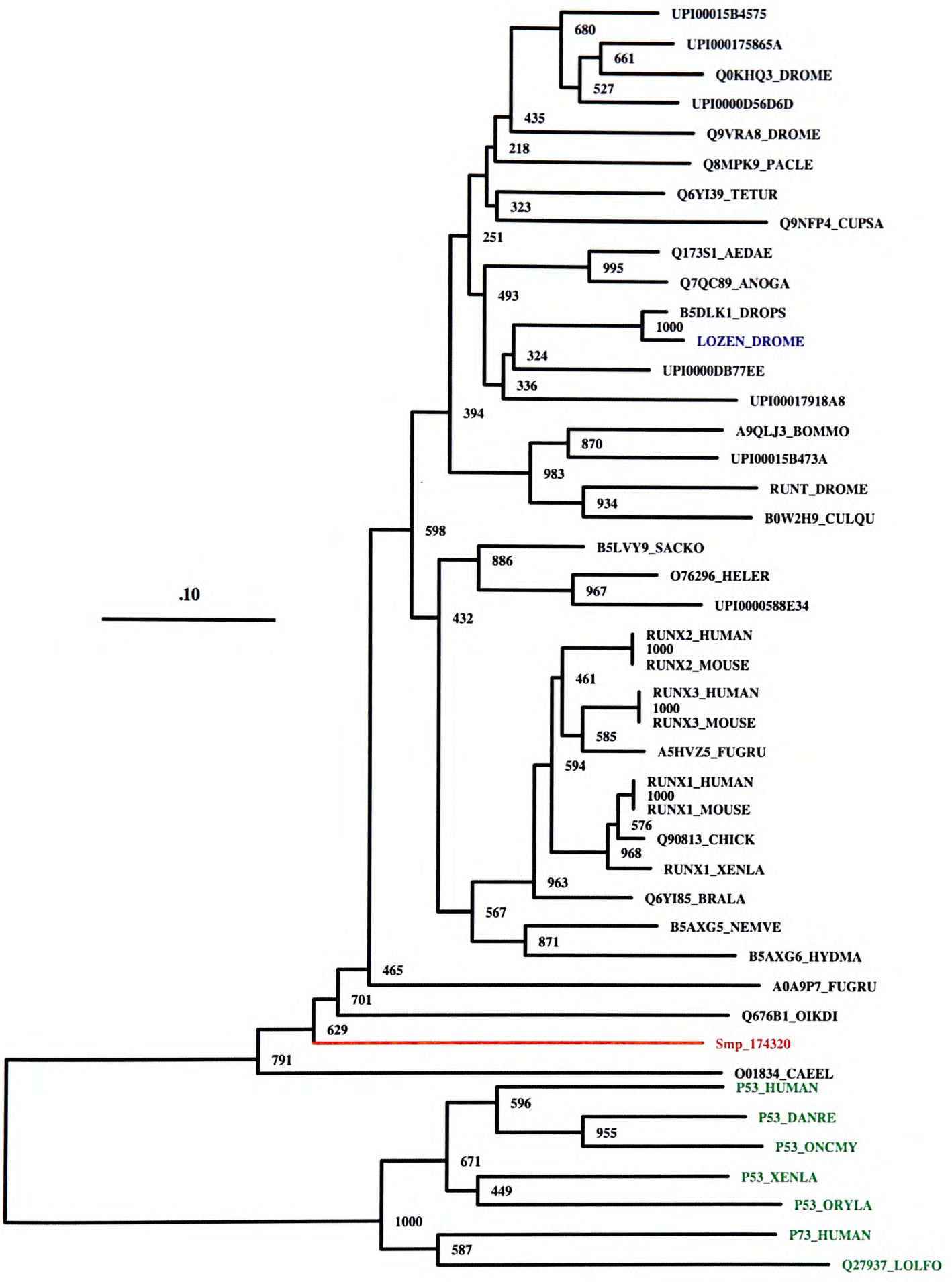


Figure 42. Phylogenetic tree of Smp_174320 generated by the neighbour joining method. Bootstrap values (based on 1000 replicates) are shown on the internal nodes. Smp_174320 (in red) is an orthologue to the *Drosophila* lozenge gene (in blue). The p53 family, highlighted in green is used as an outgroup and was obtained by structural alignment of the p53 and Runt family. Genes are shown with their UniProt identifications and with their

common species name. UPI00015B4575, *Nasonia vitripennis*; UPI000175865A, *Tribolium castaneum*; Drome, *Drosophila melanogaster*; UPI0000D56D6D, *Tribolium castaneum*; Pacle, *Pacifastacus leniusculus*; Tetur, *Tetranychus urticae*; Cupsa, *Cupiennius salei*; Aedae, *Aedes aegypti*; Anoga, *Anopheles gambiae*; Drops, *Drosophila pseudoobscura pseudoobscura*; UPI0000DB77EE, *Apis mellifera*; UPI00017918A8, *Acyrthosiphon pisum*; Bommo, *Bombyx mori*; UPI00015B473A, *Nasonia vitripennis*; Culqu, *Culex quinquefasciatus*; Sacko, *Saccoglossus kowalevskii*; Heler, *Heliocidaris erythrogramma*; UPI0000588E34, *Strongylocentrotus purpuratus*; Human, *Homo sapiens*; Mouse, *Mus musculus*; Fugru, *Fugu rubripes*; Chick, *Gallus gallus*; Xenla, *Xenopus laevis*; Brala, *Branchiostoma lanceolatum*; Nemve, *Nematostella vectensis*; Hydma, *Hydra magnipapillata*; Oikdi, *Oikopleura dioica*; Caeel, *Caenorhabditis elegans*; Danre, *Danio rerio*; Oncmy, *Oncorhynchus mykiss*; Oryla, *Oryzias latipes*; Lolfo, *Loligo forbesi*. The scale bar represents the average number of amino acid substitutions per site (0.1). The phylogenetic tree was built with the aid of Dr. Luis Sánchez-Pulido.

Another protein containing a RUNT domain is the Acute Myeloid Leukaemia 1 (AML-1 or RUNX1) (**Table 22**). AML-1, like Lozenge, also regulates multiple target genes during development. It plays an important role in many steps leading to haematopoietic cell differentiation and proliferation [227]. In the study of human leukaemia, the *AML-1* gene is a frequent target of translocations that results in a fusion protein which has been shown to interact with members of the retinoic acid nuclear receptor family to transcriptionally silence myeloid target genes [228, 229].

Based on these finding, I hypothesise that Smp_174320 is involved in the regulation of different transcription factors which may either transcriptionally activate or repress target genes in different cells during development.

Analyses of Smp_168810

Smp_168810 is a putative GA binding protein (alpha subunit) transcription factor (annotated by the TIGR/Sanger consortium) (**Figure 43**). There are currently no publications on this gene in schistosomes. There are, however, numerous publications of the GA-binding proteins (GABP) of metazoans (for instance [230-233]). As with the other proteins, I queried non-redundant GenBank protein database with Smp_168810 and found the top non-schistosome hit to be the chicken GABP homologue. This homologue matched only about a seventh of the Smp_168810 protein (*E*-value of 2×10^{-43}) at a sequence similarity of 63%. This may be due to the chicken GABP homologue being evolutionarily diverged from its *S. mansoni* homologue. Chicken GABP was postulated to be associated with a housekeeping role in ribosomal synthesis [231].

GABP is a ubiquitously expressed ETS transcription factor that controls gene expression for many different biological processes. ETS factors are a family of evolutionarily related transcription factors that are involved in developmental, carcinogenesis, cellular differentiation and apoptotic functions. ETS factors participate in protein-protein interactions with other transcriptional factors or co-activators in signal transduction pathways [233]. GABP is known to regulate both housekeeping genes as well as tissue-specific genes [231, 233] and its expression is especially abundant in liver, muscle and haematopoietic cells. Of particular interest here is that GABP is a transcriptional regulator of many key hormones and hormone receptors [233].

An example of GABP as a transcriptional regulator is the complex interaction between GABP, Sp1 (another transcriptional activator), and NRs that activates gene expression in response to retinoic acid. In metazoans, CD18 is a β2 leukocyte integrin that is required for white blood cell adhesion to the endothelium and for killing both bacteria and fungi. Myeloid cell expression of CD18 is transcriptionally activated by retinoic acid. However, the responsiveness of CD18 to retinoic acid is mediated by the interaction of GABP and Sp1 through a novel mechanism and the binding of retinoic acid to retinoic acid receptors [234].

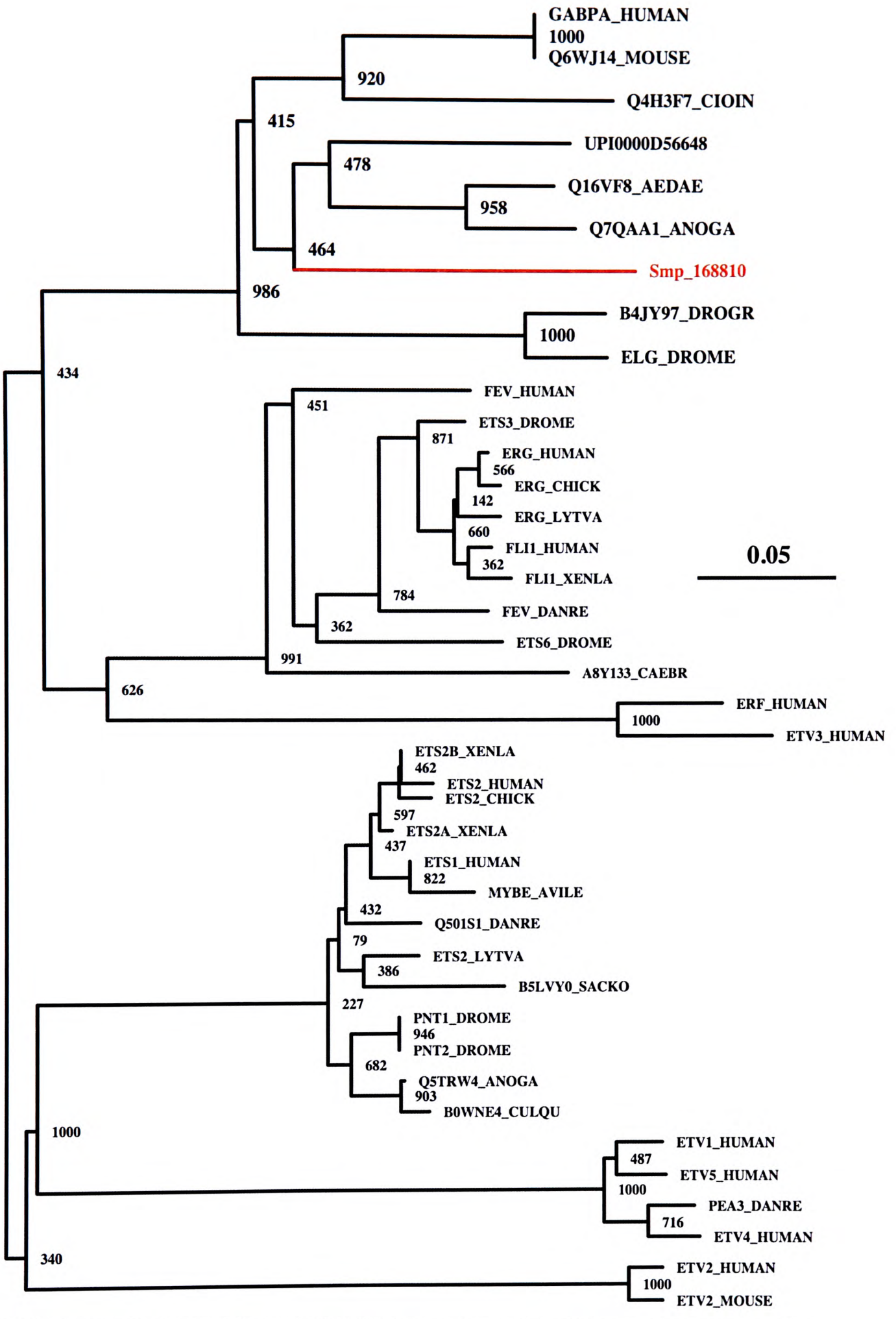


Figure 43. Phylogenetic tree of Smp_168810 (shown in red) built by neighbour joining. Bootstrap values (based on 1000 replicates) are shown on the internal nodes. Genes are shown with their UniProt identifications and with their common species name. Human, *Homo sapiens*; Mouse, *Mus musculus*; Cioin, *Ciona intestinalis*; UPI0000D56648, *Tribolium castaneum*; Aedae, *Aedes aegypti*; Anoga, *Anopheles gambiae*; Drogr, *Drosophila grimshawi*;

Drome, *Drosophila melanogaster*; Lytva, *Lytechinus variegatus*; Xenla, *Xenopus laevis*; Danre, *Danio rerio*; Caebr, *Caenorhabditis briggsae*; Avile, Avian leukemia virus E26; Sacko, *Saccoglossus kowalevskii*; Culqu, *Culex quinquefasciatus*. The scale bar represents the average number of amino acid substitutions per site (0.05). The phylogenetic tree was built with the aid of Dr. Luis Sánchez-Pulido.

Analyses of Smp_173470

Smp_173470 contains a partial hit to the BRD8 human (and mouse) protein (*E*-value of 2 x 10^{-15}) and contains a Rap1 Myb domain whose function is unknown apart from mediating possible DNA or protein interactions. The phylogenetic tree in **Figure 44**, built from conserved regions, shows the relationship between the Smp_173470 and BRD8 of mouse and human.

BRD8 (also known as p120, SMAP1 or SMAP2) was originally identified as a coactivator that enhances transcriptional activation for a thyroid hormone-activated NR [235]. *BRD8* has three alternatively spliced transcripts that encode three different isoforms [236]. Two of these isoforms are part of the Nu4A-like histone acetyltransferase (HAT) complex in humans that plays critical roles in transcriptional regulation by acetylating nucleosomal histones H4 and H2A [237]. The modification of these histones alters nucleosome-DNA interactions and promotes the interaction of proteins with chromatin, thereby regulating transcription. The Nu4 HAT multi-subunit complex was originally found in yeast and is highly conserved in eukaryotes [237]. It is important for positive regulation of transcription, cell cycle control and DNA repair.

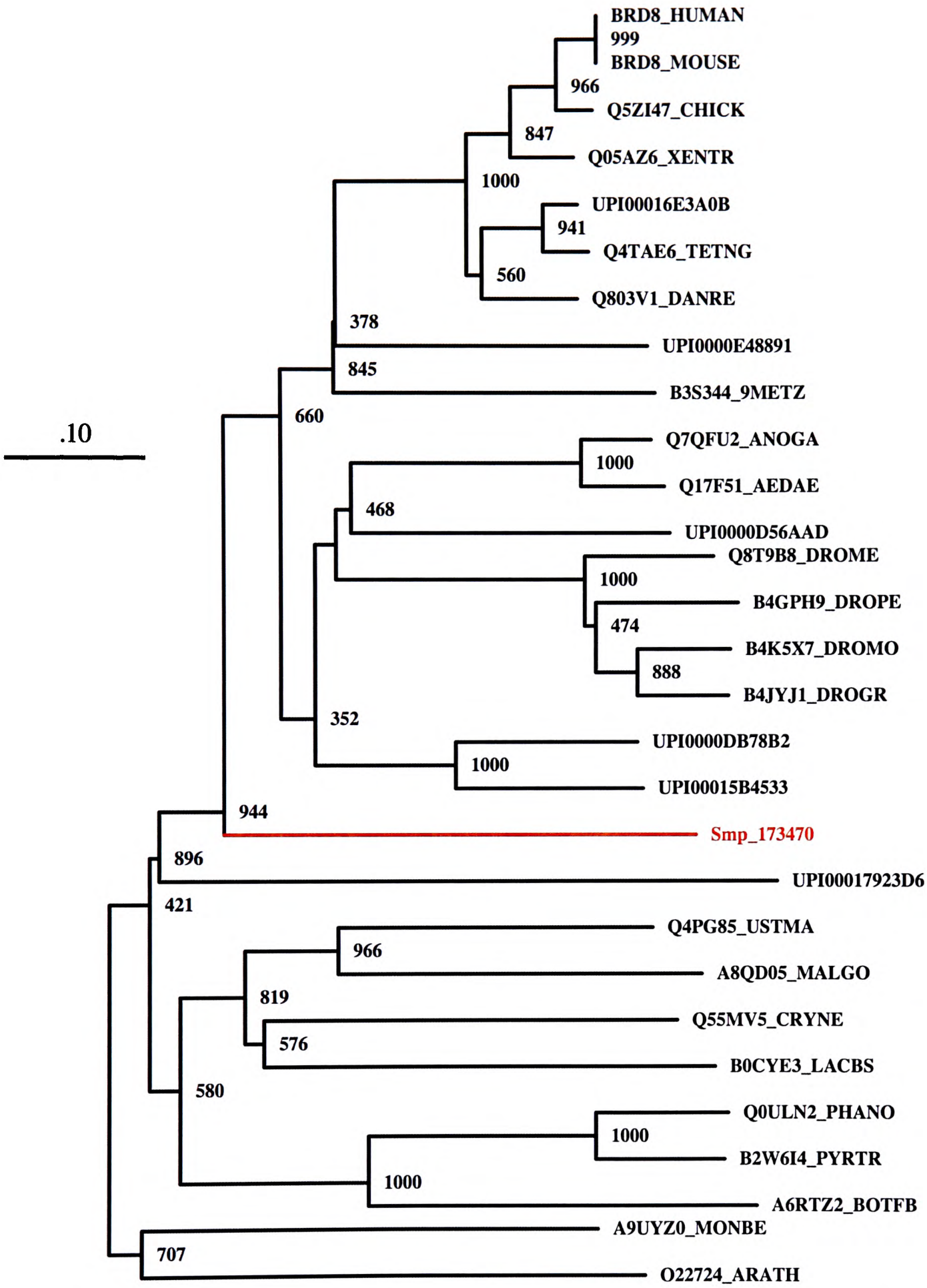


Figure 44. Phylogenetic tree of Smp_173470 (shown in red) built by neighbour joining. Bootstrap values (based on 1000 replicates) are shown on the internal nodes. Genes are shown with their UniProt identifications and with their common species name. Human, *Homo sapiens*; Mouse, *Mus musculus*; Chick, *Gallus gallus*; Xentr, *Xenopus tropicalis*; UPI00016E3A0B, *Fugu rubripes*; Tetng, *Tetraodon nigroviridis*; Danre, *Danio rerio*; UPI0000E48891, *Strongylocentrotus purpuratus*; 9Metz, *Trichoplax adhaerens*; Anoga,

***Anopheles gambiae*; Aedae, *Aedes aegypti*; UPI0000D56AAD, *Tribolium castaneum*; Drome, *Drosophila melanogaster*; Drope, *Drosophila persimilis*; Drogr, *Drosophila grimshawi*; Dromo, *Drosophila mojavensis*; UPI0000DB78B2, *Apis mellifera*; UPI00015B4533, *Nasonia vitripennis*; UPI00017923D6, *Acyrthosiphon pisum*; Ustma, *Ustilago maydis*; Malgo, *Malassezia globosa*; Cryne, *Cryptococcus neoformans*; Lacbs, *Laccaria bicolor*; Phano, *Phaeosphaeria nodorum*; Pyrtr, *Pyrenophora tritici-repentis*; Botfb, *Botryotinia fuckeliana*; Monbe, *Monosiga brevicollis*; Arath, *Arabidopsis thaliana*. The scale bar represents the average number of amino acid substitutions per site (0.1). The phylogenetic tree was built with the aid of Dr. Luis Sánchez-Pulido.**

Analyses of Smp_180860 – an additional gene identified from the previous assembly

Smp_180860 was labelled by the TIGR/Sanger consortium as a "putative nuclear hormone receptor family member nhr-41" (gene prediction version "a"). Indeed, the nematode *Caenorhabditis elegans* nhr-41 transcription factor is a homologue of Smp_180860 [210]: it shows 95% similarity to Smp_180860 (*E*-value of 0.0), labelled as TR2/TR4 orphan nuclear receptor in GenBank.

A recent experimental paper describing *S. mansoni* TR2/TR4, termed SmTR2/4, has been published [210]. The cDNA of this gene was identified from an adult female library. The complete coding sequence for SmTR2/4 translates into a 1,943 amino acid protein sequence. This appears to be the largest nuclear receptor known to date [210]. The *S. mansoni* gene sequence I obtained from the TIGR/Sanger project translates into only 1,495 amino acids which indicates that it may be incomplete. My finding that this gene is "Adult female" stage-specific contrasts that of Hu *et al.* who found the mRNA expression of SmTR2/4 ubiquitous in every stage of the life cycle, with an elevated expression level in cercariae.

Using only the conserved regions (domains) from all sequences in **Figure 41** to build a phylogenetic tree, I show that SmTR2/4 is orthologous to both hTR2 and hTR4. The human testicular receptor 2 (hTR2) and human testicular receptor 4 (hTR4) share 65% sequence similarity which is higher than any similarity detected between two members of the NR family [213]. Here, I only show the subtree showing the position of SmTR2/4 in relation to hTR2 and hTR4 in the NR subfamily II (**Figure 45**).

The *in vivo* physiological roles of TR2, TR4 and SmTR2/4 currently remain unclear. In a preliminary study of the function of SmTR2/4, Hu *et al.* characterised its DNA binding properties and showed that three of its domains exhibited transactivational activity. SmTR2/4 appears to bind to specific target nucleotide sequences both as a monomer and as a homodimer. By this interaction, SmTR2/4 regulates transcription of certain genes. One such candidate gene is suggested by Hu *et al.* to be the female specific *p14* gene, an eggshell precursor gene expressed only in sexually mature females in response to a male stimulus. Further experiments to verify this prediction and to determine other candidate genes are needed.

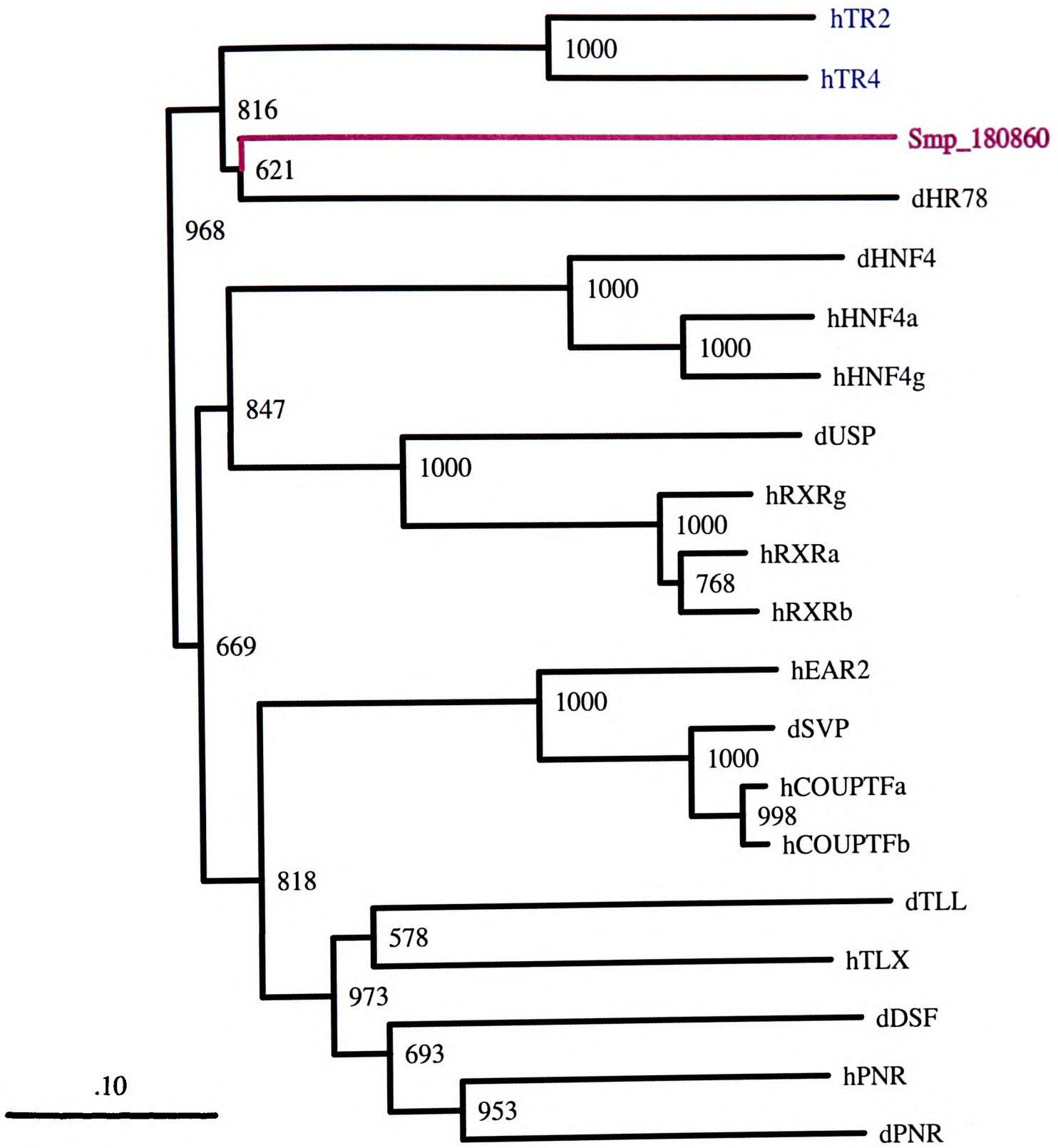


Figure 45. Phylogenetic tree of Smp_180860 generated by the neighbour joining method. Bootstrap values (based on 1000 replicates) are shown at the nodes. Only conserved regions of all the sequences in Figure 41 were used to build this tree. The sub tree of NR subfamily II is shown here with Smp_180860 being an orthologue to hTR2 and hTR4. The prefix "h" denotes human and "d" is for *Drosophila*. The GenBank accession numbers of the genes in the tree are found in [213]. The scale bar represents the average number of amino acid substitutions per site (0.1). The phylogenetic tree was built with the aid of Dr. Luis Sánchez-Pulido.

Genes	Manual annotations from PSI-BLAST	Length of *S. mansoni* protein sequence (bp)	dS	dN/dS	Number of aligned sites (bp)	Stages in which protein is expressed in
Smp_193540.2	An A1.12/9 antigen member which localizes to the neurons, putative sensory receptors and tegument of *S. mansoni* (*E*-value of 1 x 10^{-78})	1452	0.204	1.616	1158	Adult male
Smp_193540.1	An A1.12/9 antigen member which localizes to the neurons, putative sensory receptors and tegument of *S. mansoni* (*E*-value of 1 x 10^{-78})	1494	0.208	1.578	1215	Adult male
Smp_185950	Lineage-specific. Only hit was to *S. japonicum.*	1071	0.196	1.254	564	Schistosomula
Smp_117350	An A1.12/9 antigen member which localizes to the neurons, putative sensory receptors and tegument of *S. mansoni* (*E*-value of 3 x 10^{-77})	801	0.259	1.141	732	Adult male
Smp_109340	Lineage-specific. Only hit was to *S. japonicum.*	966	0.579	0.623	729	Adult male
Smp_121460	No results†	762	0.7239	0.548	561	Schistosomula
Smp_193960	The orthologue of the egg protein C3782 found in *S. japonicum* (*E*-value of 1 x 10^{-78}). Lineage specific.	828	1.316	0.514	483	Egg
Smp_005360	No results†	1113	0.346	0.507	507	Mixed sex
Smp_165100	Lineage-specific. Only hit was to S. *japonicum.*	918	0.514	0.434	552	Schistosomula
Smp_140760	Lineage-specific. Only hit was to *S. japonicum.*	2196	0.687	0.421	549	Adult Male
Smp_148680	Lineage-specific. Only hit was to *S. japonicum.*	1839	0.732	0.410	486	Egg
Smp_167020	Lineage-specific. Only hit was to S. *japonicum.*	1266	0.625	0.389	633	Mixed sex
Smp_184430	Probably a kakapo homologue (*E*-value of 4 x 10^{-4}).	1122	0.522	0.367	537	Adult male
Smp_133950	Lineage-specific. Only hit was to *S. japonicum.*	852	0.778	0.345	480	Mixed sex
Smp_140400	Lineage-specific. Only hit was to *S. japonicum.*	1692	1.139	0.328	459	Adult Male

Smp_038430	Lineage specific. Only hit is to S. *japonicum*.	1788	0.993	0.321	621	Adult male
Smp_164260	Lineage-specific. Only hit was to *S. japonicum*.	1485	1.163	0.316	801	Adult Male
Smp_064200	Lineage-specific. Only hit was to *S. japonicum*.	846	0.731	0.304	456	Adult female
Smp_154830	Probably a "synaptonemal complex protein 2" found in other metazoans. (*E*-value of 2 x 10^{-7}). The synaptonemal complex is a proteinaceous structure that links homologous chromosomes during the prophase of meiosis. The protein encoded by this gene is a major component of the synaptonemal complex and may bind DNA at scaffold attachment regions [104].	3189	0.804	0.301	570	Adult female
Smp_178410	Lineage-specific. Only hit was to *S. japonicum*.	921	0.784	0.296	513	Adult male
Smp_139750	Lineage-specific. Only hit was to *S. japonicum*.	1236	0.994	0.293	522	Mixed sex
Smp_175920	No results†	3135	0.979	0.291	624	Schistosomula
Smp_174240.2	Lineage-specific. Only hit was to *S. japonicum*.	1623	0.817	0.287	1524	Schistosomula
Smp_048080	Similar to "Seryl-aminoacyl-tRNA synthetase 2" found in the purple sea urchin (*Strongylocentrotus purpuratus*) and *Drosophila melanogaster*. Contains these domains: Seryl-tRNA synthetase domain and Class II tRNA amino-acyl synthetase-like catalytic core domain (class_II_aaRS-like_core). (E-value of 0.001)	1338	0.634	0.281	792	Mixed sex
Smp_119390	Lineage-specific. Only hit was to *S. japonicum*.	1134	0.922	0.280	639	Schistosomula

Table 23. Annotations using PSI-BLAST of the top 10% fastest evolving genes that were unannotated by InterPro. dS, dN/dS values and protein lengths of these 25 fast evolving *S. mansoni* genes are provided. Also the number of aligned residues between *S. mansoni* and its *S. japonicum* orthologue are provided. The genes are placed in descending order of their dN/dS values. "†" indicates that there were no hits in the non-redundant GenBank protein database, not even to the sequence itself (*i.e.* this sequence was not present in GenBank in addition to being schistosome-specific).

Annotating genes without InterPro annotations

PSI-BLAST was used to annotate the top 10% most rapidly evolving genes among stage-specific sequences (**Table 23**). For some genes I could not detect homologies in the non-redundant GenBank protein database. Below, I highlight those genes for which homology to other proteins in the GenBank database could be found.

I found three of the eight members of the A1.12/9 antigen family, Smp_193540.1, Smp_193540.2 and Smp_117350, as among the fastest evolving genes in schistosomes. These antigens are encoded for by repetitive DNA and localise to the putative sensory receptors of cercariae, schistosomula and adult worms. These antigens are also associated with membranes in the neurones, sensory receptors of cercariae and also in the tegument of schistosomula 4-21 days post infection [238]. The physiological roles of these antigens have yet to be examined in schistosomes, but it has been suggested through homology to the granin family that these antigens may play a role in host-parasite interactions [238]. An interesting point to note is that these were the only functionally annotated genes that have shown strong evidence of positive selection ($dN/dS > 1$) in this study. Due to the lack of sequence information from other schistosomes, I am unable to perform site-specific analyses to identify sites that have been under positive selection. I can however postulate that these antigens have roles in the defence of the pathogen. This has been seen in other eukaryotes, for example in yeast (**Chapter 3**) and mammalian systems [139], where defence of the organism has been postulated as driving positive selection.

An essential gene for development was found among the fast evolving sequences.

Smp_184430 is predicted to be a homologue of the *Drosophila* kakapo (renamed Shortstop or Shot) that was identified in a screen for wing blister mutants. It is essential for support of the complex architecture of the membrane skeleton, which amongst other cellular roles (for example, in response to external stimuli), maintains the cell shape [239]. In *Drosophila*, kakapo was also found to be essential for adhesion between and within cell layers, for instance adhesion of the epidermal cell layer to the muscles [240].

A seryl-aminoacyl-tRNA synthetase 2 homologue was found in schistosome (Smp_048080) to be fast evolving. This is a nuclear-coded mitochondrial gene that is required for translation [241, 242]. A recent study has shown for the first time the involvement of an aminoacyl-tRNA in an antibiotic biosynthetic pathway [243]. The mouse homologue of this gene has been characterised and is evolutionarily conserved across *S. cerevisiae*, human, bovine, *Neurospora crassa* and *E. coli* [244]. The protein localises to the mitochondrion and is found to be highly expressed in tissues with a high metabolic rate.

Discussion

I have shown that stage-specific sequences tend to have a significantly elevated dN/dS ($p < 0.005$). Correcting for the length biases in dN/dS estimations was achieved by:

1. selecting a threshold on the minimum number of aligned codons in a paired sequence, and by
2. randomly sampling from all aligned codons to create mosaic sequences with 1000 codons.

For the first method, a variety of thresholds on aligned pairs of codons were implemented but the number of sequences remaining fell dramatically as the threshold was increased. Finally, I settled on a minimum of 150 aligned pair of codons between *S. mansoni* and *S. japonicum* on which to determine which stages and GO-slim terms were enriched among the fastest/slowest evolving genes.

The second method of random sampling enabled me to show that compared to non-stage-specific sequences, stage-specific sequences tended to have a significantly higher dN/dS value ($p < 0.01$) (**Figure 35**). Furthermore, I was able to show that the more stage-specific a sequence was, the higher the median dN/dS value tended to become (**Figure 36**). This trend of increasing dN/dS rates with increasing stage-specificity has not been seen in mammalian systems, although mammalian tissue-specific genes have also shown to exhibit an elevated dN/dS rate compared to broadly expressed ones [22, 23]. It is assumed that any mutation that occurs in a broadly expressed protein affects many tissues thereby invariably leading to a reduction in the overall fitness of the organism. Hence, the non-synonymous mutations are selected against, keeping dN/dS rates suppressed.

In schistosomes, stage-specific genes may evolve under greater relaxed constraint or alternatively, they may be under strong selection to survive in the different environments each stage thrives in, leading to the elevation of their dN/dS rates. This is because the life cycle of the schistosome (described in **Chapter 1**) is such that the environments in which the different life forms of the pathogen are found are extremely different, and constantly changing: fresh water –> hepatopancreas -> fresh water -> blood -> organs -> faeces/urine -> fresh water. As such, the pathogens must be able to tolerate the immunological response of the host and adapt to its environment for their survival, development and propagation.

Enrichment within fastest evolving genes

I have studied stage-specific genes to understand which genes are under the most intense positive selection. I found that among the fastest evolving genes, many were found to be from the "Adult Male" stage. These "Adult Male" stage genes are often evolving so rapidly that similarities to other proteins are not detectable. In mammals, genes expressed specifically in liver, testis, kidney and thymus are rapidly evolving [22]. It remains to be seen whether genes expressed in the testes or testicular lobes of the male *S. japonicum* worms may be a contributing factor to the elevated dN/dS rates.

Annotation of fast evolving genes

In addition to having an enrichment of "Adult Male" genes within the fastest evolving genes, I found significantly enriched GO-slim categories of transcription factors. There were six genes that contributed to this GO-slim enrichment.

The metazoan homologues of these six genes play roles in many important regulatory pathways. These include cell differentiation, development and metabolism. Furthermore, the proteins encoded by the six genes have links to transcription regulation and NRs. NRs are increasingly being viewed as attractive drug targets in both metabolic diseases and also in combating schistosomiasis [207].

Currently, studies point towards hormones as the means by which the growth, sexual maturation and development of the helminth parasites occur. These hormonal signals may be from the host and/or the parasite itself. In particular, steroid, thyroid hormones and ecdysteroids have been found to influence the transcription of target genes through the means of nuclear receptors [207].

NRs may be ligand-activated or orphaned (not requiring a ligand) to cause transactivation of target genes. I have found an example of each among my six genes: Smp_144170 is an alternatively spliced form of the recently published *S. mansoni* nuclear receptor subfamily I member (SmNR1) that is ligand-activated. The orphan ligand, Smp_180860 is probably the SmTR2/4, recently discovered by others [213]. This protein has been postulated to regulate the development of the schistosome egg in sexually matured females.

Ligand-dependent interactions between NR and coactivators are required for transcriptional activation to occur. It may be possible that Smp_173470 is a coactivator based on its partial homology to proteins that are part of a multi-protein complex involved in transcriptional regulation.

Of the six fast evolving genes, Smp_097730 is most likely to be the serum response factor (SRF) transcription factor that is implicated in many pathways, including the glucose metabolic pathway. As Smp_097730 is an orthologue of the human SRF transcription factor, the function of Smp_097730 may be inferred from it. Together with a transcription factor of the ETS family, the human SRF interacts with the *LXRB* gene NR in a pathway regulated by glucose [211]. In schistosomes, there is also an ETS transcription factor member (Smp_168810) among the six fast evolving genes that is a GABP orthologue. It would be interesting to experimentally test if there was any link between the Smp_097730 (SRF) and Smp_168810 (GABP) transcription factors and to determine the target gene(s) they activate. Adult *S. mansoni* worms residing in the mesenteric veins of their mammalian hosts have been shown to absorb large amounts of glucose through their teguments by diffusion [245]. Perhaps, in response to the hormone from the host/pathogen, a glucose/hormone/retinoic acid regulated interaction between a NR, SRF-like and GABP-like may exists to transcriptionally activate or repress a target schistosome gene (illustrated in **Figure 46**).

As mentioned above, as the *lozenge* gene in *Drosophila* has been shown to regulate expression of multiple transcription factors, the *lozenge*-like orthologue in *S. mansoni* (Smp_174320) may possess a similar role. It may be a "master" transcription factor that regulates the expression of other transcription factors, which in turn activate transcription of target genes. It will need to be investigated whether these genes act cooperatively to allow the schistosome to react "dynamically" to its host environment. Depending on which host the pathogen infects, hormones of the hypothalamic-pituitary-adrenal axis either increase or decrease in levels during primary infections or

reexposure to the pathogens [246]. For example, one of these hormones is cortisol, which is a glucocorticoid hormone. These hormones regulate gene pathways in response to stress and are critical for adaptation to environmental stressors, for instance in response to injury or infection [247]. As part of the important homeostatic control mechanism, increased levels of cortisol mediate the body's alarm response to stress. This includes suppressing the immune response, thereby allowing the body an adaptive phase for countermeasures [248]. Perhaps it is during this phase that the schistosome uses to transcribe genes which may be necessary for it to enter a different life form or for protection against its environment? As the target schistosome genes of this hormone-induced pathway (**Figure 46**) are yet unknown, the fast evolving genes from **Table 23** may perhaps provide a clue of putative target genes.

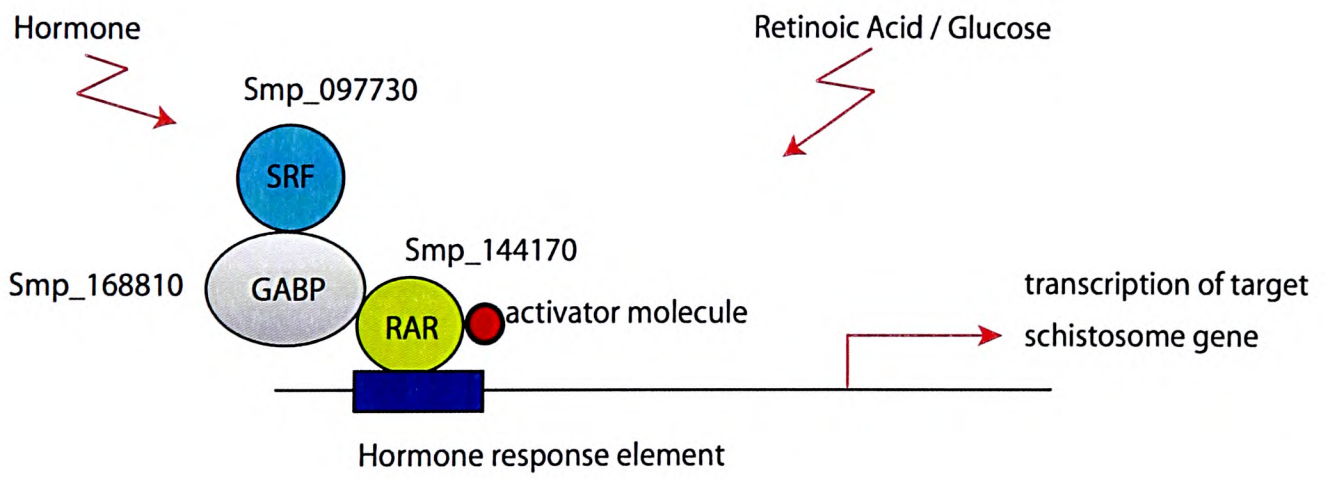


Figure 46. Illustration of a putative signal transduction pathway in which molecules such as the GABP protein Smp_168810, SRF protein Smp_097730 and the RAR-like Smp_144170 may interact. Hormones from either the schistosome or the host trigger the binding of SRF and GABP. This causes the RAR-like molecule to bind to the HRE (Hormone response element) within the promoter. This interaction may involve a transcriptional activator, which, in the presence of retinoic acid or glucose, activates the transcription of the target gene in the schistosome host.

It is noteworthy that many of the fast evolving genes (top 10% of all the stage specific genes, *i.e.* 84 genes) did not have any InterPro or GO annotations assigned to them.

This is likely to be because these sequences evolved so quickly that they are too diverged for similarity to be found among the InterPro databases. I have attempted to annotate 25 genes lacking InterPro annotations and found interesting results. For example, I have found adult male stage-specific antigens encoded by repetitive DNA that show evidence of positive selection. Aided by a double outer membrane (**Figure 3**) and these adaptive tegumental antigens, the schistosomula appear poised to combat their immunologically hostile host environment.

More experimental work is needed to determine the function of these fast evolving genes as it would provide an insight of how adaptive evolution has acted on the schistosome.

Vaccines and potential candidates for future drugs

Several vaccine candidates have been characterised from antigens found on the tegumental surface of the parasite ([249-251] and reviewed in [252]). However, some antigens are known to be differentially expressed in the life stages of the schistosomes and also between adult male and female worms [214, 253].

Secreted proteins and surface proteins expressed in the intramammalian stages have also been suggested as potential sources of vaccines [45]. Hence this study on fast evolving genes may be more than a study on evolution of schistosome genes, but may also help to identify potential vaccine candidates.

Future work

I have shown that stage-specific genes tend to have the highest dN/dS values, indicating that stage-specific genes are fast evolving. I have tried to propose mechanisms for their rapid evolution. Future work on broadly expressed genes that are fast evolving may provide further clues on how the schistosome adapts successfully to its environment. The questions I would like to ask are: *"Are there genes expressed in more than one stage that are evolving as quickly as these stage-specific genes and what are the roles of these sequences?"*

As proof of principle, I have undertaken preliminary analysis of genes expressed in combinations of the following two groups of stages:

Group 1:"Tegument" / "Cercariae" / "Eggshell" and

Group 2:"Adult female" / "Adult male" / "Mixed sex" / "Schistosomula" / "Egg"

I have found Smp_046450 (expressed in both "Adult female" and "Eggshell") and Smp_124280 (expressed in both "Adult Male" and "Eggshell") to be fast evolving, lineage-specific and evolutionarily so diverged that there was no sequence similarity to any other proteins in the GenBank non-redundant database. I did, however, find the fast evolving, lineage specific protein Smp_193960 (expressed in both "Egg" and "Eggshell") to be the orthologue of the egg protein C3782 found in *S. japonicum* (*E*-value of 1×10^{-78}). Further analyses would be required to functionally determine why these three genes were among the top 10% of fast evolving sequences.

It would also be interesting to detect and study interacting genes between host and pathogens. I would first look for fast evolving secretory proteins of the pathogens and functionally annotate these proteins. The outcome of such analyses would ultimately

depend on a lab-based biologist who would be able to experimentally determine the validity of such predictions.

CHAPTER 6: Brain Evolution

Summary

The brain imposes greater constraints on genes expressed in it compared to other organs [22]. In this chapter, I consider whether species differences in constraints on genes reflect their differing brain sizes. I find that animals which have relatively large brains express genes with higher evolutionary rates than in species with relatively small brains. Having relatively large brains is an expensive asset in terms of time (developmental to maturation), energy and anatomical complexity. As with previous studies of adaptive genes along the primate lineage, I might be expected to find genes whose evolution has been accelerated to be preferentially expressed in the cerebral cortex region, more specifically, the neocortex. The neocortex is involved with speech, cognitive and memory abilities that are particular to the higher primates. However, I found no enrichment of such genes in any of 17 brain sub-structures which included cortical and sub-cortical regions. I discuss possible reasons for this outcome in the light of a recent finding on different selective pressures applied between cortical and sub-cortical regions.

Introduction

Genes vary in their evolutionary rates according to the tissues in which they are expressed. Genes expressed in the brain evolve with significantly slower evolutionary

rates compared to those expressed in other tissues [22, 254]. In general, tissue-specific genes have higher evolutionary rates compared to more broadly expressed housekeeping genes [22, 23]. If a gene is expressed in several tissues, it will be subjected to the different constraints imposed by each tissue. These constraints may accumulate and hence genes expressed in many tissues may be subject to more constraints than genes expressed in few tissues [254], resulting in lower evolutionary rates. Both amino acid and expression changes are more constrained in genes expressed in the brain compared to other tissues [254]. However, it appears that there has been a relative acceleration of both amino acid and gene expression changes for brain specific genes along the human lineage compared to the chimpanzee [254]. It is unknown presently as to whether this is due to positive selection or relaxed purification in the human lineage, and whether these changes might be associated with the increase in brain size along this lineage. I take this investigation further to study genes expressed in the different tissues of the brain, and correlate their evolutionary rate changes with brain size (allometric) changes.

Across the animal kingdom, the main changes to the brain during its evolution have been its size. Brain size increases non-linearly with increasing body size. This allometric increase can be described from the regression analysis of log brain size on log body size [48, 255]. This means that as body size increases, brains become "absolutely" larger but "relatively" smaller. For instance, among large mammals, humans have the largest relative brain size of 2% of their body mass. In contrast, shrews, the smallest mammals, have a brain that constitutes 10% of their body mass [50].

The comparison of brain sizes can also be based on the "encephalization" component, estimated by the residual of this regression. Usually expressed as the "encephalization quotient", EQ, it measures the extent the brain size of a given species deviates from the "standard" species of the same taxon. For example, the cat has an EQ of 1 and has been used as the "standard" for mammals. Humans have the highest EQ of 7.4-7.8 which means that the human brain is 7-8 times larger than expected, given its body size [50]. However, there have been many theoretical debates about how EQ should be calculated [48, 49, 256] and how it should be used to compare cognitive abilities or intelligence (for instance, [50, 51, 257]).

Elephants and large cetaceans (for instance, whales and dolphins) exceed humans in cortical volume. Hence, humans thought to be the most intelligent species, neither have the largest brain nor cortex, in both absolute and relative terms [50, 258].

Humans perhaps have the highest information processing capacity due to the thick cortex densely populated with neurons and thick myelinated cortical fibres which aids in the rapid transmission of impulses [50].

In this project, I will be examining the genes from the following seven organisms: human (*Homo sapiens*), mouse (*Mus musculus*), dog (*Canis familiaris*), tenrec (*Echinops telfairi*), opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*) and chicken (*Gallus gallus*). These organisms were chosen because firstly, their genomes were available in the OPTIC pipeline (see below) [259]. Secondly, and more importantly, these organisms are of differing relative brain sizes with which I could use to correlate evolutionary rate changes of brain-expressed genes to their brain sizes. I partitioned the organisms as having "big brains" (dog, human and

mouse) and "small brains" (tenrec, opossum, platypus and chicken) as measured allometrically against their body size (**Figure 47**). Then I addressed the following questions: *With a focus on genes expressed in the brain, does one observe any difference between the evolutionary rates of such genes in "big brains" vs "small brains"? Are the fastest or slowest evolving brain-expressed genes preferentially expressed in certain brain sub-tissues?*

Methods and Materials

Determining large and small brains

To determine brain size relative to body size, I used the allometric equation describing brain weight and body weight (proposed by Snell in the late nineteenth century [48]):

$$E = kP^{\alpha}$$

where E and P are the brain and body weights in grams and k and α are constants. For higher vertebrates, values of k=0.07 and α =2/3 are suggested [48]. As E and P vary over several orders of magnitude, I plot a log-log graph of brain weights and body sizes with the regression line fitting the equation of log(E) = (2/3)log(P) + log(0.07) (**Figure 47**). Dog and man have relatively larger brains than expected. Tenrec, chicken, opossum and platypus have relatively smaller brains than expected. Mouse appears to be only slightly above the line and I initially categorised it as having a large brain. However, as a phylogenetic framework allowing different evolutionary rates on specific lineages to be assigned based on the relative brain sizes of the species was used, mouse was later assigned a variable rate regardless of brain size.

It is noteworthy to point out that all the brain weight and body size measurements used in this chapter are only *estimations* for that particular species. The words "brain/body weights" and "brain/body size" are also used interchangeably. Choosing which individual to measure is not a trivial matter [48, 260]. Not only do measurements differ between genders, but the age, size, state (i.e. determined by method of preservation), and measurement methods [48, 258] also influence values. Also, errors in reporting numbers are propagated through reports which cite erroneous figures without verifying their veracity from original references [48]. Hence, I have tried to obtain original values for the species used in this project from primary sources where possible and assumed these measurements to provide accurate estimations. I have also tried to verify each brain size findings with the literature, if available: for instance, man and dog [48, 50], chicken [58], mouse, platypus, tenrec and opossum [256].

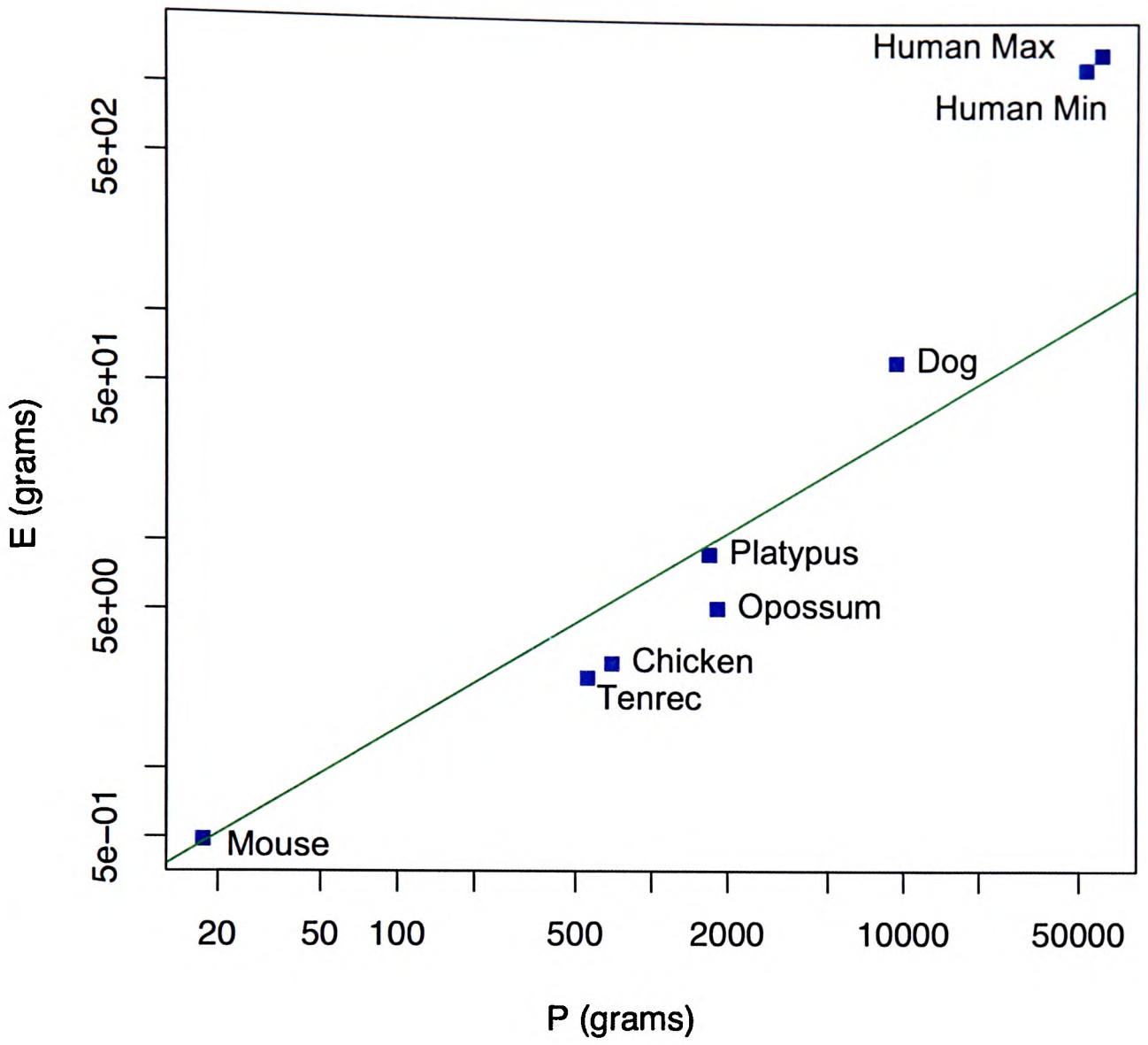


Figure 47. Relationship between brain size (E) and body size (P). Brain size and body weights of the seven species used in this study are plotted on log-log coordinates. For man, the minimum and maximum values are plotted to show a range of body sizes and brain weights. As in all vertebrates, brain size increases allometrically with body size as indicated by the green line. This regression line represents the equation log(E) = (2/3)log(P) + log(0.07) and is the logarithmic form of $E = kP^{\alpha}$, where E and P are the brain and body weight in grams, with k (= 0.07) and α (= 2/3) as constants. The deviation of the seven organisms from this line shows how much their brain size departs from the average higher vertebrate brain:body ratio. Hence, human, dog and possibly mouse have larger brains then expected and the rest have smaller brains than expected. Data for this plot have been obtained from the following sources: [48, 50, 58, 256, 260].

Obtaining sequences for 1:1 orthologues

I obtained 5,483 orthologues from seven organisms – human, dog, tenrec, opossum, chicken, mouse and platypus. These 1:1 orthologues were single copy genes present

in each of the seven organisms. These sequences were derived from the OPTIC [259] pipeline developed by Andreas Heger in our group. In brief, the genes were generated through the following steps:

1. Gene structures were downloaded from Ensembl (http://www.ensembl.org/). The versions of each genome assembly and the corresponding Ensembl gene sets for the seven organisms are summarised in **Table 24**.
2. Pairwise orthology assignments were then carried out using the PhyOP pipeline [261], developed by Leo Goodstadt, also in our group. This starts with an all-against-all pairwise alignment of protein sequences of the genes between all genome pairs. This was carried out using BLASTP [100]. Only alignments with an *E*-value $<10^{-5}$ or covering more than 75% of the shorter sequence are retained. These alignments are than weighted according to a normalised bit score. PhyOP then builds a phylogenetic tree from these alignments and orthology is inferred through congruence with the known species tree.
3. Finally, to obtain multiple orthology assignments of genes from the seven organisms, pairwise orthology assignments are extended. For instance, if gene X in species A and gene Y in species B are each orthologues of gene Z in species C, then it is implied in turn that gene X and Y are also orthologues of each other. This is identical to the orthology assignment method illustrated in **Figure** 14.
4. Alignments are then created to produce a gene tree that is again compared to the known species tree. Based on this species tree, the gene tree is split into orthologous groups.

5. 1:1 orthologue sets containing exactly one gene per species are then extracted.

Species	Assembly	Gene Set
Homo sapiens	NCBI 36 October 2005	Ensembl, August 2006
Mus musculus	NCBI m36 December 2005	Ensembl, April 2006
Canis familiaris	CanFam 2.0, May 2006	Ensembl, December 2006
Monodelphis domestica	monDom5, October 2006	Ensembl, February 2007
Ornithorhynchus anatinus	Ornithorhynchus_anatinus-5.0, December 2005	Ensembl, May 2007
Gallus gallus	WASHUC2, May 2006	Ensembl, August 2006
Echinops telfairi	echTel2, August 2005	Ensembl, August 2006

Table 24. Summary of the versions of each genome assembly and the corresponding Ensembl gene sets for the seven organisms used in this study. This information has been obtained from Andreas Heger (personal communication) and also from the OPTIC website at http://genserv.anat.ox.ac.uk/clades/amniota/statsAssemblies.

Calculating lineage-specific rates

CODEML in the PAML (version 4) package [146] was used to identify evolutionary rates for different lineages of the tree. In the null model, a single rate (ω) is assumed across the entire phylogenetic tree, regardless of brain size. This model is compared with the alternative where the tree is partitioned into lineages representing relatively large brains and lineages representing relatively small brains.

I used two alternative models in this study. The first consisted of the tree being partitioned such that there were two ω values in the tree: one for genes from species with relatively large brains and the other for species with relatively small brains. In this model, I defined large brains on the dog, human and mouse lineages. The rest of

the organisms were labelled as having small brains. The second alternative model consisted of three ω values in the tree. Here, mouse was labelled as having a variable brain size (as discussed above) and had its own value of ω. Dog and human were still labelled as having relatively large brains and the remaining animals (except mouse) as having relatively small brains.

A log-likelihood ratio test was conducted between the null and each alternative model. I tested whether the alternative models of **Figure 48B** or **Figure 48C** better fits the data than the null model in **Figure 48A**. The conservative chi-square distribution was used to determine significance with one degree of freedom (d.f.) for model B (**Figure 48B**) and two d.f. for model C (**Figure 48C**). Three false discovery rates (FDRs) of 5%, 1% and 0.1% were employed to control the number of false positives. I will be referring to **Figure 48** throughout the chapter and also in **Appendix B** to indicate which model is being referred to.

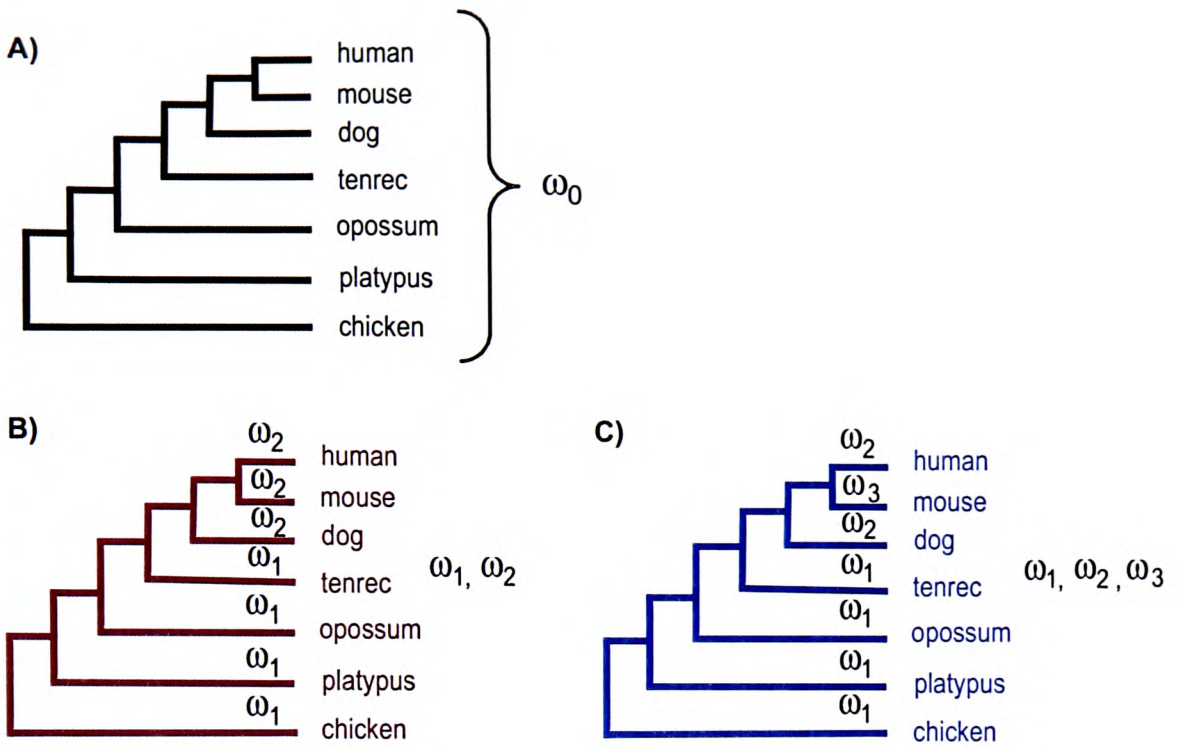


Figure 48. Null (A) and alternative models (B and C). These species trees are unrooted. In the null model, one ω is used across the tree. In the alternative models, two (B) or three (C) ω values were estimated across the tree. This partitioning of the tree allows organisms with relatively large brains to have a different rate from those with relatively small brains. Organisms with relatively small brains are defined by the "ω_1" on the branches of the tree. Relatively large brained organisms are labelled with "ω_2" on their branches. In alternative model C, mouse has its own ω as it is allowed to have a variable brain size, indicated by the "ω_3" (*i.e.* it is neither fixed as having a large or a small brain).

Obtaining brain expressed and brain specific data

The brain expression and brain specific data of genes were obtained from the GNF (Genomics Institute of the Novartis Research Foundation) Gene Expression Atlas [262]. The GNF gene expression measurements were performed on the Affymetrix GeneChip Human Genome U133 Array (HG-U133A) set. This extensive atlas profiles the expression of known, previously uncharacterized and predicted protein-coding genes in both the mouse and human genome. I used the human gene expression data set which provides the mRNA expression levels of approximately 11,000 protein-coding genes across 79 different tissues.

Using the human counterpart from each of the 5,483 1:1 orthologues sets, I found mRNA expression for only 3,454 human genes from within the GNF dataset. From these genes, I defined brain (or regions of the brain) specific genes as those with over four times the median expression level. Brain (or regions of the brain) expressed genes are defined as those with expression (average difference) levels greater than 50. These levels are derived from the average hybridization intensity of matched sequence to single mismatched sequence [263]. This average difference allows presence/absence calls to be made on the array for the removal of data that are not reliably detected and separates expression signal from noise.

The 17 brain-related categories within the GNF data that I used were:

1. Whole brain (adult),
2. Fetal brain,

and the following regions of the adult brain:

3. Amygdala,
4. Cerebellum Peduncles
5. Cingulate Cortex
6. Hypothalamus
7. Medulla Oblongata
8. Occipital Lobe
9. Parietal Lobe
10. Pons
11. Prefrontal Cortex
12. Temporal Lobe
13. Thalamus

14. Caudate Nucleus

15. Cerebellum

16. Globus Pallidus

17. Subthalamic Nucleus

Results

Lineage-specific genes

Of the two alternative models I tested using CODEML, at FDRs of 0.05, 0.01, 0.001, more genes fitted model C in **Figure 48C**. This model with three ω values in the tree more often better fitted the data than the model with two ω values in the tree. The numbers of genes that better fit each model are tabulated in **Table 25**. I show that for each model, the evolutionary rate changes of the genes do correspond to the changes in brain size (**Figure 49** and **Figure 50**). From **Figure 50**, the evolutionary rate change of mouse genes is more similar to that of relatively small-brained animals than it is of relatively larger-brained animals.

Number of significant genes		FDR < 0.05	FDR < 0.01	FDR < 0.001
	Model B	1,649 (30%)	1,077 (20%)	610 (11%)
	Model C	3,185 (58%)	2,269 (41%)	1,437 (26%)
Number of significant genes with GNF data	Model B	1,085 (66%)	722 (67%)	414 (68%)
	Model C	2,022 (63%)	1,484 (65%)	948 (66%)

Table 25. Numbers of genes which fit the alternative models (model B and C in Figure 48B and Figure 48C) better than the null model (model A in Figure 48A) as estimated by CODEML. Percentages for the number of significant genes are given out of the 5,483 1:1 orthologues. Percentages for the number of significant genes with GNF data are given out of the number of significant genes. For instance, in Model B at the FDR < 0.05, 66% is calculated by (1,085/1,649) * 100. "Model B" represents the model with two ω values in the species tree and "Model C" represents the model with three ω values in the tree (from Figure 48). Three different FDR values are used to show the large number of genes fitting the alternative models. In summary, due to more genes being significant, model C appears to fit the data better more frequently than model B.

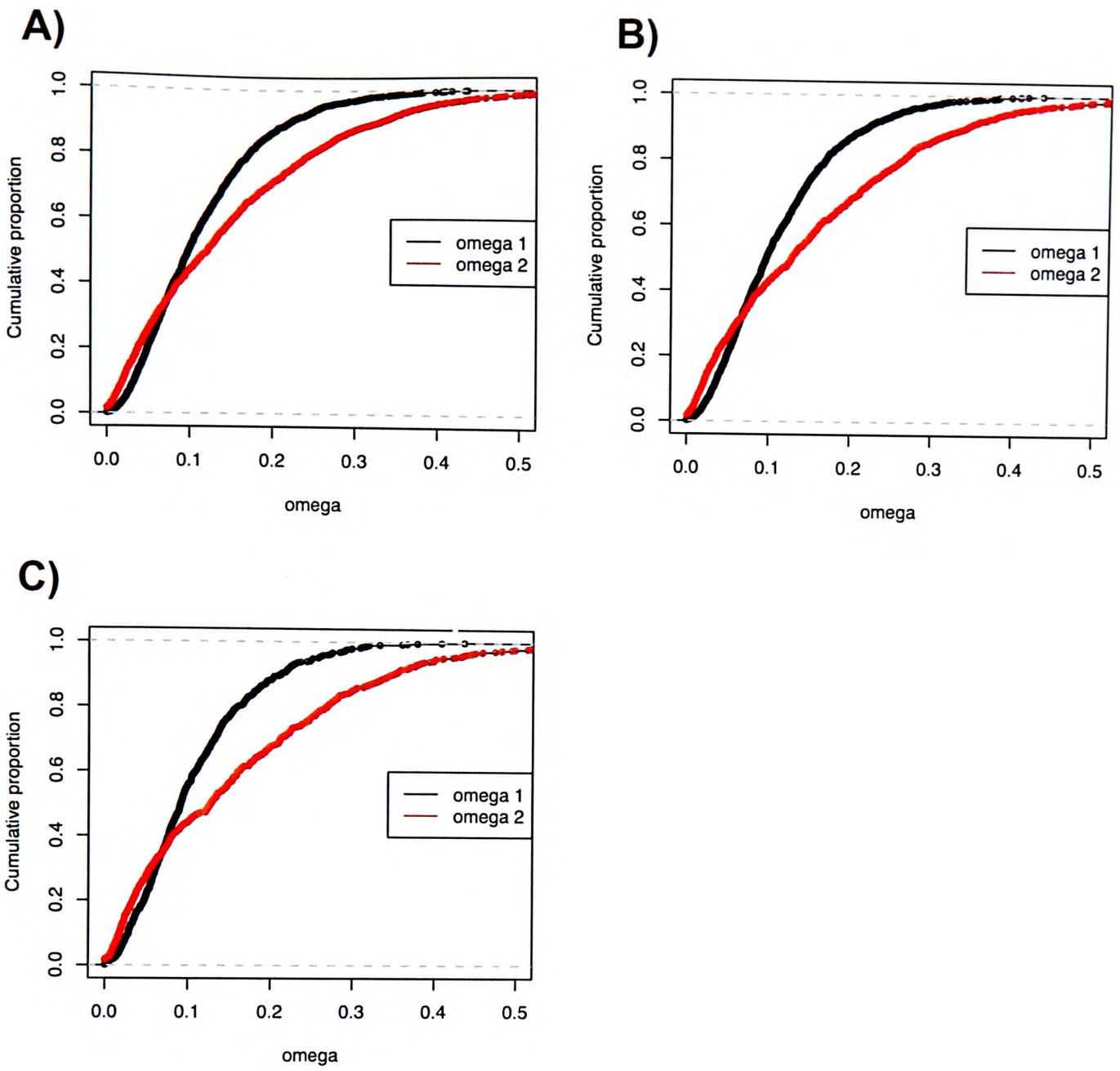


Figure 49. Cumulative frequency distribution plots of ω (omega) values predicted by CODEML. This is for the alternative model B (from Figure 48) which had two ω values in the species tree. Genes significant at each of the three FDR values – 0.05, 0.01 and 0.001 are shown as plots A, B and C respectively. The black and red lines represent the distribution of ω_1 and ω_2 values (refer to Figure 48) respectively. The plots show that the distributions of ω_1 and ω_2 differ. This indicatess that the evolutionary rates of genes on the lineages designated as having relatively large brains (ω_2) are different from those having relatively small brains (ω_1).

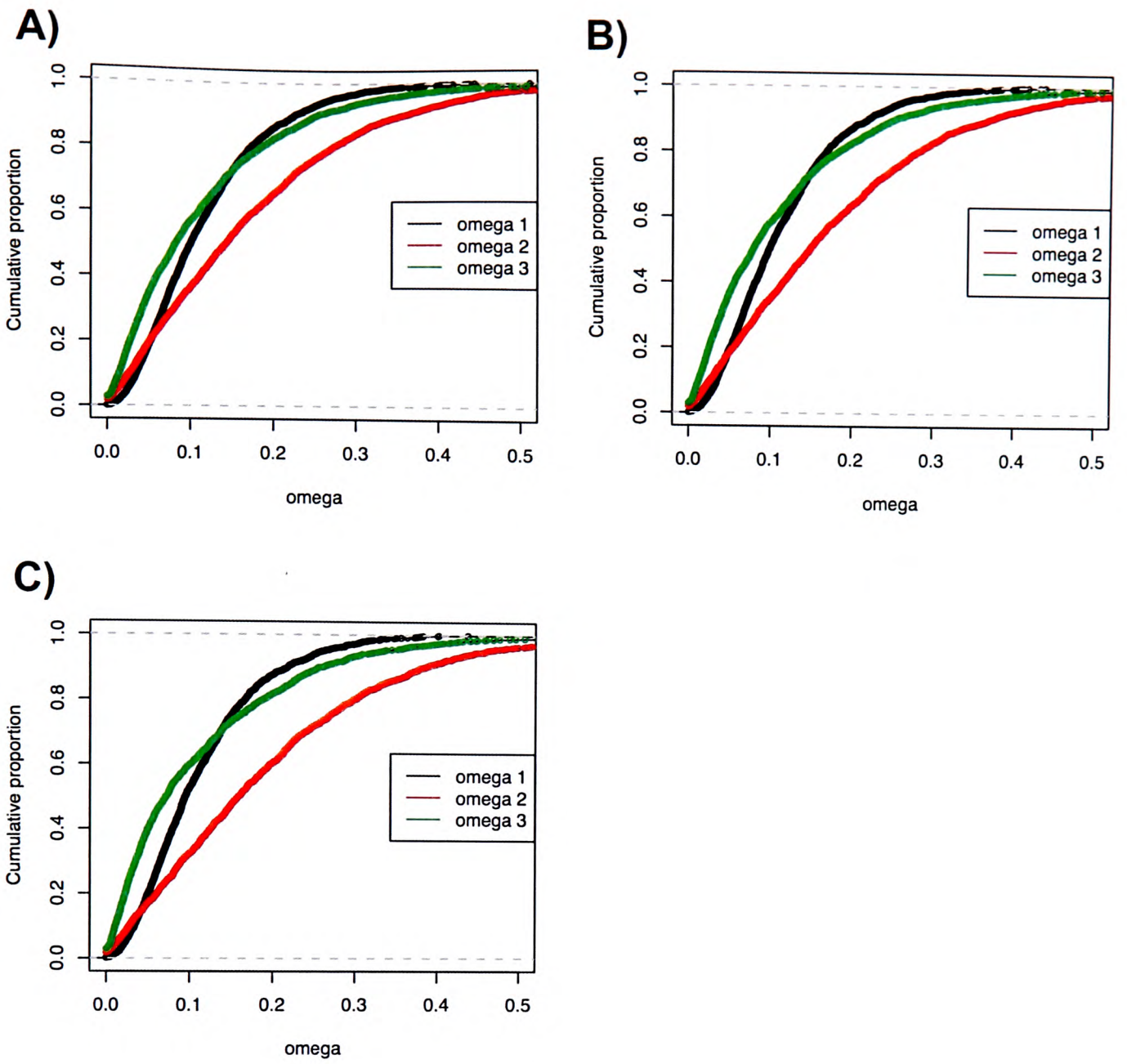


Figure 50. Cumulative frequency distribution of ω (omega) values predicted by CODEML in the alternative model C (from Figure 48) which had three ω values in the species tree. Genes significant at each of the three FDR values – 0.05, 0.01 and 0.001 are shown as plots A, B and C respectively. The black, red and green lines represent distributions of ω_1, ω_2 and ω_3 values (refer to Figure 48) respectively. From the graphs, the distribution of ω_3 appears to be more similar to that of ω_1 indicating that evolutionary rate change seen in mouse is more similar to that of relatively small-brained animals than it is of relatively larger-brained animals (*i.e.* human and dog).

Preferential expression

In both **Figure 49** and **Figure 50**, the evolutionary rates of genes corresponding to the relatively large-brained animals appear to be more elevated than the rates of relatively small-brain animals (i.e. $\omega_2 > \omega_1$). To investigate this difference, I first looked for brain expressed/specific genes (with $\omega_2 > \omega_1$) in relatively large-brained animals and

then tested if these (relatively) faster evolving genes were preferentially expressed in certain regions of the brain.

I used human to represent the relatively large-brained animals as their mRNA expression data for brain tissues were available from the GNF gene expression atlas. **Table 25** summarises the gene count for both models B and C (**Figure 48B** and **C**) tested at different FDR values. Of the 5,483 1:1 orthologues of the seven species, only 3,454 had GNF expression data. **Table 25** tabulates the numbers and percentages of the significant genes which had GNF data at each of the three FDR values.

Genes significant at each of the three FDR values were tested for enrichment in different brain regions using Fisher's exact test. For each of these three FDRs, results of the number of genes specific to (and/or expressed in) each of the 17 brain categories defined in the GNF data are tabulated in **Appendix B**. None of the 17 categories contained a significant enrichment ($p < 0.05$) for fast evolving genes (see **Appendix B** for the breakdown of genes tested in each category).

Discussion

As discussed in **Chapter 1**, most animals have small brains because having large brains is costly both developmentally and metabolically. This may contribute to the finding that genes expressed in the brain tend to be more under greater constraints than other tissues [254]. I wanted to extend this finding to determine if the evolutionary rate differences between brain expressed/specific genes corresponded to the allometric change in brain sizes. To this end, I partitioned seven organisms

(human, mouse, dog, tenrec, chicken, opossum and platypus) into those having large and small brains based on a simple allometry equation describing body size to brain weights.

Rejection of the null hypothesis

I found that there were indeed differences between gene evolutionary rates that segregated according to species' brain size categories. There was a significant tendency for gene rates to be higher in species with relatively larger brains than in species with relatively smaller brains. This effect could be explained in terms of species' effective population sizes (N_e): species with large N_e (e.g. mouse) have a lower evolutionary rate compared to species with small N_e (e.g. human) [264]. The N_e of mouse (*M. musculus*) is ~5–8 $\times 10^5$ [265]) compared to the small N_e of human (*H. sapiens*) of ~10^4 [266]. To date, humans have the smallest N_e of all the vertebrates genomes which have been sequenced [267].

Why small N_e values are associated with higher evolutionary rates

As selection is less effective in small populations, many more neutral or slightly deleterious mutations are fixed in a species with small N_e by random drift [92]. This causes an increase in the non-synonymous substitution rates of species with small N_e compared to those with large N_e values. The ratio of non-synonymous to synonymous substitution rates (ω) thus is expected to be higher for a species with small N_e [268].

Smaller brain size on the mouse lineage

In the instance where mouse was allowed to have a variable brain size (model C in **Figure 48**), the data for more genes were shown to be significantly in favour of

mouse being labelled as relatively small-brained rather than relatively big-brained. It is noteworthy to point out that as shown in **Figure 47**, mouse appeared to have larger than average brain size, but depending on which species of mouse is used in the calculations, mouse can be found to have relatively smaller than expected brains [50, 258]. Furthermore, the allometric relationship between body weight and brain size to determine large and small brains provides only an estimation of expected size and as discussed above, is thus subjective. The finding that data for mouse (*Mus musculus*) are more consistent with those for relatively small-brained animals implies that in the evolution of small brains to large brains, there was a change in relative brain size leading to relatively smaller brains on the mouse lineage.

Adaptive evolution has been predicted to occur along the primate lineage in genes such as the *Abnormal spindle-like microcephaly associated (ASPM)* [269] and *Microcephalin (MCPH1)* gene [141] which are associated with microcephaly, *ADCYAP1 (*adenylate-cyclase-activating polypeptide 1) which is involved in neural precursor proliferation, and *AHI1 (*Abelson helper integration site 1), an axon guidance gene to name but a few (see [270] for a full review of the different genes and their roles). Some of these genes are known be expressed outside of the brain (e.g. *ASPM* [271] and *MCPH1* [272], and yet mutations in these genes result in notable effects in the brain or in only certain regions of the brain. For instance, positive selection in the *ASPM* gene has been linked to major changes of the cerebral cortex but not to major changes in the whole brain or cerebellum size [273]. Different parts of the brain have independent evolutionary size changes and functional constraints [47] and it is to this end that I examined if the genes showing higher

evolutionary rates in relatively large-brained animals than in relatively small-brained animals are preferentially expressed in certain brain regions.

As the neocortex is the region which has contributed to the evolution of enhanced cognitive abilities of the higher primates and has also enlarged disproportionally during mammalian evolution, I had expected any fast evolving genes to be preferentially expressed in the cortical regions, more specifically, the neocortex. Thus I looked for the preferential expression of these fast evolving genes in adult brain (as a whole) and its sub-structures and also in (whole) fetal brain by application of Fisher's exact test. There was no significant enrichment of such genes in any of the 17 GNF brain categories analysed.

I would like to discuss my results in the light of a recent finding by Tuller *et al.* [274] who showed that there were different constraints within the brain. Genes expressed in the cortical regions were under stronger selective pressure than sub-cortically expressed genes [274]. Tuller *et al.* went on to show that genes highly expressed in the cortex tended to be highly expressed in sub-cortical regions, thus increasing the functional constraints on their evolutionary rates. Thus despite the evolutionary differences between genes expressed in relatively large and relatively small brains, perhaps these genes were so highly conserved and widely expressed that I was unable to detect a brain region in which the fast evolving genes were preferentially expressed. Perhaps also that there are no adaptive changes in genes expressed in these regions or perhaps adaptive changes are mostly located in regulatory non-coding sequences.

Future work

Further work on this chapter should include an analysis which precludes the role of any one species (for instance, human) accelerating the evolutionary rate within any group (relatively large/small brains) in the phylogenetic framework discussed above. This could be achieved by allowing each node to model ω_2 exclusively and to compare the results iteratively. This would enable one to identify the lineage on which accelerated (or decelerated) evolutionary rates was occurring. An alternative solution would be to recursively apply the three parameter model ω_3, ω_2 and ω_1 discussed above across the entire tree. This would allow every node to have the chance of having a variable evolutionary rate. In addition, the evolutionary rate of genes on each lineage may thus be identified independently and a comparison of all the rates across the phylogenetic tree be more accurately compared.

On a different note, this study was carried out with the GNF expression data which only describe mouse and human transcriptomes. To date, there are no available brain expression data of such an extensive nature as the GNF data for all other organisms used here. As such, I used the human transcriptome expression data to represent the relatively large-brained animals. However, these data may not have been truly representative of genes expressed in dog (and/or mouse). This project may thus be better suited to be applied to the primate lineage (large brains) alongside with mouse as the small-brained counterpart.

CHAPTER 7: Discussion

Codons that have accumulated more amino acid changing substitutions than expected from the rate of unselected substitutions are said to be under positive selection. Such codons are of interest because they indicate changes in the genetic code that may be beneficial for the organism. Such changes are preferentially passed on to subsequent generations and may eventually become fixed in the population. Identifying such sites allows a deeper understanding of molecular and cellular functions that provide the cutting edge of adaptation [144]. Currently, changes at the whole organism level have contributed to genomic signatures of positive selection [139, 275]; human-based examples are the recent acquisitions of speech and an enlarged brain [138, 141]. Yet it is the codon level that provides the most clues regarding which protein sites are most, or least, susceptible to adaptation. The effects of these evolutionary changes are seen in the altered molecular, and finally, cellular and organismal functions.

In this thesis, I have examined the evolutionary rate changes across a diverse group of organisms. The main findings were:

- Positive selection was detected among genes which are involved in the survival and defence of the organism. This was discussed in **Chapter 3** where yeast genes were enriched in functions that were crucial in times of stress. This was also observed in experiments outlined in **Chapter 5** that showed gene enrichment among codons for antigens, which may aid schistosomes in surviving in their hostile host environments.

- Both highly and lowly expressed yeast genes were under constraints at both their synonymous *and* non-synonymous sites (**Chapter 4**). I showed that constraints on the protein sequence (*i.e.* constraint on non-synonymous sites) occurred to perhaps ensure the correct folding of the resulting protein to produce the required tertiary structure.
- Stage-specific genes have higher evolutionary rates than broadly expressed genes (**Chapter 5**). I also showed that in schistosomes, the more stage-specific a gene was, the more elevated its dN/dS rate tended to be.
- Fast-evolving, stage-specific schistosome genes were enriched in nuclear receptor genes. These nuclear receptors may be part of pathways used by these pathogens in response to changes to the host environment (**Chapter 5**).
- Evolutionary rate changes of brain expressed genes correlated with the changes to the allometric change in brain size (**Chapter 6**). Relatively large-brained animals exhibit elevated evolutionary rates of brain expressed/specific genes compared to relatively small-brained animals. These faster evolving genes were not preferentially expressed in any of the brain-substructures I examined.

The diversity of the projects undertaken here has meant that an understanding of the diverse data sets and the biological contexts wherein they were studied had to be acquired. Although the projects differ greatly in the use of large-scale biological data sets, similar methodological approaches have been used to analyse the underlying evolutionary mechanisms. In particular, I have looked for accelerated evolutionary rates of sites to determine the types and intensity of selective forces that have operated on the protein sequence and its codons.

Accelerated evolutionary rate has been analyzed and linked to many factors, such as expression level of genes, the dispensability and essentiality of a gene, gene duplication, and the number of protein-protein interactions of the gene's protein (summarized in a recent review [276]). One factor that has attracted controversy is the age of genes and its effect on the evolutionary rate [277, 278]. It has been argued that there is an inverse relationship between the age of genes and the evolutionary rate for human-mouse orthologous protein pairs [277]. A different interpretation however sees this as an artefact due to the difficulty in detecting protein orthologues of fast evolving genes in distantly related species [278]. Most of the fast evolving genes studied in this thesis have evolved so quickly that they are too diverged for similarities to be found among other proteins and/or their domains (**Chapters 3** and **5**). I have attempted to ascribe functions to these proteins based on either domain homology or gene orthology. From this, the effects of adaptive evolution can be understood in a molecular and functional context.

Understanding the selective pressures which suppress non-synonymous substitution rates has also been important. As much as adaptive changes can occur to increase the fitness of an organism, restricting change and purifying mutations ensures the proper functioning of pathways and systems in the organism. In **Chapter 4**, I show these constraints acting at the protein level and in **Chapter 6**, these constraints are seen to act on an entire organ, namely the brain.

In summary, these studies have all contributed in their own way to understanding how evolution through selection or mutation has had a part to play in every organism, regardless of size or position in the food chain.

Future work using evolutionary studies

Possible future work for the projects described in this thesis has been discussed at the end of each relevant chapter. However, in the light of these studies, I would like to highlight recently published studies and also possible projects that one might undertake using evolutionary rate analyses as a tool.

As pointed out in the future work section of **Chapter 6**, the brain evolution project I undertook would have been better suited to a primate lineage study. However, the methods and concepts used in that chapter are applicable to other studies. A recent study on schizophrenia shows how studies on positive selection can be used to study human brain disease [279]. This interesting study focused on the primate brain's prefrontal cortex, which is associated with cognitive abilities, and genes involved in brain metabolism. The results showed that brain metabolism is significantly altered in schizophrenic patients and that these genes may be under positive selection on the human lineage. It would be interesting to further apply evolutionary rate studies on other human diseases, for instance, learning disability. Work done in our group has shown that large copy number variations in the genome have been significantly linked to learning disability. It would be interesting to study if certain regions were under positive selection on the human lineage. This could be done by comparing the extent of linkage disequilibrium (LD) by using the publicly available nucleotide polymorphism data. Positive selection is known to increase LD around the selected variant, hence this indirect method can be used to test for positive selection between pairs of genes from different haplotypes [280].

Evolutionary rate studies are not restricted only to coding sequences. Studies done in our group showed that 1.3-4% of functional material was located outside of human protein coding sequences [174]. Of this, about 1 MB (0.03% of the human genome) has evolved adaptively compared to the chimpanzee and this number has been suggested to be an under-estimation [174]. The amount of positive selection on such non-coding sequence appears to differ between species. For instance, positive selection on introns and intergenic regions has been proposed to have contributed to the estimated 20% of divergence between *Drosophila melanogaster* and *D. simulans* [136]. In another study performed by our group, accelerated evolutionary rates were detected as significantly abundant among introns in human brain expressed genes (Lunter and Ponting – in preparation). The role of selection on these non-coding sequences has yet to be elucidated, but as more work is conducted on non-coding DNA, evolutionary rate studies will become more necessary to uncover the remaining fingerprints of selection.

LIST OF ABBREVIATIONS

ω	omega; dN/dS; K_A/K_S; ratio of non-synonymous to synonymous substitutions rates
κ	Kappa; transition/transversion ratio
ω^+	Positively selected site
ω_1	ω of relatively small-brained animals
ω_2	ω of relatively large-brained animals
ω_3	Variable ω
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
c1000.s1000	1000 sequences containing 1000 codons
CAI	Codon Adaptation Index
cDNA	Complementary DNA
dN	K_A; number of non-synonymous substitutions per non-synonymous site
DNA	Deoxyribonucleic acid
dS	K_S; number of synonymous substitutions per synonymous site
EST	Expressed sequence Tag
FDR	False Discovery Rate
GNF	Genomics Institute of the Novartis Research Foundation
GO	Gene Ontology
GO-slim	Gene
kb	Kilobase
K_A	dN
K_S	dS
Mb	Megabase
MIT	Massachusetts Institute of Technology
ML	Maximum Likelihood
NCBI	The National Center for Biotechnology Information
N_e	Effective population size
NR	Nuclear Receptor
ORF	Open reading frame
PAML	Phylogenetic Analysis by Maximum Likelihood

PERL	Practical Extraction and Report Language
PSI-BLAST	Position-specific iterative BLAST
SGD	Saccharomyces Genome Database
SLR	Sitewise Likelihood Ratio
tRNA	Transfer RNA (Ribonucleic acid)
WashU	Washington University in St. Louis

APPENDIX A

The residue solvent accessibility (RSA) and solubility indicator for each of the sites in the protein YBR052C (**Chapter 2**) are shown. These scores are partitioned into three states as defined previously by Rost and Sander [153]: buried (<9% relative accessibility), intermediate (9%–35% relative accessibility), and exposed (≥36% relative accessibility).

Site number	RSA	Relative Solubility
4	78.7	>35%
5	4.5	<9%
6	0	<9%
7	0	<9%
8	0	<9%
9	0	<9%
10	0.9	<9%
11	9	(9-35)%
12	16.2	(9-35)%
13	58.5	>35%
14	88	>35%
15	42.9	>35%
16	25.5	(9-35)%
17	1.4	<9%
18	16	(9-35)%
19	39.2	>35%
20	0	<9%
21	0	<9%
22	44	>35%
23	10.1	(9-35)%
24	0	<9%
25	7.5	<9%
26	47.6	>35%
27	0	<9%
28	0	<9%

29	59.3	>35%
30	49.5	>35%
31	15.5	(9-35)%
32	83.3	>35%
33	18.3	(9-35)%
34	51.5	>35%
35	23.6	(9-35)%
36	29	(9-35)%
37	6.3	<9%
38	21.8	(9-35)%
39	10.6	(9-35)%
40	2.8	<9%
41	42.6	>35%
42	70.8	>35%
43	44.4	>35%
44	91.5	>35%
45	135.2	>35%
58	119	>35%
59	26.4	(9-35)%
60	73.2	>35%
61	60.8	>35%
62	9.4	(9-35)%
63	50	>35%
64	45.1	>35%
65	50	>35%
66	4.3	<9%
67	0.6	<9%
68	38.5	>35%
69	59.2	>35%
70	2.2	<9%
71	27.4	(9-35)%
72	0	<9%
73	1.2	<9%
74	0	<9%
75	0	<9%
76	0	<9%
77	0	<9%
78	14	(9-35)%
79	16.9	(9-35)%
80	41.1	>35%
81	109.6	>35%
82	21.4	(9-35)%
83	62.4	>35%

84	11.7	(9-35)%
85	0.9	<9%
86	56.2	>35%
87	4.4	<9%
88	0	<9%
89	42	>35%
90	70.3	>35%
91	14.7	(9-35)%
92	9.1	(9-35)%
93	60.7	>35%
94	53.6	>35%
95	4.2	<9%
96	66.9	>35%
97	54.6	>35%
98	1.8	<9%
99	66.1	>35%
100	78.7	>35%
101	76.1	>35%
102	33.3	(9-35)%
103	16.2	(9-35)%
104	2.4	<9%
105	62	>35%
106	41.7	>35%
107	22	(9-35)%
108	1.5	<9%
109	0	<9%
110	0	<9%
111	0	<9%
112	0	<9%
113	0	<9%
114	7.7	<9%
115	5.6	<9%
116	57.5	>35%
117	47.7	>35%
118	64	>35%
119	88	>35%
120	33.3	(9-35)%
121	28.6	(9-35)%
122	5.6	<9%
123	45.9	>35%
124	54.2	>35%
125	0	<9%
126	0	<9%

127	56.7	>35%
128	20.8	(9-35)%
129	0	<9%
130	13.4	(9-35)%
131	60.4	>35%
132	18.4	(9-35)%
133	1.8	<9%
134	62.2	>35%
135	76.6	>35%
136	9.8	(9-35)%
137	36.9	>35%
138	4.3	<9%
139	52.4	>35%
140	39.4	>35%
141	14.6	(9-35)%
142	14.3	(9-35)%
143	13	(9-35)%
144	46.3	>35%
145	124.3	>35%
156	52.4	>35%
157	31	(9-35)%
158	11.3	(9-35)%
159	58.8	>35%
160	44.6	>35%
161	0	<9%
162	1.9	<9%
163	2.3	<9%
164	2.2	<9%
165	27.4	(9-35)%
166	7.5	<9%
167	15.5	(9-35)%
168	100.9	>35%
169	71.8	>35%
170	51.2	>35%
171	84.4	>35%
172	35.1	>35%
173	71.5	>35%
174	15.9	(9-35)%
175	26.4	(9-35)%
176	73.7	>35%
177	69	>35%
178	2.6	<9%
179	26.8	(9-35)%

180	58.5	>35%
181	30.5	(9-35)%
182	0	<9%
183	24.2	(9-35)%
184	39.6	>35%
185	0	<9%
186	0	<9%
187	31.7	(9-35)%
188	35.9	>35%
189	0	<9%
190	0	<9%
191	43.3	>35%
192	24.6	(9-35)%
193	5.7	<9%
194	40.5	>35%
195	83.9	>35%
196	57.9	>35%

APPENDIX B

This appendix provides the results of the Fisher's exact test that was used to determine if fast-evolving genes corresponding to the relatively large brained-animals were preferentially expressed in sub-structures of the brain (**Chapter 6**). The test was performed at three different false discovery rates (FDR) of 5%, 1% and 0.1%. Model B and Model C are the two alternative models tested against the null model (model A) (please refer to **Figure 48** in **Chapter 6**).

The numbers in the table are the gene counts that were significant at that FDR. Fast (evolving genes) are defined as having an evolutionary rate (ω) of

$$\omega_2 - \omega_1 > 0$$

and slow (evolving genes) are defined as having rates which are

$$\omega_1 - \omega_2 > 0$$

The definitions of brain and non-brain are given under each of the GNF categories tested.

None of the GNF categories was significantly enriched with the fast-evolving genes and hence the results of the Fisher's Exact test of "Not significant" (*i.e.* $p > 0.05$).

Adult whole brain

In both tables, "brain" indicates adult whole brain. "Non-brain" indicates the other significant genes are not of the "adult whole brain" category.

Brain expressed genes

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	574	397	395	241	234	128
	Non-Brain	71	43	55	31	31	21
	Not significant						
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	1229	579	917	408	602	235
	Non-Brain	150	63	113	45	79	32
	Not significant						

Adult whole brain specific genes

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	19	18	16	11	11	3
	Non-Brain	626	422	434	261	254	146
	Not significant						
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	45	29	36	20	29	8
	Non-Brain	1334	613	994	433	652	259
	Not significant						

Fetal adult brain

In both tables, "brain" indicates fetal whole brain. "Non-brain" indicates the other significant genes are not of the "fetal whole brain" category.

Brain expressed

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	574	397	395	241	234	128
	Non-Brain	71	43	55	31	31	21
	Not significant						
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	1229	579	917	408	602	235
	Non-Brain	150	63	113	45	79	32
	Not significant						

Brain specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	21	18	17	11	11	3
	Non-Brain	624	422	433	261	254	146
	Not significant						
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	48	29	37	20	27	7
	Non-Brain	1331	613	993	433	654	260
	Not significant						

Sub-brain tissue specific

In each of the ensuing tables, "brain" indicates genes containing the tissue-in-question-specific GNF data. "Non-brain" indicates the other significant genes which do not belong to the particular tissue category.

Amygdala specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	5	2	4	1	3	1
	Non-Brain	640	438	446	271	262	148
	Not significant						
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	8	5	7	2	5	1
	Non-Brain	1371	637	1023	451	676	266
	Not significant						

Cerebellum Peduncles specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	2	1	1	1	1	1
	Non-Brain	643	439	449	271	264	148
	Not significant						
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	12	3	7	1	3	1
	Non-Brain	1367	639	1023	452	678	266
	Not significant						

Cingulate Cortex specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	1	1	1	1	0	1
	Non-Brain	644	439	449	271	265	148
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	3	3	2	2	2	2
	Non-Brain	1376	639	1028	451	679	265
	Not significant			Not significant		Not significant	

Hypothalamus specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	3	4	3	2	1	1
	Non-Brain	642	436	447	270	264	148
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	9	6	6	3	5	1
	Non-Brain	1370	636	1024	450	676	266
	Not significant			Not significant		Not significant	

Medulla Oblongata specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	0	0	0	0	0	0
	Non-Brain	645	440	450	272	265	149
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	0	1	0	0	0	0
	Non-Brain	1379	641	1030	453	681	267
	Not significant			Not significant		Not significant	

Occipital Lobe specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	0	0	0	0	0	0
	Non-Brain	645	440	450	272	265	149
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	3	0	2	0	1	0
	Non-Brain	1376	642	1028	453	680	267
	Not significant			Not significant		Not significant	

Parietal Lobe specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	2	3	1	1	0	0
	Non-Brain	643	437	449	271	265	149
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	10	3	9	3	6	1
	Non-Brain	1369	639	1021	450	675	266
	Not significant			Not significant		Not significant	

Pons specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	6	3	6	2	2	1
	Non-Brain	639	437	444	270	263	148
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	10	5	9	3	7	2
	Non-Brain	1369	637	1021	450	674	265
	Not significant			Not significant		Not significant	

Prefrontal Cortex specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	3	2	3	2	1	2
	Non-Brain	642	438	447	270	264	147
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	10	3	8	2	4	2
	Non-Brain	1369	639	1022	451	677	265
	Not significant			Not significant		Not significant	

Temporal Lobe specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	6	5	6	5	4	4
	Non-Brain	639	435	444	267	261	145
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	15	6	13	5	11	4
	Non-Brain	1364	636	1017	448	670	263
	Not significant			Not significant		Not significant	

Thalamus specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	7	4	6	2	2	1
	Non-Brain	638	436	444	270	263	148
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	15	4	12	3	9	1
	Non-Brain	1364	638	1018	450	672	266
	Not significant			Not significant		Not significant	

Caudate Nucleus specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	4	1	2	1	2	0
	Non-Brain	641	439	448	271	263	149
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	5	3	5	1	2	1
	Non-Brain	1374	639	1025	452	679	266
	Not significant			Not significant		Not significant	

Cerebellum specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	3	4	2	4	2	4
	Non-Brain	642	436	448	268	263	145
	Not significant		Not significant		Not significant		
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	7	4	3	4	3	4
	Non-Brain	1372	638	1027	449	678	263
	Not significant		Not significant		Not significant		

Globus Pallidus specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	5	5	3	3	1	3
	Non-Brain	640	435	447	269	264	146
	Not significant		Not significant		Not significant		
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	16	12	12	9	7	6
	Non-Brain	1363	630	1018	444	674	261
	Not significant		Not significant		Not significant		

Subthalamic Nucleus specific

	FDR < 0.05		FDR < 0.01		FDR < 0.001	
Model B		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	9	7	8	5	3	4
	Non-Brain	636	433	442	267	262	145
	Not significant			Not significant		Not significant	
Model C		Fast	Slow	Fast	Slow	Fast	Slow
	Brain	19	8	15	6	10	4
	Non-Brain	1360	634	1015	447	671	263
	Not significant			Not significant		Not significant	

References

1. Goffeau, A., et al., *Life with 6000 genes*. Science, 1996. **274**(5287): p. 546, 563-7.
2. Mewes, H.W., et al., *Overview of the yeast genome*. Nature, 1997. **387**(6632 Suppl): p. 7-65.
3. Meyers, R.A., *Encyclopedia of molecular cell biology and molecular medicine*. 2nd ed. 2004, Weinheim Chichester: Wiley-VCH ; John Wiley.
4. Landry, C.R., et al., *Ecological and evolutionary genomics of Saccharomyces cerevisiae*. Mol Ecol, 2006. **15**(3): p. 575-91.
5. Zeyl, C. and G. Bell, *The advantage of sex in evolving yeast populations*. Nature, 1997. **388**(6641): p. 465-8.
6. Wood, V., et al., *The genome sequence of Schizosaccharomyces pombe*. Nature, 2002. **415**(6874): p. 871-80.
7. Sherman, F. *An Introduction to the Genetics and Molecular Biology of the Yeast Saccharomyces cerevisiae*. 2000 [cited; Available from: http://dbb.urmc.rochester.edu/labs/sherman_f/yeast/Index.html.
8. Roemer, T., et al., *The Spa2-related protein, Sph1p, is important for polarized growth in yeast*. J Cell Sci, 1998. **111 (Pt 4)**: p. 479-94.
9. Pruyne, D. and A. Bretscher, *Polarization of cell growth in yeast*. J Cell Sci, 2000. **113 (Pt 4)**: p. 571-85.
10. Pardo, M., et al., *A proteomic approach for the study of Saccharomyces cerevisiae cell wall biogenesis*. Electrophoresis, 2000. **21**(16): p. 3396-410.
11. Morey, N.J., C.N. Greene, and S. Jinks-Robertson, *Genetic analysis of transcription-associated mutation in Saccharomyces cerevisiae*. Genetics, 2000. **154**(1): p. 109-20.
12. Madden, K. and M. Snyder, *Cell polarity and morphogenesis in budding yeast*. Annu Rev Microbiol, 1998. **52**: p. 687-744.
13. Levin, D.E., *Cell wall integrity signaling in Saccharomyces cerevisiae*. Microbiol Mol Biol Rev, 2005. **69**(2): p. 262-91.
14. Kurtzman, C.P. and C.J. Robnett, *Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses*. FEMS Yeast Res, 2003. **3**(4): p. 417-32.
15. Klis, F.M., et al., *Dynamics of cell wall structure in Saccharomyces cerevisiae*. FEMS Microbiol Rev, 2002. **26**(3): p. 239-56.
16. Herman, P.K., *Stationary phase in yeast*. Curr Opin Microbiol, 2002. **5**(6): p. 602-7.
17. Colman-Lerner, A., T.E. Chin, and R. Brent, *Yeast Cbk1 and Mob2 activate daughter-specific genetic programs to induce asymmetric cell fates*. Cell, 2001. **107**(6): p. 739-50.
18. Chenevert, J., *Cell polarization directed by extracellular cues in yeast*. Mol Biol Cell, 1994. **5**(11): p. 1169-75.
19. Young, K.H., *Yeast two-hybrid: so many interactions, (in) so little time*. Biol Reprod, 1998. **58**(2): p. 302-11.
20. Kellis, M., et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements*. Nature, 2003. **423**(6937): p. 241-54.
21. Kuma, K., N. Iwabe, and T. Miyata, *Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes*

demonstrated by protein kinase and immunoglobulin supergene families. Mol Biol Evol, 1995. **12**(1): p. 123-30.
22. Winter, E.E., L. Goodstadt, and C.P. Ponting, *Elevated rates of protein secretion, evolution, and disease among tissue-specific genes*. Genome Res, 2004. **14**(1): p. 54-61.
23. Duret, L. and D. Mouchiroud, *Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate*. Mol Biol Evol, 2000. **17**(1): p. 68-74.
24. Dekker, J., *GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p*. Genome Biol, 2007. **8**(6): p. R116.
25. Seoighe, C. and K.H. Wolfe, *Yeast genome evolution in the post-genome era*. Curr Opin Microbiol, 1999. **2**(5): p. 548-54.
26. Eyre-Walker, A. and L.D. Hurst, *The evolution of isochores*. Nat Rev Genet, 2001. **2**(7): p. 549-55.
27. Kimura, M., *The neutral theory of molecular evolution: a review of recent evidence*. Jpn J Genet, 1991. **66**(4): p. 367-86.
28. SGD. 2008 [cited; Available from: http://www.yeastgenome.org/cache/genomeSnapshot.html.
29. Jones, T., et al., *The diploid genome sequence of Candida albicans*. Proc Natl Acad Sci U S A, 2004. **101**(19): p. 7329-34.
30. van het Hoog, M., et al., *Assembly of the Candida albicans genome into sixteen supercontigs aligned on the eight chromosomes*. Genome Biol, 2007. **8**(4): p. R52.
31. Goffeau, A., *Four years of post-genomic life with 6,000 yeast genes*. FEBS Lett, 2000. **480**(1): p. 37-41.
32. Cliften, P.F., et al., *Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis*. Genome Res, 2001. **11**(7): p. 1175-86.
33. Casaregola, S., et al., *Genomic exploration of the hemiascomycetous yeasts: 7. Saccharomyces servazzii*. FEBS Lett, 2000. **487**(1): p. 47-51.
34. Bon, E., et al., *Genomic exploration of the hemiascomycetous yeasts: 6. Saccharomyces exiguus*. FEBS Lett, 2000. **487**(1): p. 42-6.
35. Moller, K., L. Olsson, and J. Piskur, *Ability for anaerobic growth is not sufficient for development of the petite phenotype in Saccharomyces kluyveri*. J Bacteriol, 2001. **183**(8): p. 2485-9.
36. Chitsulo, L., et al., *The global status of schistosomiasis and its control*. Acta Trop, 2000. **77**(1): p. 41-51.
37. The schistosome, U.o.C. 1998 5th October 1998 [cited; Available from: http://www.path.cam.ac.uk/~schisto/SchistoLife/Miracidium.html.
38. McLaren, D.J. and D.J. Hockley, *Blood flukes have a double outer membrane*. Nature, 1977. **269**(5624): p. 147-9.
39. Ross, A.G., et al., *Schistosomiasis*. N Engl J Med, 2002. **346**(16): p. 1212-20.
40. Wilson, M.S., et al., *Immunopathology of schistosomiasis*. Immunol Cell Biol, 2007. **85**(2): p. 148-54.
41. Brindley, P.J. and A. Sher, *The chemotherapeutic effect of praziquantel against Schistosoma mansoni is dependent on host antibody response*. J Immunol, 1987. **139**(1): p. 215-20.
42. Andrews, P., *Praziquantel: mechanisms of anti-schistosomal activity*. Pharmacol Ther, 1985. **29**(1): p. 129-56.

43. King, C.H., E.M. Muchiri, and J.H. Ouma, *Evidence against rapid emergence of praziquantel resistance in Schistosoma haematobium, Kenya*. Emerg Infect Dis, 2000. **6**(6): p. 585-94.
44. Hu, W., et al., *Evolutionary and biomedical implications of a Schistosoma japonicum complementary DNA resource*. Nat Genet, 2003. **35**(2): p. 139-47.
45. Verjovski-Almeida, S., et al., *Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni*. Nat Genet, 2003. **35**(2): p. 148-57.
46. Haas, B.J., et al., *Schistosoma mansoni genome: closing in on a final gene set*. Exp Parasitol, 2007. **117**(3): p. 225-8.
47. Robert, A.B., *Primate brain evolution: Integrating comparative, neurophysiological, and ethological data*. Evolutionary Anthropology: Issues, News, and Reviews, 2006. **15**(6): p. 224-236.
48. Jerison, H.J., *Evolution of the brain and intelligence*. 1973, New York ; London: Academic Press. xiv, 482 p.
49. Allman, J.M., *Evolving brains*. Scientific American Library paperback. 1999, New York: Scientific American Library. xi, 224 p.
50. Roth, G. and U. Dicke, *Evolution of the brain and intelligence*. Trends Cogn Sci, 2005. **9**(5): p. 250-7.
51. Deaner, R.O., et al., *Overall brain size, and not encephalization quotient, best predicts cognitive ability across non-human primates*. Brain Behav Evol, 2007. **70**(2): p. 115-24.
52. Garamszegi, L.Z., et al., *Maternal effects and the evolution of brain size in birds: overlooked developmental constraints*. Neurosci Biobehav Rev, 2007. **31**(4): p. 498-515.
53. Aiello, L.C., *The Expensive-Tissue Hypothesis: The Brain and the Digestive System in Human and Primate Evolution*. Current anthropology, 1995. **36**(2): p. 199.
54. Jones, K.E. and A.M. MacLarnon, *Affording larger brains: testing hypotheses of mammalian brain evolution on bats*. Am Nat, 2004. **164**(1): p. E20-31.
55. Pitnick, S., K.E. Jones, and G.S. Wilkinson, *Mating system and brain size in bats*. Proc Biol Sci, 2006. **273**(1587): p. 719-24.
56. Clark, D.A., P.P. Mitra, and S.S. Wang, *Scalable architecture in mammalian brains*. Nature, 2001. **411**(6834): p. 189-93.
57. Krubitzer, L., H. Kunzle, and J. Kaas, *Organization of sensory cortex in a Madagascan insectivore, the tenrec (Echinops telfairi)*. J Comp Neurol, 1997. **379**(3): p. 399-414.
58. Garamszegi, L.Z., et al., *Sperm competition and sexually size dimorphic brains in birds*. Proc Biol Sci, 2005. **272**(1559): p. 159-66.
59. Schillaci, M.A., *Sexual selection and the evolution of brain size in primates*. PLoS ONE, 2006. **1**: p. e62.
60. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
61. Venter, J.C., et al., *The sequence of the human genome*. Science, 2001. **291**(5507): p. 1304-51.
62. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome*. Nature, 2002. **420**(6915): p. 520-62.
63. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution*. Nature, 2004. **428**(6982): p. 493-521.
64. Aparicio, S., et al., *Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes*. Science, 2002. **297**(5585): p. 1301-10.

65. Misra, S., et al., *Annotation of the Drosophila melanogaster euchromatic genome: a systematic review*. Genome Biol, 2002. **3**(12): p. RESEARCH0083.
66. Hoskins, R.A., et al., *Heterochromatic sequences in a Drosophila whole-genome shotgun assembly*. Genome Biol, 2002. **3**(12): p. RESEARCH0085.
67. Galagan, J.E., et al., *The genome sequence of the filamentous fungus Neurospora crassa*. Nature, 2003. **422**(6934): p. 859-68.
68. Yu, J., et al., *A draft sequence of the rice genome (Oryza sativa L. ssp. indica)*. Science, 2002. **296**(5565): p. 79-92.
69. Goff, S.A., et al., *A draft sequence of the rice genome (Oryza sativa L. ssp. japonica)*. Science, 2002. **296**(5565): p. 92-100.
70. Clark, A.G., et al., *Evolution of genes and genomes on the Drosophila phylogeny*. Nature, 2007. **450**(7167): p. 203-18.
71. Cliften, P., et al., *Finding functional features in Saccharomyces genomes by phylogenetic footprinting*. Science, 2003. **301**(5629): p. 71-6.
72. Fleischmann, R.D., et al., *Whole-genome random sequencing and assembly of Haemophilus influenzae Rd*. Science, 1995. **269**(5223): p. 496-512.
73. Burke, D.T., G.F. Carle, and M.V. Olson, *Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors*. Science, 1987. **236**(4803): p. 806-12.
74. Venter, J.C., et al., *Shotgun sequencing of the human genome*. Science, 1998. **280**(5369): p. 1540-2.
75. LoVerde, P.T., et al., *Schistosoma mansoni genome project: an update*. Parasitol Int, 2004. **53**(2): p. 183-92.
76. NCBI. *A science primer*. 2004 [cited; Available from: http://www.ncbi.nlm.nih.gov/About/primer/est.html
77. Quackenbush, J., et al., *The TIGR gene indices: reconstruction and representation of expressed gene sequences*. Nucleic Acids Res, 2000. **28**(1): p. 141-5.
78. Christoffels, A., et al., *STACK: Sequence Tag Alignment and Consensus Knowledgebase*. Nucleic Acids Res, 2001. **29**(1): p. 234-8.
79. Sasaki, Y.F., D. Ayusawa, and M. Oishi, *Construction of a normalized cDNA library by introduction of a semi-solid mRNA-cDNA hybridization system*. Nucleic Acids Res, 1994. **22**(6): p. 987-92.
80. Altschul, S.F., et al., *Basic local alignment search tool*. J Mol Biol, 1990. **215**(3): p. 403-10.
81. Darwin, C., *On the origin of species by means of natural selection or the preservation of favoured races in the struggle of life*. 1859. London: Murray.
82. Theissen, G., *Secret life of genes*. Nature, 2002. **415**(6873): p. 741.
83. Webber, C. and C.P. Ponting, *Genes and homology*. Curr Biol, 2004. **14**(9): p. R332-3.
84. Fitch, W.M., *Homology a personal view on some of the problems*. Trends Genet, 2000. **16**(5): p. 227-31.
85. Wikipedia. *Mutation*. 2008 [cited; Available from: http://en.wikipedia.org/wiki/Mutation.
86. Tajima, S. and I. Suetake, *Regulation and function of DNA methylation in vertebrates*. J Biochem, 1998. **123**(6): p. 993-9.
87. Keller, I., D. Bensasson, and R.A. Nichols, *Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes*. PLoS Genet, 2007. **3**(2): p. e22.

88. Ikehata, H., et al., *Distribution of spontaneous CpG-associated G:C --> A:T mutations in the lacZ gene of Muta mice: effects of CpG methylation, the sequence context of CpG sites, and severity of mutations on the activity of the lacZ gene product*. Environ Mol Mutagen, 2000. **36**(4): p. 301-11.
89. Zhang, J., *Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes*. J Mol Evol, 2000. **50**(1): p. 56-68.
90. Kristina Strandberg, A.K. and L.A. Salter, *A comparison of methods for estimating the transition:transversion ratio from DNA sequences*. Mol Phylogenet Evol, 2004. **32**(2): p. 495-503.
91. Yang, Z. and A.D. Yoder, *Estimation of the transition/transversion rate bias and species sampling*. J Mol Evol, 1999. **48**(3): p. 274-83.
92. Kimura, M., *The neutral theory of molecular evolution*. 1983, Cambridge: Cambridge University Press. xv, 367 p.
93. Kimura, M., *Evolutionary rate at the molecular level*. Nature, 1968. **217**(129): p. 624-6.
94. Zhang, L. and W.-H. Li, *Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes*. Mol Biol Evol, 2004. **21**(2): p. 236-239.
95. Yang, Z., et al., *Codon-substitution models for heterogeneous selection pressure at amino acid sites*. Genetics, 2000. **155**(1): p. 431-49.
96. Ponting, C.P., *Issues in predicting protein function from sequence*. Brief Bioinform, 2001. **2**(1): p. 19-29.
97. Wikipedia. *Sequence alignment*. 5 October 2008 [cited; Available from: http://en.wikipedia.org/wiki/Sequence_alignment.
98. Korf, I., et al., *Blast*. 2003, Sebastopol, Calif. ; Farnham: O'Reilly. xviii, 339.
99. Henikoff, S. and J.G. Henikoff, *Performance evaluation of amino acid substitution matrices*. Proteins, 1993. **17**(1): p. 49-61.
100. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
101. Karlin, S. and S.F. Altschul, *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes*. Proc Natl Acad Sci U S A, 1990. **87**(6): p. 2264-8.
102. Altschul, S.F. and E.V. Koonin, *Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases*. Trends Biochem Sci, 1998. **23**(11): p. 444-7.
103. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase*. Nucleic Acids Res, 2004. **32**(Database issue): p. D115-9.
104. NCBI, N.C.f.B.I. 9th May, 2008 [cited; Available from: http://www.ncbi.nlm.nih.gov/.
105. Slater, G.S. and E. Birney, *Automated generation of heuristics for biological sequence comparison*. BMC Bioinformatics, 2005. **6**: p. 31.
106. Birney, E., M. Clamp, and R. Durbin, *GeneWise and Genomewise*. Genome Res, 2004. **14**(5): p. 988-95.
107. Wallace, I.M., et al., *M-Coffee: combining multiple sequence alignment methods with T-Coffee*. Nucleic Acids Res, 2006. **34**(6): p. 1692-9.
108. Corpet, F., *Multiple sequence alignment with hierarchical clustering*. Nucleic Acids Res, 1988. **16**(22): p. 10881-90.
109. Morgenstern, B., *DIALIGN: multiple DNA and protein sequence alignment at BiBiServ*. Nucleic Acids Res, 2004. **32**(Web Server issue): p. W33-6.

110. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matirx choice*. Nucl. Acids Res., 1994. **16**: p. 1861-1864.
111. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment*. J Mol Biol, 2000. **302**(1): p. 205-17.
112. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
113. Notredame, C., *Recent progress in multiple sequence alignment: a survey*. Pharmacogenomics, 2002. **3**(1): p. 131-44.
114. Durbin, R., *Biological sequence analysis : probabilistic models of proteins and nucleic acids*. 1998, Cambridge: Cambridge University Press. xi, 356.
115. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees*. Mol Biol Evol, 1987. **4**(4): p. 406-25.
116. Swofford, D.L., et al., *Phylogenetic Inference*, in *Molecular Systematics*. 1996, Sinauer Associaets, Inc.: MA, USA. p. 407-514.
117. Felsenstein, J., *Confidence limits on phylogenies: an approach using the bootstrap*. Evolution, 1985. **39**(4): p. 783.
118. Efron, B., E. Halloran, and S. Holmes, *Bootstrap confidence levels for phylogenetic trees*. Proc Natl Acad Sci U S A, 1996. **93**(23): p. 13429-34.
119. Nei, M., *Selectionism and neutralism in molecular evolution*. Mol Biol Evol, 2005. **22**(12): p. 2318-42.
120. Massingham, T. and N. Goldman, *Detecting amino acid sites under positive selection and purifying selection*. Genetics, 2005. **169**(3): p. 1753-62.
121. Yang, Z., *Adaptive molecular evolution*, in *Handbook of Statistical Genetics*, M. Bishop, C. Cannings, and D. Balding, Editors. 2003, Wiley: New York. p. 229-254.
122. Nielsen, R. and Z. Yang, *Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene*. Genetics, 1998. **148**(3): p. 929-36.
123. Yang, Z. and R. Nielsen, *Synonymous and nonsynonymous rate variation in nuclear genes of mammals*. J Mol Evol, 1998. **46**(4): p. 409-18.
124. Yang, Z., W.S. Wong, and R. Nielsen, *Bayes empirical bayes inference of amino acid sites under positive selection*. Mol Biol Evol, 2005. **22**(4): p. 1107-18.
125. Wong, W.S., et al., *Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites*. Genetics, 2004. **168**(2): p. 1041-51.
126. Yang, Z., *The power of phylogenetic comparison in revealing protein function*. PNAS, 2005. **102**(9): p. 3179-3180.
127. Anisimova, M., J.P. Bielawski, and Z. Yang, *Accuracy and power of bayes prediction of amino acid sites under positive selection*. Mol Biol Evol, 2002. **19**(6): p. 950-8.
128. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*. Nat Genet, 2000. **25**(1): p. 25-9.
129. Sokal, R.R. and F.J. Rohlf, *Biometry : the principles and practice of statistics in biological research*. 3rd ed ed. 1995, New York: W.H. Freeman. xix, 887.
130. Easton, V.J.a.M., J.H. *Statistics Glossary*. 1997 September 1997 [cited; Available from: http://www.stats.gla.ac.uk/steps/glossary/nonparametric.html.

131. Benjamini, Y., *Controlling the false discovery rate-a practical and powerful approach for multicomparison testing*. Journal of the Royal Statistical Society. Series C, Applied statistics, 1995. **57**: p. 289.
132. Swanson, W.J., et al., *Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals*. PNAS %R 10.1073/pnas.051605998, 2001. **98**(5): p. 2509-2514.
133. Clark, N.L. and W.J. Swanson, *Pervasive Adaptive Evolution in Primate Seminal Proteins*. PLoS Genetics, 2005. **1**(3): p. e35.
134. Opazo, J.C., et al., *Adaptive evolution of the insulin gene in caviomorph rodents*. Mol Biol Evol, 2005. **22**(5): p. 1290-8.
135. Liu, J.C., et al., *Episodic evolution of growth hormone in primates and emergence of the species specificity of human growth hormone receptor*. Mol Biol Evol, 2001. **18**(6): p. 945-53.
136. Andolfatto, P., *Adaptive evolution of non-coding DNA in Drosophila*. Nature, 2005. **437**(7062): p. 1149-52.
137. Welch, J.J., *Estimating the genome-wide rate of adaptive protein evolution in Drosophila*. Genetics, 2006.
138. Enard, W., et al., *Molecular evolution of FOXP2, a gene involved in speech and language*. Nature, 2002. **418**(6900): p. 869-72.
139. Emes, R.D., et al., *Comparison of the genomes of human and mouse lays the foundation of genome zoology*. Hum Mol Genet, 2003. **12**(7): p. 701-9.
140. Wolfe, K.H. and W.H. Li, *Molecular evolution meets the genomics revolution*. Nat Genet, 2003. **33 Suppl**: p. 255-65.
141. Mekel-Bobrov, N., et al., *Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens*. Science, 2005. **309**(5741): p. 1720-2.
142. Jordan, I.K., et al., *Essential genes are more evolutionarily conserved than are nonessential genes in bacteria*. Genome Res, 2002. **12**(6): p. 962-8.
143. Drummond, D.A., A. Raval, and C.O. Wilke, *A single determinant dominates the rate of yeast protein evolution*. Mol Biol Evol, 2006. **23**(2): p. 327-37.
144. Pal, C., B. Papp, and M.J. Lercher, *An integrated view of protein evolution*. Nat Rev Genet, 2006. **7**(5): p. 337-48.
145. Chamary, J.V., J.L. Parmley, and L.D. Hurst, *Hearing silence: non-neutral evolution at synonymous sites in mammals*. Nat Rev Genet, 2006. **7**(2): p. 98-108.
146. Yang, Z., *PAML: a program package for phylogenetic analysis by maximum likelihood*. Comput Appl Biosci, 1997. **13**(5): p. 555-6.
147. Anisimova, M., J.P. Bielawski, and Z. Yang, *Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution*. Mol Biol Evol, 2001. **18**(8): p. 1585-92.
148. Tatusova, T.A. and T.L. Madden, *BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences*. FEMS Microbiol Lett, 1999. **174**(2): p. 247-50.
149. Yon Rhee, S., et al., *Use and misuse of the gene ontology annotations*. Nat Rev Genet, 2008. **9**(7): p. 509-15.
150. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes*. J Mol Biol, 2001. **305**(3): p. 567-80.
151. Emanuelsson, O., et al., *Locating proteins in the cell using TargetP, SignalP and related tools*. Nat Protoc, 2007. **2**(4): p. 953-71.

152. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000. **28**(1): p. 235-42.
153. Rost, B. and C. Sander, *Conservation and prediction of solvent accessibility in protein families*. Proteins, 1994. **20**(3): p. 216-26.
154. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*. Biopolymers, 1983. **22**(12): p. 2577-637.
155. Goodstadt, L. and C.P. Ponting, *CHROMA: consensus-based colouring of multiple alignments for publication*. Bioinformatics, 2001. **17**(9): p. 845-6.
156. *Signal Peptides*. 2008 [cited; Available from: http://en.wikipedia.org/wiki/Signal_peptide.
157. Emes, R.D., et al., *Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents*. Genome Res, 2004. **14**(4): p. 591-602.
158. Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae*. Nature, 2000. **403**(6770): p. 623-7.
159. Finn, R.D., et al., *The Pfam protein families database*. Nucleic Acids Res, 2008. **36**(Database issue): p. D281-8.
160. Lim, C.S., et al., *Regulation of SPIN90 phosphorylation and interaction with Nck by ERK and cell adhesion*. J Biol Chem, 2003. **278**(52): p. 52116-23.
161. Kim, D.J., et al., *Interaction of SPIN90 with the Arp2/3 complex mediates lamellipodia and actin comet tail formation*. J Biol Chem, 2006. **281**(1): p. 617-25.
162. Morton, C.J. and I.D. Campbell, *SH3 Domains: Molecular [`]Velcro'*. Current Biology, 1994. **4**(7): p. 615-617.
163. Eisenmann, K.M., et al., *Dia-Interacting Protein Modulates Formin-Mediated Actin Assembly at the Cell Cortex*. Current Biology, 2007. **17**(7): p. 579-591.
164. Suzuki, Y. and T. Gojobori, *A method for detecting positive selection at single amino acid sites*. Mol Biol Evol, 1999. **16**(10): p. 1315-1328.
165. Urwin, R., et al., *Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen PorB*. Mol Biol Evol, 2002. **19**(10): p. 1686-94.
166. Andrade, S.L., et al., *Crystal structure of the NADH:quinone oxidoreductase WrbA from Escherichia coli*. J Bacteriol, 2007. **189**(24): p. 9101-7.
167. Brown, C.J., et al., *Evolutionary rate heterogeneity in proteins with long disordered regions*. J Mol Evol, 2002. **55**(1): p. 104-10.
168. Dunker, A.K., et al., *Intrinsic disorder and protein function*. Biochemistry, 2002. **41**(21): p. 6573-82.
169. Vitkup, D., C. Sander, and G.M. Church, *The amino-acid mutational spectrum of human genetic disease*. Genome Biol, 2003. **4**(11): p. R72.
170. Wolfe, K.H., *Comparative genomics and genome evolution in yeasts*. Philos Trans R Soc Lond B Biol Sci, 2006. **361**(1467): p. 403-12.
171. Fabrizio, P. and V.D. Longo, *The chronological life span of Saccharomyces cerevisiae*. Aging Cell, 2003. **2**(2): p. 73-81.
172. Copley, R.R., L. Goodstadt, and C. Ponting, *Eukaryotic domain evolution inferred from genome comparisons*. Curr Opin Genet Dev, 2003. **13**(6): p. 623-8.
173. Malpertuy, A., et al., *Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes*. FEBS Lett, 2000. **487**(1): p. 113-21.

174. Ponting, C.P. and G. Lunter, *Signatures of adaptive evolution within human non-coding sequence*. Hum Mol Genet, 2006. **15 Spec No 2**: p. R170-5.
175. dos Reis, M., R. Savva, and L. Wernisch, *Solving the riddle of codon usage preferences: a test for translational selection*. Nucleic Acids Res, 2004. **32**(17): p. 5036-44.
176. Akashi, H., *Translational selection and yeast proteome evolution*. Genetics, 2003. **164**(4): p. 1291-303.
177. Carbone, A., A. Zinovyev, and F. Kepes, *Codon adaptation index as a measure of dominating codon bias*. Bioinformatics, 2003. **19**(16): p. 2005-15.
178. Comeron, J.M. and M. Kreitman, *The correlation between synonymous and nonsynonymous substitutions in Drosophila: mutation, selection or relaxed constraints?* Genetics, 1998. **150**(2): p. 767-75.
179. Kliman, R.M., N. Irving, and M. Santiago, *Selection conflicts, gene expression, and codon usage trends in yeast*. J Mol Evol, 2003. **57**(1): p. 98-109.
180. Powell, J.R. and E.N. Moriyama, *Evolution of codon usage bias in Drosophila*. Proc Natl Acad Sci U S A, 1997. **94**(15): p. 7784-90.
181. Sharp, P.M., et al., *Variation in the strength of selected codon usage bias among bacteria*. Nucleic Acids Res, 2005. **33**(4): p. 1141-53.
182. Sharp, P.M., et al., *Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity*. Nucleic Acids Res, 1988. **16**(17): p. 8207-11.
183. Duan, J. and M.A. Antezana, *Mammalian mutation pressure, synonymous codon choice, and mRNA degradation*. J Mol Evol, 2003. **57**(6): p. 694-701.
184. Sharp, P.M. and W.H. Li, *Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons*. Nucleic Acids Res, 1986. **14**(19): p. 7737-49.
185. Bennetzen, J.L. and B.D. Hall, *Codon selection in yeast*. J Biol Chem, 1982. **257**(6): p. 3026-31.
186. Wright, F., *The 'effective number of codons' used in a gene*. Gene, 1990. **87**(1): p. 23-9.
187. Akashi, H., *Gene expression and molecular evolution*. Curr Opin Genet Dev, 2001. **11**(6): p. 660-6.
188. Hirsh, A.E., H.B. Fraser, and D.P. Wall, *Adjusting for selection on synonymous sites in estimates of evolutionary distance*. Mol Biol Evol, 2005. **22**(1): p. 174-7.
189. Bulmer, M., *Are codon usage patterns in unicellular organisms determined by selection-mutation balance?* Journal of Evolutionary Biology, 1988. **1**(1): p. 15-26.
190. Sharp, P.M. and W.H. Li, *The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias*. Mol Biol Evol, 1987. **4**(3): p. 222-30.
191. Urrutia, A.O. and L.D. Hurst, *Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection*. Genetics, 2001. **159**(3): p. 1191-9.
192. Heger, A. and C.P. Ponting, *Evolution of codon bias among twelve drosophila species*. 2007.

193. Shields, D.C. and P.M. Sharp, *Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases*. Nucleic Acids Res, 1987. **15**(19): p. 8023-40.
194. Sharp, P.M. and W.H. Li, *The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications*. Nucleic Acids Res, 1987. **15**(3): p. 1281-95.
195. Wagner, A., *Inferring lifestyle from gene expression patterns*. Mol Biol Evol, 2000. **17**(12): p. 1985-7.
196. Datta, A. and S. Jinks-Robertson, *Association of increased spontaneous mutation rates with high levels of transcription in yeast*. Science, 1995. **268**(5217): p. 1616-9.
197. Qin, H., et al., *Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes*. Genetics, 2004. **168**(4): p. 2245-60.
198. Comeron, J.M. and M. Aguade, *An evaluation of measures of synonymous codon usage bias*. J Mol Evol, 1998. **47**(3): p. 268-74.
199. Beyer, A., et al., *Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale*. Mol Cell Proteomics, 2004. **3**(11): p. 1083-92.
200. Drummond, D.A., et al., *Why highly expressed proteins evolve slowly*. Proc Natl Acad Sci U S A, 2005. **102**(40): p. 14338-43.
201. Cooper, A., *Thermodynamics of Protein Folding and Stability*, in *Protein: A Comprehensive Treatise*, G. Allen, Editor. 1999, JAI Press Inc. p. 217-270.
202. Jona, G., L.L. Livi, and O. Gileadi, *Mutations in the RING domain of TFB3, a subunit of yeast transcription factor IIH, reveal a role in cell cycle progression*. J Biol Chem, 2002. **277**(42): p. 39409-16.
203. Borden, K.L., *RING domains: master builders of molecular scaffolds?* J Mol Biol, 2000. **295**(5): p. 1103-12.
204. NHS. *Schistosomiasis (Bilharzia)*. 2008 [cited; Available from: http://www.nhs.uk/Conditions/schistosomiasis/Pages/Questionstoaskpage.aspx?url=Pages/Questionstoask.aspx.
205. WHO, s. 2008 2008 [cited; Available from: http://www.who.int/vaccine_research/diseases/soa_parasitic/en/index5.html.
206. Cioli, D. and L. Pica-Mattoccia, *Praziquantel*. Parasitol Res, 2003. **90 Supp 1**: p. S3-9.
207. de Mendonca, R.L., et al., *Hormones and nuclear receptors in schistosome development*. Parasitol Today, 2000. **16**(6): p. 233-40.
208. Wu, W., et al., *Schistosoma mansoni (Platyhelminthes, Trematoda) nuclear receptors: sixteen new members and a novel subfamily*. Gene, 2006. **366**(2): p. 303-15.
209. Bertin, B., et al., *The monomeric orphan nuclear receptor Schistosoma mansoni Ftz-F1 dimerizes specifically and functionally with the schistosome RXR homologue, SmRXR1*. Biochem Biophys Res Commun, 2005. **327**(4): p. 1072-82.
210. Hu, R., et al., *SmTR2/4, a Schistosoma mansoni homologue of TR2/TR4 orphan nuclear receptor*. Int J Parasitol, 2006. **36**(10-11): p. 1113-22.
211. Nilsson, M., et al., *Elk1 and SRF transcription factors convey basal transcription and mediate glucose response via their binding sites in the human LXRB gene promoter*. Nucleic Acids Res, 2007. **35**(14): p. 4858-68.
212. Robinson-Rechavi, M., H. Escriva Garcia, and V. Laudet, *The nuclear receptor superfamily*. J Cell Sci, 2003. **116**(Pt 4): p. 585-6.

213. Wu, W., et al., *Identification and characterization of a nuclear receptor subfamily I member in the Platyhelminth Schistosoma mansoni (SmNR1).* FEBS J, 2007. **274**(2): p. 390-405.
214. Liu, F., et al., *New perspectives on host-parasite interplay by comparative transcriptomic and proteomic analyses of Schistosoma japonicum.* PLoS Pathog, 2006. **2**(4): p. e29.
215. Korf, I., et al., *Integrating genomic homology into gene structure prediction.* Bioinformatics, 2001. **17 Suppl 1**: p. S140-8.
216. Apweiler, R., et al., *InterPro--an integrated documentation resource for protein families, domains and functional sites.* Bioinformatics, 2000. **16**(12): p. 1145-50.
217. Camon, E., et al., *The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro.* Genome Res, 2003. **13**(4): p. 662-72.
218. Biswas, M., et al., *Applications of InterPro in protein annotation and genome analysis.* Brief Bioinform, 2002. **3**(3): p. 285-95.
219. Nei, M. and T. Gojobori, *Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.* Mol Biol Evol, 1986. **3**(5): p. 418-26.
220. Affolter, M., et al., *The Drosophila SRF homolog is expressed in a subset of tracheal cells and maps within a genomic region required for tracheal development.* Development, 1994. **120**(4): p. 743-53.
221. Casero, M.C. and L. Sastre, *A serum response factor homologue is expressed in ectodermal tissues during development of the crustacean Artemia franciscana.* Mech Dev, 2000. **96**(2): p. 229-32.
222. Kim, S.W., et al., *Retinoid-dependent antagonism of serum response factor transactivation mediated by transcriptional coactivator proteins.* Oncogene, 2001. **20**(45): p. 6638-42.
223. Poser, S., et al., *SRF-dependent gene expression is required for PI3-kinase-regulated cell proliferation.* EMBO J, 2000. **19**(18): p. 4955-66.
224. Finn, R.D., et al., *Pfam: clans, web tools and services.* Nucleic Acids Res, 2006. **34**(Database issue): p. D247-51.
225. Flores, G.V., et al., *Lozenge is expressed in pluripotent precursor cells and patterns multiple cell types in the Drosophila eye through the control of cell-specific transcription factors.* Development, 1998. **125**(18): p. 3681-7.
226. Daga, A., et al., *Patterning of cells in the Drosophila eye by Lozenge, which shares homologous domains with AML1.* Genes Dev, 1996. **10**(10): p. 1194-205.
227. Zhang, Y.W., et al., *A novel transcript encoding an N-terminally truncated AML1/PEBP2 alphaB protein interferes with transactivation and blocks granulocytic differentiation of 32Dcl3 myeloid cells.* Mol Cell Biol, 1997. **17**(7): p. 4133-45.
228. Fazi, F., et al., *Heterochromatic gene repression of the retinoic acid pathway in acute myeloid leukemia.* Blood, 2007. **109**(10): p. 4432-40.
229. Stacey, M.W., et al., *Nuclear receptor co-repressor gene localizes to 17p11.2, a frequently deleted band in malignant disorders.* Genes Chromosomes Cancer, 1999. **25**(2): p. 191-3.
230. Suzuki, F., et al., *Functional interactions of transcription factor human GA-binding protein subunits.* J Biol Chem, 1998. **273**(45): p. 29302-8.

231. Toku, S., I.K. Quaye, and T. Tanaka, *Isolation and characterization of chicken GA-binding protein*. Biochim Biophys Acta, 2002. **1579**(1): p. 50-4.
232. Yang, Z.F., S. Mott, and A.G. Rosmarin, *The Ets transcription factor GABP is required for cell-cycle progression*. Nat Cell Biol, 2007. **9**(3): p. 339-46.
233. Rosmarin, A.G., et al., *GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein-protein interactions*. Blood Cells Mol Dis, 2004. **32**(1): p. 143-54.
234. Bush, T.S., et al., *GA-binding protein (GABP) and Sp1 are required, along with retinoid receptors, to mediate retinoic acid responsiveness of CD18 (beta 2 leukocyte integrin): a novel mechanism of transcriptional regulation in myeloid cells*. Blood, 2003. **101**(1): p. 311-7.
235. Monden, T., et al., *p120 acts as a specific coactivator for 9-cis-retinoic acid receptor (RXR) on peroxisome proliferator-activated receptor-gamma/RXR heterodimers*. Mol Endocrinol, 1999. **13**(10): p. 1695-703.
236. Pruitt, K.D. and D.R. Maglott, *RefSeq and LocusLink: NCBI gene-centered resources*. Nucleic Acids Res, 2001. **29**(1): p. 137-40.
237. Doyon, Y., et al., *Structural and functional conservation of the NuA4 histone acetyltransferase complex from yeast to humans*. Mol Cell Biol, 2004. **24**(5): p. 1884-96.
238. Havercroft, J.C. and A.L. Smith, *Localization of the A1.12/9 antigen family to the neurones, putative sensory receptors and tegument of Schistosoma mansoni*. Parasite Immunol, 1993. **15**(7): p. 361-71.
239. Okuda, T., et al., *Molecular cloning of macrophin, a human homologue of Drosophila kakapo with a close structural similarity to plectin and dystrophin*. Biochem Biophys Res Commun, 1999. **264**(2): p. 568-74.
240. Gregory, S.L. and N.H. Brown, *kakapo, a gene required for adhesion between and within cell layers in Drosophila, encodes a large cytoskeletal linker protein related to plectin and dystrophin*. J Cell Biol, 1998. **143**(5): p. 1271-82.
241. Zanotto, E., Z.H. Shah, and H.T. Jacobs, *The bidirectional promoter of two genes for the mitochondrial translational apparatus in mouse is regulated by an array of CCAAT boxes interacting with the transcription factor NF-Y*. Nucleic Acids Res, 2007. **35**(2): p. 664-77.
242. Yokogawa, T., et al., *Characterization and tRNA recognition of mammalian mitochondrial seryl-tRNA synthetase*. J Biol Chem, 2000. **275**(26): p. 19913-20.
243. Garg, R.P., et al., *Investigations of valanimycin biosynthesis: elucidation of the role of seryl-tRNA*. Proc Natl Acad Sci U S A, 2008. **105**(18): p. 6543-7.
244. Gibbons, W.J., Jr., et al., *Genomic organization, expression, and subcellular localization of mouse mitochondrial seryl-tRNA synthetase*. Biochem Biophys Res Commun, 2004. **317**(3): p. 774-8.
245. Skelly, P.J. and C.B. Shoemaker, *Rapid appearance and asymmetric distribution of glucose transporter SGTP4 at the apical surface of intramammalian-stage Schistosoma mansoni*. Proc Natl Acad Sci U S A, 1996. **93**(8): p. 3642-6.
246. Stavitsky, A.B., *Regulation of granulomatous inflammation in experimental models of schistosomiasis*. Infect Immun, 2004. **72**(1): p. 1-12.
247. Michailidou, Z., et al., *Glucocorticoid receptor haploinsufficiency causes hypertension and attenuates hypothalamic-pituitary-adrenal axis and blood pressure adaptions to high-fat diet*. FASEB J, 2008. **22**(11): p. 3896-907.

248. Reiche, E.M.V., S.O.V. Nunes, and H.K. Morimoto, *Stress, depression, the immune system, and cancer*. The Lancet Oncology, 2004. **5**(10): p. 617-625.
249. Bergquist, N.R. and D.G. Colley, *Schistosomiasis vaccine:research to development*. Parasitol Today, 1998. **14**(3): p. 99-104.
250. Gobert, G.N., *Immunolocalization of schistosome proteins*. Microsc Res Tech, 1998. **42**(3): p. 176-85.
251. Tran, M.H., et al., *Tetraspanins on the surface of Schistosoma mansoni are protective antigens against schistosomiasis*. Nat Med, 2006. **12**(7): p. 835-40.
252. Abath, F.G. and R.C. Werkhauser, *The tegument of Schistosoma mansoni: functional and immunological features*. Parasite Immunol, 1996. **18**(1): p. 15-20.
253. Davis, A.H., et al., *Isolation of cDNA clones for differentially expressed genes of the human parasite Schistosoma mansoni*. Proc Natl Acad Sci U S A, 1986. **83**(15): p. 5534-8.
254. Khaitovich, P., et al., *Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees*. Science, 2005. **309**(5742): p. 1850-4.
255. Seyfarth, R.M. and D.L. Cheney, *What are big brains for?* Proc Natl Acad Sci U S A, 2002. **99**(7): p. 4141-2.
256. Changizi, M.A. and S. Shimojo, *Parcellation and area-area connectivity as a function of neocortex size*. Brain Behav Evol, 2005. **66**(2): p. 88-98.
257. Marino, L., *Absolute brain size: did we throw the baby out with the bathwater?* Proc Natl Acad Sci U S A, 2006. **103**(37): p. 13563-4.
258. Macphail, E.M., *Brain and intelligence in vertebrates*. Oxford science publications. 1982, Oxford: Clarendon Press. viii, 423 p.
259. Heger, A. and C.P. Ponting, *OPTIC: orthologous and paralogous transcripts in clades*. Nucleic Acids Res, 2008. **36**(Database issue): p. D267-70.
260. Crile, G. and D.P. Quiring, *A record of body weight and certain organ and gland weights of 3690 animals*. The Ohio Journal of Science, 1940. **40**: p. 219-259.
261. Goodstadt, L. and C.P. Ponting, *Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human*. PLoS Comput Biol, 2006. **2**(9): p. e133.
262. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes*. Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6062-7.
263. Su, A.I., et al., *Large-scale analysis of the human and mouse transcriptomes*. Proc Natl Acad Sci U S A, 2002. **99**(7): p. 4465-70.
264. Ohta, T., *Near-neutrality in evolution of genes and gene regulation*. Proc Natl Acad Sci U S A, 2002. **99**(25): p. 16134-7.
265. Keightley, P.D., M.J. Lercher, and A. Eyre-Walker, *Evidence for widespread degradation of gene control regions in hominid genomes*. PLoS Biol, 2005. **3**(2): p. e42.
266. Takahata, N., *Allelic genealogy and human evolution*. Mol Biol Evol, 1993. **10**(1): p. 2-22.
267. Ponting, C.P., *The functional repertoires of metazoan genomes*. Nat Rev Genet, 2008. **9**(9): p. 689-98.
268. Ohta, T., *Amino acid substitution at the Adh locus of Drosophila is facilitated by small population size*. Proc Natl Acad Sci U S A, 1993. **90**(10): p. 4548-51.
269. Evans, P.D., et al., *Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans*. Science, 2005. **309**(5741): p. 1717-20.

270. Vallender, E.J., N. Mekel-Bobrov, and B.T. Lahn, *Genetic basis of human brain evolution*. Trends Neurosci, 2008.
271. Kouprina, N., et al., *The microcephaly ASPM gene is expressed in proliferating tissues and encodes for a mitotic spindle protein*. Hum Mol Genet, 2005. **14**(15): p. 2155-65.
272. Trimborn, M., et al., *Mutations in microcephalin cause aberrant regulation of chromosome condensation*. Am J Hum Genet, 2004. **75**(2): p. 261-6.
273. Ali, F. and R. Meier, *Positive selection in ASPM is correlated with cerebral cortex evolution across primates but not with whole brain size*. Mol Biol Evol, 2008.
274. Tuller, T., M. Kupiec, and E. Ruppin, *Evolutionary Rate and Gene Expression Across Different Brain Regions*. Genome Biol, 2008. **9**(9): p. R142.
275. Castillo-Davis, C.I., et al., *The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint*. Genome Res, 2004. **14**(5): p. 802-11.
276. Cai, J., et al., *Accelerated Evolutionary Rate May Be Responsible for the Emergence of Lineage-Specific Genes in Ascomycota*. Journal of Molecular Evolution, 2006.
277. Alba, M.M. and J. Castresana, *Inverse relationship between evolutionary rate and age of mammalian genes*. Mol Biol Evol, 2005. **22**(3): p. 598-606.
278. Elhaik, E., N. Sabath, and D. Graur, *The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence*. Mol Biol Evol, 2006. **23**(1): p. 1-3.
279. Khaitovich, P., et al., *Metabolic changes in schizophrenia and human brain evolution*. Genome Biol, 2008. **9**(8): p. R124.
280. Sabeti, P.C., et al., *Detecting recent positive selection in the human genome from haplotype structure*. Nature, 2002. **419**(6909): p. 832-7.