Picking the High Hanging Fruit:

Automated Ways to Annotate Awkward Genes

by Seán Ó hÉigeartaigh

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

Department of Genetics
University of Dublin
Trinity College

February 2012

Declaration

This thesis is submitted by the undersigned for the degree of Doctor of Philosophy from Trinity College Dublin. I declare that this thesis has not been submitted as an exercise for a degree at this or any other university. Unless otherwise stated, the work in this thesis is entirely my own work; the work of others undertaken in collaboration is duly acknowledged in the text wherever included. I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed…….....

Seán Ó hÉigeartaigh

28th October, 2011

Summary:

In Chapter 2 I describe the development of software called SearchDOGS (*D*atabase of *O*rthologous *G*enomic *S*egments). By identifying regions of conserved local synteny across species using the synteny information contained in the Yeast Gene Order Browser (YGOB) and combining this information with standart BLAST sequence similarity searches, SearchDOGS is able to identify unannotated genes in published yeast genomes with a very high degree of sensitivity. It is particularly effective for identifying short or highly diverged genes that are often missed using standard methods. Using this approach, we have identified 595 unannotated genes across eleven yeast species, incuding two previously unidentified genes in *S. cerevisiae*. Among these, we identify a number of genes coding for the mating pheromone **a**-factor in six species including *Kluyveromyces lactis*; these tiny genes are notoriously difficult to identify by standard methods.

In Chapter 3 I describe the adaptation of SearchDOGS to identify missing genes in bacterial genomes. Bacterial SearchDOGS is a standalone, downloadable package that can be used in conjunction with any set of bacterial genomes that span a suitable evolutionary range, including unpublished or private data. The software automatically generates a pillar homology structure between the genomes in order to calculate the synteny information that is central to the SearchDOGS procedure. HTML results files are generated for each species, including BLAST links, *Ka/Ks* protein sequence conservation estimates and other relevant information for each candidate gene identified, in order to allow the user to make an informed decision regarding the validity of each candidate gene. Using this approach, I identified 171 gene candidates in the *Shigella boydii* sb227 genome, including 62 candidates of length <60 codons.

In Chapter 4 I undertake a comparative analysis in the Saccharomycetaceae of another type of "awkward gene" that is difficult to annotate and sometimes poorly understood: genes that undergo programmed ribosomal frameshifting. I expand on

previous studies of three yeast chromosomal genes, *OAZ1*, *EST3* and *ABP140*, that were previously known to contain a programmed frameshifting signal. I describe a further example of unusual gene evolution, *URA6*, that may be a case of a gene split or a programmed ribosomal frameshift. In the case of *ABP140*, I identify previously unidentified cases of retention of truncated ohnologs following whole genome duplication. The *URA6* locus is particularly notable as it appears to require an unlikely number of events to produce the distribution of full-length and split/frameshifted orthologs regardless of whether this is an example gene split or frameshifting locus.

Appendix II includes the manuscript "Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents" by Jonathan L. Gordon, David Armisén, Estelle Proux-Wéra, Seán S. ÓhÉigeartaigh, Kevin P. Byrne, and Kenneth H. Wolfe, recently accepted for publication by the Proceedings of the National Academy of Science". In a project involving multiple members of the laboratory, we annotated the genomes of seven yeast species that we sequenced, and studied the evolution of the mating-type locus in these species. My role in this project was in the editing and correction of sequence data for the new Saccharomycetaceae family species that were included in this study (soon to be publicly available on YGOB). Also, SearchDOGS was included as an annotation step in the Yeast Genome Annotation Pipeline (YGAP) that was used to annotate these genomes.

Acknowledgements:

My first debt of gratitude must be to my parents Cian agus Aingeal, who have patiently supported me through twenty-three years of being in education of one form or another! Go raibh míle maith agaibh. Gan bhur gcabhair agus bhur dtacaíocht, ní bheadh mé in ann é seo a dhéanamh.

I'm also grateful for nearly four years of great company in the Wolfe Laboratory, and a lot of invaluable advice, insight and friendship from everyone who was there – Estelle, David, Kevin, Áine, Jonathan, Nadia, Gavin, Jeff and Nora, thank you all, (and thanks too to Karsten, always generous with his time and assistance)!

Lastly, I owe a great deal to my supervisor, Ken Wolfe, not only for sharing his tremendous breadth of knowledge and insight, but more importantly for helping me to figure out the right questions to ask. I feel this is an ability that will stand to me in future life, no matter where it takes me.

Table of Contents

Table of Figures

Table of Tables:

ABBREVIATIONS USED

WGD: Whole Genome Duplication

YGOB: Yeast Gene Order Browser

YGAP: Yeast Genome Annotation Pipeline

ORF: Open reading frame.

HSP: High-scoring pair

EST: Expressed sequence tag

SNP: Single nucleotide polymorphism

Chapter 1

Sequencing, annotation and comparative analysis of yeast and bacterial genomes.

Introduction

In this chapter I review the factors that make *Saccharomyces cerevisiae* an ideal model organism and the Saccharomycetaceae an ideal family of species in which to study genomic evolution. I give an overview of the current state of play in prokaryotic and eukaryotic genome sequencing and annotation, and discuss the major problems that exist in the accurate annotation of protein-coding genes. Finally, I discuss in detail approaches for differentiating pseudogenes from genuine protein-coding genes, and the challenge of plucking the "high hanging fruit": accurately identifying the shortest and most highly diverged protein-coding genes within genomes.

1.1 The Saccharomycetaceae family yeasts as model organisms.

The term "yeasts" is a loose classification that describes fungal species that generally exist in a unicellular form and reproduce by budding or fission (Knop 2011). While species falling in this category are scattered across several fungal lineages (Dujon 2010), the bulk of genomics studies have involved the model organism *Saccharomyces cerevisiae* and other species within the subphylum Saccharomycotina (Souciet et al. 2009) contained within the phylum Ascomycota (Kurtzman 2011). The largest of five fungal phyla, the Ascomycota are characterised by the production of an ascus, a structure formed during meiosis that contains the spores (Heckman et al. 2001). The Saccharomycotina are one of three subphyla currently described within the phylum Ascomycomycota (Fitzpatrick et al. 2006; James et al. 2006). While this

taxonomy represents the best of our current knowledge, fungal classifications are constantly in flux as more evidence becomes available from the sequencing of additional genomes and more comprehensive gene sequence analyses. Chapters 2 and 4 of this thesis are mainly focused on the species contained within the subphylum Saccharomycetaceae, a family of the order Saccharomycetales within the Saccharomycotina (Kurtzman 2011) (Figure 1.1). These unicellular yeasts can exist in a haploid or a diploid state, and reproduce by budding (Knop 2011). Chapter 4 also contains comparisons with a group of related species within the Saccharomycotina; these "CTG" group yeasts are so called because they are characterised by a reassignation of the CTG codon to be translated as serine rather than leucine (Fitzpatrick et al. 2006). *Yarrowia lipolytica*, also included in the study, comes from a third lineage within the Saccharomycotina.

Both the first eukaryotic chromosome to be sequenced (Oliver et al. 1992) and the first eukaryotic genome to be completed (Goffeau et al. 1996) belonged to the budding yeast *Saccharomyces cerevisiae*. The genomic characteristics that made *S. cerevisiae* an ideal candidate for a model organism also make the Saccharomycotina an ideal group of species to undertake comparative genomic study on: their genomes are small, have few introns and little noncoding DNA, and show a high level of gene order conservation. The genomes of the species fully annotated to date range from 9 to 21Mb and 4,700 to ~6,500 genes (Dujon 2010), small and compact when compared to the genomes of other eukaryotes such as human (3Gb; ~23,000 genes) (Consortium 2004) and *Arabidopsis thaliana* (157 Mb; 25,000+ genes) (Bennett et al. 2003; Bevan and Walsh 2005), and even the more closely related Pezizoascomycotina such as *Neurospora crassa* (40 Mb; ~10,000 genes) (Galagan et al. 2003). The *S. cerevisiae* genome has a protein-coding content of just under 70% (Dujon 1996), whereas less than 2% of the human genome is protein-coding (Elgar and Vavouri 2008). Only 5.5% of *S. cerevisiae* genes listed in the Saccharomyces Genome Database (SGD) are intron-containing (Nash et al. 2007), and 9 genes have been found to contain more than one intron.

Within the Saccharomycotina, protein sequence divergence indicates that the level of divergence between *S. cerevisiae* and *Yarrowia lipolytica* (the Saccharomycotina species most distantly related to *S. cerevisiae* to be sequenced so far) is equivalent to that between human and the sea squirt *Ciona intestinalis* (Dujon et al. 2004). While the evolutionary distance within the Saccaharomycotina is of the order of that covered by the entire phylum Chordata, the rate of genome rearrangement in yeasts is roughly one-third of that of vertebrates, and is estimated to be roughly 2 rearrangements per million years (Drillon and Fischer 2011). The genomes of the sequenced Saccharomycetaceae in particular show a very high degree of colinearity and retained local gene order (Byrne and Wolfe 2005). These genomic traits not only make it possible to carry out powerful comparative genomic studies on this group of species, but also to harness synteny and gene content information from existing genome annotations to assist in annotating newly sequenced species. When local conservation of gene order between species exists, this information can also be used in conjunction with BLAST searches to improve the existing annotations of these species. This is the principle behind SearchDOGS, new software described in Chapter 2 of this thesis.

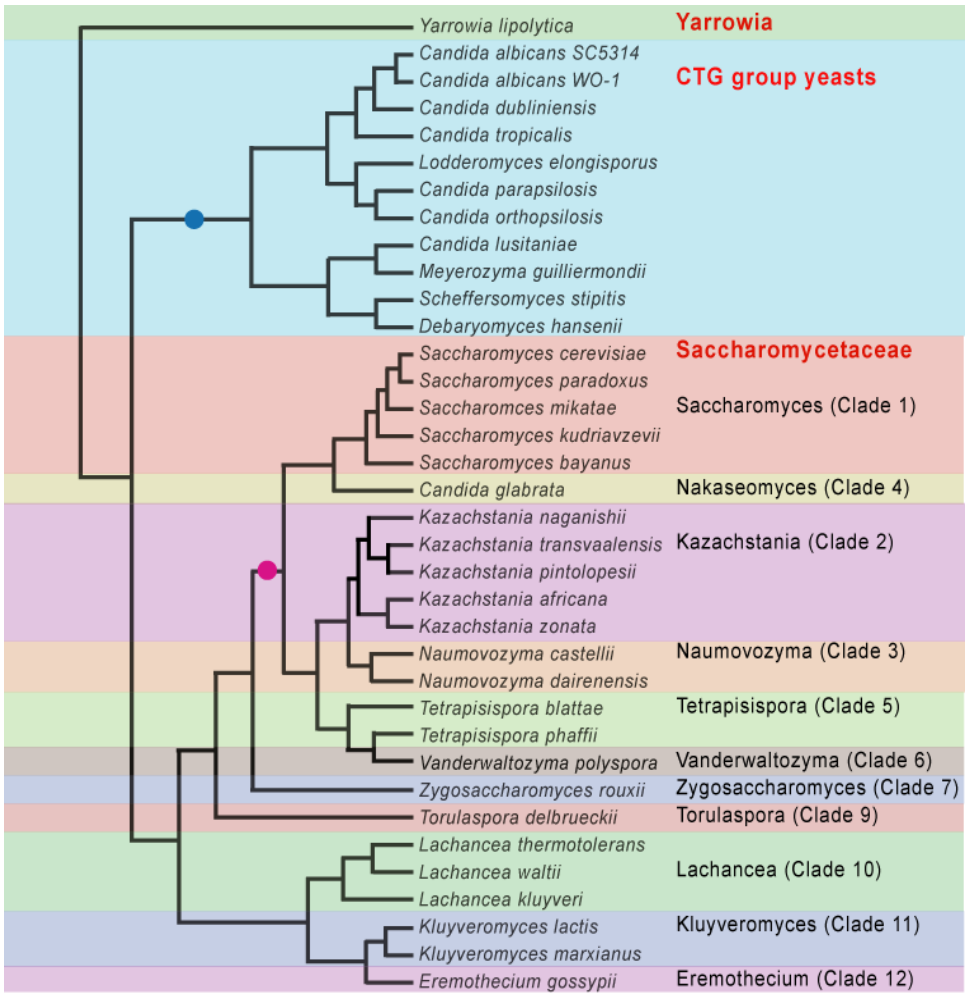

Figure 1.1 Phylogenetic tree of the Saccharomycotina yeast species studied in Chapters 2, 4 and Appendix II of this thesis. The clades of the Saccharomycetaceae family yeasts, which are the focus of these chapters, are shown. The only clade not to have a representative species in the studies in Chapters 4 and Appendix II is the *Zygotorulaspora* clade (clade 8). The pink circle indicates the position of the WGD within the Saccharomycetaceae tree, and the blue circle indicates the position at which CTG codon reassignation occurred in the CTG group yeasts. The tree is based on Kurtzman (2003), Hedke (2006), Fitzpatrick (2006) and Gordon et al, (in preparation; Appendix II), and is not drawn to scale.

Several studies have also used a wide range of *S. cerevisiae* and *S. paradoxus* wild isolates (Liti et al. 2009; Schacherer et al. 2009), as well as *S. cerevisiae* laboratory strains (Schacherer et al. 2007) for population genomics analyses. These studies provide valuable insight on population structures and the extent of geographical isolation that exists between strains in these species, as well as giving an estimate of the amount of genomic variation (insertions and deletions, copy number variation, single nucleotide polymorphisms (SNPs), transposable elements, and protein-coding

gene content) that exists within the population of a single species (Liti and Schacherer 2011).

The Saccharomycotina, and the Saccharomycetaceae species in particular, have both shaped human history and been shaped by it to a unique extent; two separate domestication events having been suggested for *S. cerevisiae* (Fay and Benavides 2005). The Saccharomycetaceae are of considerable economic, medical, and biotechnological importance. Nicknamed "baker's yeast" for its traditional role in bread rising, countless strains and hybrids of *S. cerevisiae* are also used in alcohol production, from beer to sake (Liti et al. 2009; Nakao et al. 2009; Pal et al. 2009). *Candida glabrata* is the second causative agent of human candidiasis (Dujon et al. 2004), and several of the other Saccharomycetaceae can function as opportunistic pathogens in immunocompromised patients (Goldstein and McCusker 2001). Yeasts have massive potential for the production of fuels, drugs and other valuable compounds through techniques such as metabolic engineering (Keasling 2010), as evidenced by the production of the antimalarial drug precursor artemisinic acid using engineered *S. cerevisiae* (Ro et al. 2006). Furthermore, the powerful tool of RNAi may soon be available in yeast, as a functioning RNAi pathway has recently been identified in *N. castellii*, and can be reconstituted in *S. cerevisiae* by the addition of a plasmid bearing *N. castellii* Dicer and Argonaute genes (Drinnenberg et al. 2009).

1.2 Whole Genome Duplication in the Saccharomycetaceae.

The major event in the recent evolutionary history of the Saccharomycetaceae is the occurrence of a whole genome duplication (WGD) approximately 100 million years ago (Wolfe and Shields 1997; Scannell et al. 2006) in the branch leading to the Saccharomyes species, as well as *V. polyspora*, *N. castellii* and *C. glabrata*. This involved the ancestor of these species becoming a short-lived tetraploid through endo-duplication (also known as autopolyploidy; duplication of the full complement of chromosomes in a cell) or fusion with a close relative (allopolyploidy), either at the haploid or the diploid state (Kellis et al. 2004). By analysing the position of *S.*

cerevisiae's 16 centromeres within the context of the sister relationship of chromosomal regions, it can be shown that they form eight pairs, indicating that the WGD was either an autotetraploidy by an eight-chromosome ancestor or an allotetraploidy involving two eight-chromosome ancestors (Wolfe 2006). The lack of evidence for any structural (gene order) differences between the two copies of the ancestral chromosomes at the moment of WGD (Gordon et al. 2009) also indicates that an autotetraploidy event is more likely to have been the cause of WGD.

Polyploidy is a very unstable state for a genome (Mayer and Aguilera 1990; Otto and Whitton 2000), and as a result the polyploid genome rapidly returned to normal ploidy, most likely through a combination of a large number of deletion events and epigenetic silencing (Sankoff et al. 2011). Thus, although there was an initial doubling of gene content in the genome, the need to return to normal ploidy coupled with the high level of gene redundancy has led to the loss of one from each pair of duplicate genes created by WGD (termed ohnologs (Wolfe 2000)). An ancestral genome of ~10,000 genes immediately after whole genome duplication has been reduced in this way to 5,606 genes in *S. cerevisiae*, 1104 of which form 552 ohnolog pairs (Byrne and Wolfe 2005; Oheigeartaigh et al. 2011).

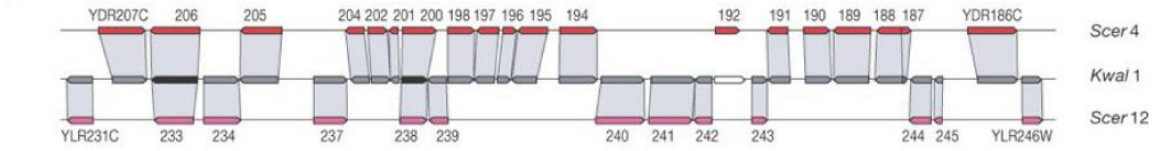

Figure 1.2 *L. waltii* chromosome one (Kwal 1; middle track) has shared synteny blocks with *both S. cerevisiae* chromosome 4 (Scer 4; top track, genes in red) and *S. cerevisiae* chromosome 12 (Scer 12; bottom track, genes in pink). *L. waltii* genes containing only *one S. cerevisiae* ortholog are coloured grey. Instances in which both *S. cerevisiae* ohnologs created by WGD have been retained result in orthology between the *L. waltii* locus and two separate *S. cerevisiae* loci; these *L. waltii* genes are coloured black. From Kellis et al. (2004).

While post-WGD genomes contain on average only 10% more genes than their non-WGD counterparts, the whole genome duplication led to a doubling of chromosome number in the branch on which it occurred, and post-WGD species retain roughly double the number of chromosomes as do non-WGD species (Wolfe and Shields 1997; Kellis et al. 2004). Loss of genes from ohnolog pairs has been shown not to

have a significant bias towards one or other of the duplicate chromosomes that existed immediately post-WGD (Scannell et al. 2007) resulting in an interleaved pattern of gene retention where a full set of genes is retained but divided at random between the two duplicate chromosomes (Kellis et al. 2004). The remaining genes on each duplicate chromosome retain the same order as the ancestor. As a result, each chromosomal region in a non-WGD species shares synteny with regions on two different chromosomes in a related post-WGD species (Figure 1.2).

Why might an event as traumatic as a whole genome duplication become fixed in the population? A WGD event burdens an organism with an increased metabolic load (Wagner 2005; Gerstein et al. 2006) and reproductive isolation (Greig et al. 2002a; Greig et al. 2002b) but can have major benefits from an evolutionary standpoint. Unlike a small-scale duplication, a WGD allows entire metabolic pathways to be duplicated and adapted, such as in the example of the glycolytic pathway in *S. cerevisiae* (Conant and Wolfe 2007). Six genes involved in the ten reactions of glycolysis were retained in duplicate following the WGD, and, coupled with several adaptations that both pre- and post-dated the WGD, appear to have allowed *S. cerevisae* to increase glycolytic flux in response to high glucose concentrations (Conant and Wolfe 2006; Conant and Wolfe 2007). Given that the WGD roughly coincided with a massive radiation of fruit-bearing angiosperms in the mid-Cretaceous (Wang et al. 2009), it may be that the WGD allowed the ancestor of *Saccharomyces* and the other post-WGD genera the opportunity to develop a competitive advantage, rapidly adapting and subfunctionalizing its glycolytic pathway in response to a large increase in available glucose in its environment (Conant and Wolfe 2007; Merico et al. 2007). The results of these adaptations were not only to increase the amount of glucose *S. cerevisiae* can metabolise in a high-glucose environment, but also to poison its competitors by the production of alcohol (known as the Crabtree effect) (Piskur et al. 2006).

The WGD is also likely to have contributed to rapid speciation following the event, through a mechanism in which reproductive isolation was quick to develop between polypoid strains that lost different members of duplicate genes (Scannell et al. 2006). Speciation between the lineages leading to *V. polyspora* and *S. cerevisiae* occurred shortly after the WGD, at a time when the common ancestor of the lineages still retained over 9,000 genes (Scannell et al. 2007). While almost half of the surviving single copy genes in present-day *V. polyspora* and *S. cerevisiae* are paralogs, similar levels of gene loss following their speciation have resulted in both species converging on very similar genome sizes (5510 genes and 5606 genes respectively).

What happens to individual genes following whole genome duplication? In the majority of cases, retention of two copies of a gene serves no purpose, or may actually be deleterious to the cell if it produces energetically wasteful excess gene products or interferes with stochiometric ratios (Hurles 2004; Sankoff et al. 2011). Thus, while one copy remains highly conserved, typically the other copy is not under selection to be retained, or there may even be positive selection for disruption or silencing of this copy. Through a combination of epigenetic silencing and the introduction of missense or nonsense mutations (Eckardt 2001; Sankoff et al. 2011), the majority of genes become non-functional and are eventually lost from the genome. An ohnolog pair may be retained if increased dosage for that gene is advantageous to the cell (Sugino and Innan 2006). It is also possible that a gene might become subfunctionalized, i.e. that complementary mutations in both copies of an ohnolog pair might result in the functions of the product of the ancestral gene being split between the ohnolog pair such that both copies are essential (Force et al. 1999; Lynch and Force 2000). Alternatively, duplication of a gene or set of functionally related genes may allow one or both ohnologs/sets of ohnologs to adapt so that the two genes/sets of genes are optimised for different conditions, such as in the example of the *S. cerevisiae* glycolytic pathway described above. This is classed as subfunctionalization as well, although it has elements of neofunctionalization (see below) to it. This type of ohnolog pair adaptation can also enhance regulatory control (Gu et al. 2005). Finally, in some cases the redundant ohnolog copy that is free to

evolve may develop a new function that is advantageous to the cell, while the other ohnolog retains all the ancestral functions, a principle first proposed by Ohno (1970) and now termed neofunctionalization. Byrne and Wolfe (2007) found that 56% of ohnolog pairs retained in *S. cerevisiae*, *C. glabrata* and *N. castellii* showed significantly asymmetric protein sequence evolution indicative of neofunctionalization, indicating that this was a major driving force in adaptation and diversification among the post-WGD species.

1.3 The YGOB browser

The Yeast Gene Order Browser (YGOB), hosted by Wolfe laboratory (http://wolfe.gen.tcd.ie/ygob), is an online tool for the visual display of annotated Saccharomycetaceae family genomes (Byrne and Wolfe 2005). It currently contains five post-WGD genomes (*Vanderwaltozyma polyspora* (formerly called *Kluyveromyces polysporus*), *Naumovozyma castellii* (formerly *Saccharomyces castellii*), *Candida glabrata** (*despite the name, *Candida glabrata* clades with the Saccharomycetaceae and is closely related to the Saccharomyces species (Fitzpatrick et al. 2006)), *S. bayanus*, and *S. cerevisiae* as well as six non-WGD genomes (*Zygosaccharomyces rouxii*, *Kluyveromyces lactis*, *Eremothecium gossypii* (formerly called *Ashbya gosyipii*), *Lachancea kluyveri* (formerly *Saccharomyces kluyveri*), *Lachancea thermotolerans*, *Lachancea waltii* (both formerly classed as *Kluyveromyces* species). The extensive renaming of Saccharomycetaceae species reflects our growing knowledge of the correct phylogenetic positioning of these species as more genomes are sequenced and more comprehensive gene sequence comparisons are carried out. The current species naming is based on recent work by Kurtzman and Robnett (2003), Hedtke and colleagues (2006) and others.

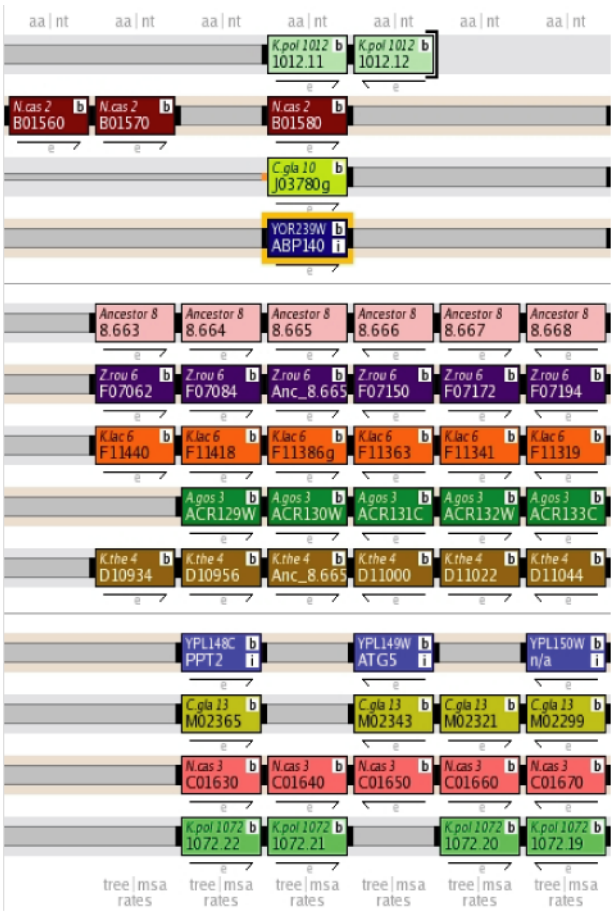

Figure 1.3 Yeast Gene Order Browser (YGOB) screenshot of the genomic context of *S. cerevisiae ABP140* (highlighted in yellow). Here the *S. cerevisiae* genome (dark and light blue) is the focus, and the other genomes are rearranged to match the *S. cerevisiae* gene order. Each post-WGD species (here *V. polyspora, N. castellii, C.glabrata and S.cerevisiae*) have two tracks (top and bottom) representing two regions of synteny with a single region in the non-WGD species in the middle (*Z. rouxii, K. lactis, E. gossyppii, L. thermotolerans* and the Ancestral Gene Order). This screenshot represents a selection of the genomes currently available in YGOB; more species can be switched "on" or "off".

YGOB allows the visual comparison of regions of local synteny between the included species [Figure 1.3]. Each gene is represented as a standard-sized icon that is not representative of the actual size of a gene, and the size of intergenic regions represents the presence/absence of annotated orthologs in a given species rather than actual size. When viewing YGOB a gene from a single species is the focus of attention, and all other species are reordered to represent their synteny with the focal species. Due to six of the species having undergone WGD relative to the others, two syntenic tracks are displayed for each of the post-WGD species, illustrating that each region of a non-WGD genome is syntenic with two regions of a post-WGD genome. YGOB calculates and makes it possible to visualise whether a gene in a species is present in its expected syntenic context or absent, and thus is a useful tool in identifying instances in which a species appears to lack an ortholog in the expected genomic location. SearchDOGS was designed to work in conjunction with the YGOB data structures to investigate each of these "missing ortholog" gaps, and evaluate whether an unannotated ortholog exists at each of these loci.

YGOB also contains the Ancestral Gene Order (Gordon et al. 2009), a construct that represents the complement and order of genes that was present in the common ancestor of the set of species contained within YGOB. The Ancestral Gene Order was constructed by analysing the distribution of orthologs at each individual locus across the genomes contained within YGOB, as well as the conservation of blocks of synteny (local gene order) between species. Using the principle of parsimony, Gordon and colleagues were able to manually infer the loci at which an ancestral gene was likely to have existed, and the most likely order of genes in this ancestor. The ancestral gene order is a valuable tool for examining the various inter- and intrachromosomal inversions that have taken place along the lineages leading to the modern-day species featured in YGOB. It has also been adopted as a "gold standard" for some researchers developing computational tools to automatically reconstruct ancestral gene orders for groups of related species (Zheng et al. 2008).

1.4. Recent sequencing and annotation of Saccharomycetaceae yeasts.

The Wolfe Laboratory has recently sequenced five new post-WGD Saccharomycetaceae species (*Kazachstania africana, Kazachstania naganishii Naumovozyma dairenensis, Tetrapisispora blattae* and *Tetrapisispora phaffii*) as well as the non-WGD species *Torulaspora delbrueckii* (Gordon et al, in preparation; Appendix II) (Figure 1.1). In addition, *Naumovozyma castellii* was resequenced to a higher quality than the previous draft sequence (Cliften et al. 2003). Sequencing of these species was carried out using the Roche/454 "next generation" sequencing (NGS) platform (Droege and Hill 2008) (NGS technologies described briefly in Section 1.5), and a *de novo* assembly. The Roche/454 platform is based on pyrosequencing technology (Margulies et al. 2005), and several features of the technology, including slightly longer reads relative to other NGS platforms, make it an appropriate platform for the sequencing of genomes *de novo* (i.e. without a reference sequence) (Horner et al. 2010).

Automated annotation of these new genomes was carried out using the Yeast Genome Annotation Pipeline (YGAP), recently developed by the Wolfe Laboratory (Proux-Wéra et al, in preparation), and soon to be available for public use. YGAP is based on the principle that by comparing against the genomes currently contained in the YGOB browser, we should be able to infer the existence and order of the large majority of genes in a related genome to be annotated, as well as telling us what intron/exon structure to expect. It combines homology and synteny information from these species to identify the location of genes. An initial protein-coding gene-finding step uses the protein sequences of genes in YGOB pillars in TBLASTN searches against the genome sequence to be annotated to identify the location of homologs in this species. In order to identify small and highly diverged genes that may have been missed in this step, a highly sensitive synteny-based gene-finding program called SearchDOGS (described in Chapter 2) is run. Following these steps, species-specific singletons of greater than 150 amino acids are identified using an *ab initio* ORF-finding procedure. YGAP also annotates tRNA genes using tRNAscan-SE (Lowe and Eddy 1997) and flags retrotransposons. The completed genome annotation is viewable in an automatically generated "Mini-YGOB" browser viewable only to the user and containing the annotated genome along with the genomes of *S. cerevisiae* and *E. gossypii* as well as the Ancestral gene order (Gordon et al. 2009) as references.

The species to be sequenced were chosen in order to get a comprehensive sampling of the species clades contained within the Saccharomycetaceae for the purposes of comparative genomic analysis. With the inclusion of the six recently sequenced genomes, 11 of the 12 clades currently described for the Saccharomycetaceae family yeasts (Kurtzman 2011) will have at least one representative species in YGOB (Figure 1.1). These recent additions to our annotated species set allowed for a detailed study of the evolution of the mating-type locus in the family Saccharomycetaceae, presented in Appendix II.

1.5 Next generation sequencing

The first DNA-based genome to be sequenced was that of the bacteriophage φX174 in 1977 (Sanger et al. 1977a). From then until recently, various adaptations of this pioneering method (Sanger et al. 1977b) known as Sanger sequencing have been used for the majority of all DNA sequencing (Kircher and Kelso 2010) However, in recent years, a number of new massively parallel sequencing technologies have come onto the market. These technologies, which have been developed as a result of advances in optical devices, the field of nanotechnology, and innovations in the application of more traditional techniques in molecular biology, are characterised by the production of large numbers of short reads per instrument run (in some cases over 1 billion) (Horner et al. 2010; Metzker 2010). They include the Roche/454 platform (Droege and Hill 2008) the Illumina Genome Analyser (Bennett 2004), Applied Biosystems SOLiD (Porreca et al. 2006) and the Helicos Heliscope sequencer (Harris et al. 2008). A detailed review of these methods is beyond the scope of this introduction. but for thorough discussion of the techniques involved, comparisons of the various technologies currently available and the challenges of managing and assembling vast amounts of short reads into complete genomic sequences, I recommend reviews by Horner et al (2010), Metzker (2010) and Kircher & Kelso (2010).

In short, these technologies, termed "next generation sequencing (NGS) technologies", can outperform traditional Sanger methods in daily throughput by a factor of 100-1,000, and at a cost of 4%-0.1% per million base pairs compared to Sanger methods (Kircher and Kelso 2010). In addition to *de novo* sequencing and genome resequencing, these techniques can also be applied to gene expression studies, with sequencing of cDNA providing a comprehensive snapshot of the transcriptome, as well as a number of other applications including the characterisation of small RNA populations and epigenetic studies (Morrissy et al. 2009; Horner et al. 2010).

Refinements in the current NGS techniques are resulting in a steady improvement in throughput and accuracy (Kircher and Kelso 2010). However, upcoming technologies may soon supercede the current "next generation" platforms; these include methods based on nanopore technology such as Oxford Nanopore's BASE platform (Clarke et al. 2009) and IBM's proposed silicon-based nanopore technology (IBM 2009), as well as methods based on tunnelling electron microscopy (Horner et al. 2010). Pacific Biosciences have also recently released a powerful SMRT (*S*ingle *M*olecule *R*eal *T*ime) sequencing platform, producing read lengths of over 1MB and involving sequencing runs that can take as little as 30 minutes of instrument time (Korlach et al.). Motivated by the potential for personalised medicine, NIH/NHGRI have set a much-publicised goal of a $1,000 genome (http://grants.nih.gov/grants/guide/rfa-files/rfa-hg-09-011.html) (Bennett et al. 2005; Mardis 2006; Wolinsky 2007).

As a result of these advances, the rate at which genomes are being sequenced is increasing exponentially. The first free-living organism to be fully sequenced was the bacterium *Haemophilus influenzae* in 1995 (Fleischmann et al. 1995), and the first completely sequenced eukaryotic genome followed in 1996. By 2007/8, roughly 41 archaeal, 468 bacterial and 49 eukaryotic had been sequenced (Ansong et al. 2008a). In September 2011 Entrez Genome (http://www.ncbi.nlm.nih.gov) listed 1750 complete bacterial genomes and another 5230 in progress, an increase of 293 complete and 1313 incomplete sequencing and annotations projects from Feburary 2011 alone. This highlights the ever-increasing need for fast, accurate and readily available annotation tools in order to avoid a) a bottleneck at the step of annotations of these new genomes; and b) inaccurate or incomplete genome annotations adding misinformation to the databases upon which comparative genomic approaches are heavily reliant.

In the coming sections, I give an overview of currently used automated annotation methods and outline the problems and challenges that exist for the accurate annotation of genomes. Chapter 2 describes the development of software named SearchDOGS that is designed to find unannotated genes in existing annotations in

yeasts, and Chapter 3 describes the adaptation of this software into a tool for improving the annotations of bacterial genomes. As a standalone, freely downloadable package, it has the advantage that it can be used with both published and unpublished or confidential genomic data.

1.6 Automated annotation methods in yeasts and prokaryotes

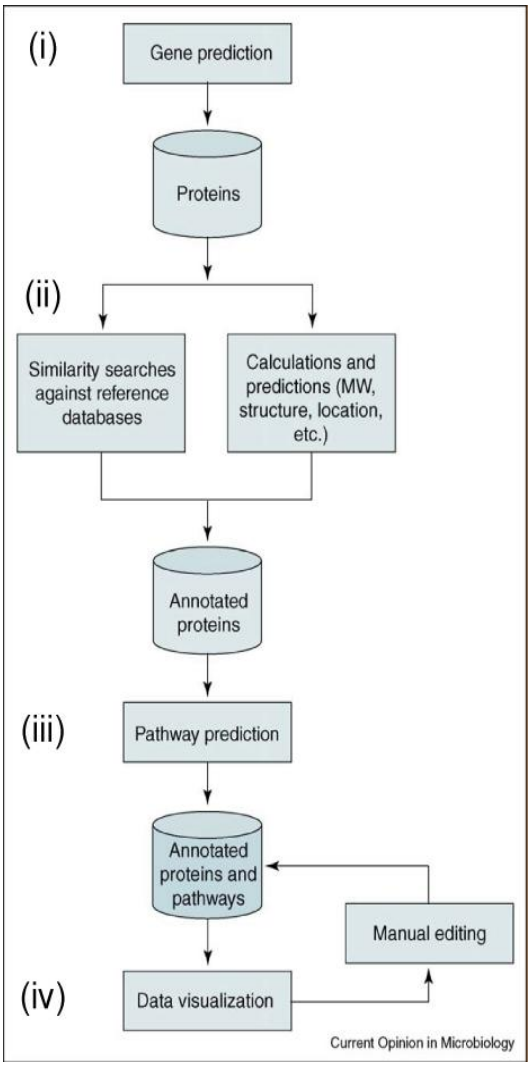

Figure 1.4 Flowchart of a typical annotation pipeline. (i) Composition-based gene finders are used to generate an initial candidate gene set for the genome to be annotated. (ii) The protein sequences of these candidate genes are searched against reference databases to identify homologs, and additional information is generated for the proteins. (iii) Metabolic networks are reconstructed for the genome. (iv) A visual interface is provided for the user in order to facilitate manual curation of the automated results. Flowchart is adapted from Stothard and Wishart (2006).

Virtually all current genome annotation projects make use of automated annotation programs and pipelines, although the need for manual intervention and correction still exists. A typical annotation platform involves the steps shown in Figure 1.4. For *de novo* annotation, the starting point is a genome sequence in chromosomal/scaffold/contig form.

The initial stage involves protein-coding gene prediction usually using a combination of programs that fall roughly into two categories: a) composition-based and b) homology-based gene prediction. Composition-based gene finders (also sometimes called "intrinsic" or *ab initio* gene finders) use mathematical tools such as hidden Markov models and the Viterbi algorithm among others to identify the location of genes based on properties of the nucleotide sequence that correspond to the likelihood of the existence of a gene (Besemer and Borodovsky 2005). They make no explicit use of protein or DNA information outside the sequence being studied. Examples of commonly used gene-finding tools of this class include Glimmer (Delcher et al. 1999) and CRITICA (Badger and Olsen 1999) for prokaryotes, GENSCAN (Burge and Karlin

1997) and GeneID (Parra et al. 2000) for eukaryotes, and GeneMark (Besemer and Borodovsky 2005), which has been adapted for both prokaryotes and eukaryotes. Sequence similarity-based gene finders use sequence similarity searches such as BLAST (Altschul et al. 1997) or HMMer (Finn et al.) to find protein-coding regions, typically searching sequences in the genome to be annotated against large protein databases in order to identify homology to annotated genes in other species.

Protein-coding gene annotation is accompanied by the annotation of tRNA genes by programs such as tRNAscan-SE (Lowe and Eddy 1997) and sometimes programs to detect other noncoding RNA genes and genomic features (Achaz et al. 2007; Lagesen et al. 2007; Langille et al. 2008; Gardner et al. 2009). Predictions of protein function are inferred from homology to similar genes in other species with assigned functions. Often platforms include tools to provide a host of other information about protein-coding genes, such as chemical properties of the protein (e.g. isoelectric point and molecular mass), localisation within the cell and modular structure of the protein (Medigue and Moszer 2007). It is also becoming increasingly common to include software to reconstruct metabolic networks in the annotated genome (Maltsev et al. 2006; Reed et al. 2006; Schneider et al. 2010). The last step is to create a visual interface in order to present this information to the user and to allow manual modification.

Annotation pipelines often integrate results from a number of protein-coding gene prediction programs to generate a results set that is as complete and accurate as possible. A typical approach, such as that used for bacterial annotation platform Basys (Van Domselaar et al. 2005) is to generate an initial set of candidate genes using a composition-based program. These genes are then searched against protein databases such as UniProt (Apweiler et al. 2004) to identify homologs. The protein translations of the intergenic regions are sometimes also searched against protein databases in a subsequent step to identify missing genes from the initial annotation set (Van Domselaar et al. 2005). Comparative genomic annotation pipelines can attack the annotation challenge in the reverse direction, using the set of annotated

genes in related genomes in TBLASTN searches to identify genes that are likely to exist in the genome to be annotated (Parra et al. 2007). Another comparative approach is to compare two unannotated genomes, and thus identify regions of conserved sequence that are likely to correspond to coding exons; this approach is the basis for the TWINSCAN (Korf et al. 2001) and SGP2 (Parra et al. 2003) programs. Other platforms, such as microbial pipelines PhydBac (Enault et al. 2005) and MaGe (Vallenet et al. 2006), use comparative genomic approaches to more accurately assign function to predicted genes.

Platforms have also been developed that integrate EST (*E*xpressed *S*equence *T*ag) datasets for the species to be annotated in order to improve their gene prediction accuracy; examples include eukaryotic pipelines AUGUSTUS (Stanke et al. 2008), N-SCAN/EST (Wei and Brent 2006) and EuGéne-M (Foissac and Schiex 2005). Manual annotation tools such as Artemis (Rutherford et al. 2000) can be used to curate the results of these pipelines.

In addition to those already mentioned, some examples of commonly used microbial annotation pipelines are Microscope (which incorporates MaGe) (Vallenet et al. 2009), AGES (Kumar et al. 2011), AGMIAL (Bryson et al. 2006), AmiGene (Bocs et al. 2003) and the Microbial Annotation Pipeline of the Integrated Microbial Genomes system (Markowitz et al. 2010). Tools used for eukaryotic annotation include the Ensembl pipeline (Curwen et al. 2004), CEGMA (Parra et al. 2007) and tools provided at the UCSC genome browser (Karolchik et al. 2008) and at NCBI (Wheeler et al. 2008). For yeasts, the RAPYD platform has recently been made available (Schneider et al. 2010), and the Yeast Genome Annotation Pipeline (YGAP) is soon to be released (Proux-Wéra et al, in preparation).

1.7 Problems and sources of error in genome annotation.

Genome annotation is fraught with a multitude of potential sources of error. Here, I divide errors in protein-coding gene annotation into the following categories:
Primary sources of error: Errors in the sequencing of a genome.
Secondary sources: Errors in the identification of protein-coding genes. Identifying promoter regions and other such features represents an even more difficult challenge, but these topics are outside the scope of this introduction.
Tertiary sources: Problems associated with gene information contained within genome databases.

Primary sources of annotation error

The first stage at which error can creep into a genome annotation is at the sequencing stage. Error profiles vary between sequencing technologies, but most often errors come in the form of a single base pair insertion/deletion/substitution, although larger insertion/deletion mistakes can be made in genomic regions containing a lot of repetitive DNA sequence (Kircher and Kelso 2010; Quinlan et al. 2008; Wicker et al. 2006). While next generation sequencing techniques significantly outperform Sanger methods in throughput and can provide genome sequences at a much lower cost (Kircher and Kelso 2010), the error rate associated is generally higher than the rate of 10^{-4} to 10^{-5} (i.e. one error every 10,000 to 100,000 bases) associated with Sanger technologies (Ewing and Green 1998). Error rates associated with Roche/454 sequencing are 10^{-3} to 10^{-4} for substitution errors (Margulies et al. 2005; Quinlan et al. 2008), and small indel errors are relatively common (Quinlan et al. 2008). Average error rates for Illumina, Applied Biosystems and Helicos are higher, ranging from 10^{-2}-10^{-3} for Illumina and Applied Biosystems to a few percent for Helicos (Kircher and Kelso 2010). Most errors can be resolved by using a high enough genome coverage, although some consistent miscalls have been reported for homopolymers over 10nt in length using Roche/454 (Wicker et al. 2006; Green et al.

2008; Quinlan et al. 2008). However, a higher coverage means more expense, and so the quality of sequencing of genomes varies considerably. Higher levels of sequence error in lower coverage genome sequences result in erroneous stop codons or frameshifts being introduced into gene sequences. Where full-length homologs exist in related species for comparison, differentiating these sequencing errors from genuine species-specific gene truncation events can be extremely difficult.

Due to the cost involved in taking a genome from the "draft" stage to the "finished" sequence stage, many genomes, including the majority of eukaryotic genomes, are released in "draft" form, and may never be "finished" (Salzberg 2007). As these genomes are split over many contigs, gene annotation is complicated by instances of genes being split over two or more contigs, and this can result in the annotation of fragments of genes, genes annotated twice, or genes annotated with either duplicated or missing regions due to contigs being incorrectly aligned.

One last factor to consider, although it does not represent a sequencing error, is that sometimes the strain of a species sequenced will contain substitutions or indels specific to the sequence of that laboratory strain. In this instance, conclusions about a particular gene inactivation in the annotated representative may not hold for the population in general For example, the commonly studied S288c lab strain of *S. cerevisiae* has a truncated allele of the gene *SAL1* which, in conjunction with rare alleles of three other genes, contributes to a high rate of spontaneous mitochondrial instability relative to natural isolate strains (Dimitrov et al. 2009). Furthermore, sequencing of a number of *S. cerevisiae* isolates from different sources identified 38 ORFs in different individuals that are missing from the reference S288c genome and are likely to represent real genes (Liti et al. 2009).

Secondary sources of annotation error

False positive versus false negative gene calls

There are several categories of error that affect automated gene annotations in both prokaryotes and eukaryotes. The first is the rate of false positive annotations (annotation of spurious features) and false negatives (genuine genes missed). Annotation pipelines often use a number of *ab initio* and homology based programs in order to reduce false positives by using a combined results set (Medigue and Moszer 2007; Parra et al. 2007), and reduce false negatives by providing additional support for "borderline cases". However a multitude of studies have shown that a large number of false positives tend to make it into the annotation, and a smaller but nonetheless significant number of false negatives are missed (Skovgaard et al. 2001; Nielsen and Krogh 2005; Castellana et al. 2008; de Groot et al. 2009; Gallien et al. 2009; Payne et al. 2010).

Start and stop coordinates

Accurate annotation of gene coordinates, particularly of start coordinates, is difficult to automate. For a large proportion of genes the programs must choose between a selection of start codons to create genes of different lengths. Furthermore, while nearly all genes in eukaryotes begin with a starting ATG, prokaryotic genes commonly use the alternative start codons GTG and TTG (7.7% and 1.7% of protein coding genes in *E. coli* K12 respectively), with rare instances of genes beginning with ATT, CTG and ATC (two known cases each of ATT and CTG starts codons respectively, and one of an ATC start codon in *E. coli*) (Riley et al. 2006). This means that prokaryotic gene start prediction is particularly problematic, as in many cases there is a large selection of start codons that must be differentiated between. Nielsen and Krogh compared the annotations of 143 prokaryotic genomes to results

generated using their Easygene software, and found that in some genomes up to 60% of genes may have incorrectly annotated start codons, with an overly strong preference being shown for the most upstream start codon (Nielsen and Krogh 2005). While this estimate is of course dependent on the accuracy of the Easygene algorithm, it is indicative of the level of disagreement that tends to exist between gene-calling methods. A comparison of automated genome annotations for *Halorhabdus utahensis* obtained using three different annotation systems showed that while the same stop codon is identified for genes in over 90% of cases, the three programs choose the same start and stop codons in only 48% of cases (Bakke et al. 2009). Incorrectly annotated start codons result in improperly defined upstream intergenic regions, meaning that promoter regions and regulatory binding sites can be missed, as well as the cellular localisation signals that are typically located in the N-terminal of a protein.

In eukaryotes, the assumption that translation of a protein-coding gene begins at a starting AUG may also be erroneous, as may be the assumption that a single start codon is used in all cases of translation of a gene. A recent ribosomal profiling study in mouse embryonic stem cells by Ingolia et al. (2011) found that the majority of a set of 5,000 well-expressed transcripts contained more than one detectable site of translation, with a significant proportion beginning from noncanonical AUG-like codons.

Stop codons are predicted with a much higher degree of accuracy than start codons in prokaryotes. However, there may be many instances in which stop codons are read through in both prokaryotes and eukaryotes. Translational stop codon readthrough is known to be common in viral genomes (Namy and Rousset 2011), but was thought until recently to be very uncommon in eukaryotes. However, using both protein sequence conservation and deep RNA sequence data, Jungweis et al. recently reported a set of 283 candidates for translational readthrough in Drosophila.
In one specific class of proteins called selenoproteins, the TGA stop codon codes for selenocysteine and is read through. Tools have been developed to facilitate

characterisation of selenoproteins in eukaryotes (Castellano et al. 2001; Kryukov et al. 2003; Castellano et al. 2004) but to date not in prokaryotes (Ansong et al. 2008a), and as a result truncated selenoprotein misannotations are likely to exist in the databases.

Intron and splice sites

Prediction of correct coordinates is complicated significantly by the frequency of introns in eukaryotic genomes. While consensus intron splice and branch sites are well-characterised in some species including *S. cerevisiae* (Spingola et al. 1999), the frequency of deviation from the strict consensus sequences and locations of these sequences makes their accurate identification difficult without manual curation. In principle the nucleotide sequence characteristics employed by *ab initio* composition-based programs should differentiate to some extent between the coding DNA of exons and the non-coding introns, and homology-based programs should be "primed" to look for introns based on the intron locations in related homologs (the latter, of course, requires that intron coordinates be correctly predicted in these homologs). Homology-based approaches can often predict intron location by identifying gaps in the sequence alignment when candidate genes are aligned to their homologs; other approaches such as eukaryotic annotation tool AUGUSTUS integrate EST data wherever available to improve the accuracy of intron prediction (Stanke et al. 2008), whereas CEGMA attempts to identify the exon-intron structure of a core set of genes in eukaryotic genomes by using highly curated information from the core genes of six model organisms (Parra et al. 2007). Nonetheless, introns coordinates are often predicted with low accuracy, and are sometimes either missed altogether (Allen et al. 2004; Castellana et al. 2008) or are overpredicted (Jeffries et al. 2007). A missed intron can result in a spuriously truncated protein call if readthrough of the intronic region hits a stop codon, an apparent frameshift if the exons are in different frames, or the insertion of a chunk of spurious protein-coding sequence if exons are in the same frame.

Going from gene sequence to protein sequence is less than straightforward in eukaryotes. It is estimated that up to 70% of human genes are subject to alternative splicing (Brett et al. 2002; Kriventseva et al. 2003; Stamm et al. 2005) meaning that a single gene can produce proteins with different enzymatic activities, binding properties and cellular localisation (Stamm et al. 2005). Some proteins are the result of *trans*-splicing, meaning that they are the result of mRNAs from two or more different genes (Eul et al. 1995). Detection of alternatively spliced gene structures using *ab initio* and sequence-similarity methods is problematic and in need of refinement, as described by Foissac and Schiex (2005). Pipelines such as AUGUSTUS report rates of correct prediction of "at least one splice form" of 57-77% by combining multiple sources of information incuding gene and transcript annotations from related species syntenically mapped to the target genome, evolutionary conservation of DNA, mRNA and ESTs of the target species, and retroposed genes (Stanke et al. 2008).

Overlapping ORFs

In most instances in prokaryotes and eukaryotes in which two adjacent ORFs overlap by more than a few base pairs, only one is considered likely to code for a real gene. Automated annotation programs frequently have difficulty in choosing the correct ORF. In cases where an ORF is mostly overlapped by an ORF in another reading frame, some annotation pipelines will always choose to annotate the larger ORF at the expense of the smaller; however a proteomics study in *Yersinia pestis* found that in 25% of cases where a smaller ORF was completely overlapped by a larger one, the smaller ORF was the true gene (Payne et al. 2010). In many instances the overlap itself is likely to be false, especially in prokaryotes due to the difficulty in predicting gene start codons described above; an investigation of overlapping ORFs by Palleja et al. (2008) in prokaryotes identified over 900 cases in which reported overlaps over 60bp in length represented annotation errors rather than real overlaps. This is a likely result of automated preferences for the most upstream start codon leading to gene predictions that are overly long in the 5' direction (Nielsen and Krogh 2005).

In some cases both overlapping gene candidates may be "true" genes. Chung et al. identified 40 candidate genes with evolutionarily conserved overlapping coding regions, and consider this a conservative estimate of the number of genuine overlapping genes that may exist in the human genome (Chung et al. 2007).

Differentiating pseudogenes from true genes.

See Section 1.9.

Identification of short and highly diverged genes

See Section 1.10.

Unusual gene features

Genomes also feature occasional gene structures that are unlikely to be identified without manual curation by someone with knowledge of the type of gene and the biology of the species in question. One such example is genes containing a programmed ribosomal frameshift. In these genes, during translation the ribosome "skips" forwards or slides backwards one base pair (or occasionally two), thus jumping to another frame (Farabaugh 2010). In the example of the *E. coli* gene *dnaX*, this allows one gene to code for two components of the DNA polymerase III complex by producing equimolar amounts of a full length and a shorter, frameshifted gene product (Larsen et al. 1997). Programmed frameshifting is common in viruses and transposable elements (Atkins 2010), and four examples to date have been identified in *S. cerevisiae* (described in detail in Chapter 4). There are also instances of genes containing sequences that stimulate the ribosome to dissociate from and reassociate with the mRNA at a downstream location, leading to up to 50bp segments of the mRNA transcript being bypassed during transcription (Herr et al. 2000; Maldonado and Herr 1998).

Orthology versus paralogy

Once homology between a newly annotated gene and genes in other species has been identified, it is important to be able to ascertain whether the relationship is one of orthology or paralogy. This is significant to functional inference as orthologs generally have conserved functions whereas paralogs often diverge (Studer and Robinson-Rechavi 2009). Synteny information can be used to indicate whether a pair of genes are orthologs or paralogs, and this type of information is beginning to be integrated into annotation platforms (Vallenet et al. 2006; Stanke et al. 2008) (Proux-Wéra et al, in preparation).

Prokaryotes: Phylogenetic position of the genome to be annotated

The first prokaryotic genomes to be sequenced belonged to the proteobacteria (Fleischmann et al. 1995; Fraser et al. 1995; Cole et al. 1998); consequently the majority of current gene prediction algorithms have been "trained" on proteobacterial datasets. While these algorithms provide relatively accurate and robust gene calls in proteobacterial genomes, their accuracy can be expected to decrease as more distantly related prokaryotes are sequenced (Ansong et al. 2008a). As previously noted, the accuracy of gene start predictions may be especially poor in GC-rich genomes; using the prokaryotic genefinder Easygene, Nielsen and Krogh predicted the accuracy of start site prediction in the annotation of *Aeropyrum pernix* at 41%, a genome with a GC content of 56.3% (Nielsen and Krogh 2005). In eukaryotes, as the noncoding:coding ratio and fraction of genes featuring introns and alternative splicing increases, the accuracy of automated gene-calling tools can be expected to decrease.

Tertiary sources of annotation error

The third category of error source is that of errors drawn from the information currently stored in databases and available genome annotations. Homology-based gene calling methods are reliant on the quality of annotation of the homologs they compare ORFs against. A homolog misannotated with an overly long N terminal due to the choice of an incorrect start codon may lead to the rejection as truncated pseudogenes of genuine homologs in other species (see Section 1.9). If an intron or a splice site has not been identified in an annotated homolog at a given locus, is it likely to escape detection in future annotations of homologs. Indeed the gene itself needs to have been identified for it to help identify future homologs; a "missing gene" in a related species could provide false evidence in favour of rejecting a candidate ORF in a closely related species. Conversely, the annotation of spurious ORFs and and non-functional gene fragments as genuine genes in some genomes has led to automated methods spreading "false genes" through further genome annotations due

to the support of homology for *ab initio* false predictions (Section 1.9). This is also an issue with the transferral of false protein function predictions across homologs (Bork and Bairoch 1996; Salzberg 2007).

More recent prokaryotic genome annotations carried out by pipelines combining *ab initio*-, homology-, and sometimes synteny-based gene-calling approaches are considerably more accurate than older annotations (Poptsova and Gogarten 2010; Salzberg 2007; Vallenet et al. 2006); however the older and less accurate annotations remain in the databases and represent potential reservoirs of error for homology-based methods. Furthermore, the increase in accuracy of automated methods may be counterbalanced by the relative decrease in extensive manual curation that accompanied earlier genome annotation (Bocs et al. 2003).

Another issue, discussed more extensively in Salzberg et al. (2007), is that of how genome annotations are maintained and updated in GenBank/EMBL/DDBJ (Stoesser et al. 1997; Tateno et al. 2002; Benson et al. 2009). A genome annotation deposited in GenBank is the property of the individual/group that submitted it, and cannot be edited by third parties. Sequencing or annotation errors identified by third parties will go uncorrected unless the original "owners" update the GenBank entry. As a result, many genome annotations are inaccurate and outdated. Some databases such as the TIGR Comprehensive Microbial Resource have been developed with the aim of having alternative annotations and reannotations (Peterson et al. 2001); however GenBank and its sister databases remain by far the most widely-used and supported sources of genome annotation information.

1.8 Experimental validation of protein-coding genes and the emerging technique of proteogenomics.

Genome databases are saturated with automatic gene annotations for which no experimental evidence exists. The term "hypothetical" and "conserved hypothetical" refers to predicted genes having no known homologs or homology only to other

hypothetical genes in related species, and typically 30-50% of all genes in a genome fall into this category (Ansong et al. 2008a). Brown and Sjolander estimated in 2006 that only 3% of genes in the UniProt database not labeled as "hypothetical" or "unknown" had experimental support (Brown and Sjolander 2006), and with the rapidly growing number of sequenced and automatically annotated genomes that number is surely lower now. Experimental validation of protein-coding gene predictions takes place at the levels of expression and of translation.

Genome expression studies take the complement of mRNA expressed in a cell under a certain condition or set of conditions and reverse transcribe and sequence each mRNA. In yeast, DNA microarray and SAGE (serial analysis of gene expression) studies were initially used to characterise the expression of the annotated *S. cerevisiae* gene set under a range of environmental conditions (Velculescu et al. 1997; Hughes et al. 2000). These studies have been complemented by a large-scale, cDNA analysis over the entire yeast genome, characterising the full transcriptome and thus allowing the identification of transcripts from hitherto unannotated ORFs (Miura et al. 2006; Nagalakshmi et al. 2008; Yassour et al. 2009). Some annotation platforms such as AUGUSTUS and N-SCAN/EST integrate this expression data to improve the accuracy of gene prediction (Wei and Brent 2006; Stanke et al. 2008). However, low levels of transcription occur over much of entire genomes; 85% of the *S. cerevisiae* genome is expressed in rich media (David et al. 2006). This may make it difficult to unambiguously differentiate the expression of protein-coding genes transcribed at low levels from "background noise" transcription (Royce et al. 2005).

A drawback of these approaches is that evidence of transcription is not unambiguous proof of the functionality of a gene as it does not determine if the expressed gene is translated into a protein (Ansong et al. 2008a). These methods also do not indicate what levels of a given protein are active in a cell. Gene expression levels are not always an accurate predictor of protein abundance as some proteins are targeted for degradation at a much higher rate than others, and the function of a protein can also be affected by post-translational modifications (Ansong et al. 2008a).

Until the advent of high-throughput proteomic techniques, gene annotation techniques based on analysing the protein complement of a cell, such as peptide sequence analysis of individual proteins by Edman degradation (Edman 1949), were costly and labour –intensive. However the current high-throughput liquid chromatography-tandem mass (LC-MS/MS) spectrometry based proteomics approaches make analysis of a genome's worth of proteins feasible and more affordable. In this approach, a mixture of proteins recovered from an organism is digested by proteases, producing a set of peptides that are separated by multidimensional liquid chromatography and then analysed by tandem mass spectrometry (MS/MS) (Washburn et al. 2001; Aebersold and Mann 2003). For each peptide, an MS/MS spectrum of fragment masses unique to that peptide is produced, serving as a "fingerprint" for that peptide. Protein sequences are then identified by comparing these spectra against theoretical masses of possible protein sequences using programs such as SEQUEST (Eng JK 1994), Mascot (Perkins et al. 1999) or X! tandem (Craig and Beavis 2004), or by *de novo* analysis (Washburn et al. 2001; Aebersold and Mann 2003).

It is also possible to search these spectra against the entire six-frame translation of a genome, and in this way directly identify protein-coding genes without relying on predicted gene sequences, a technique first demonstrated by Yates et al. in 1995 (Yates et al. 1995). In 1997 the genome of *H. influenzae* was the first bacterial genome to be curated in this way (Link et al. 1997a). This approach is significantly more computationally intensive than simply using a set of predicted protein sequences, and can become extremely computationally time-consuming for large eukaryotic genomes where the ratio of non-coding to coding DNA is very high; the human proteome has an estimated size of 25 million residues whereas the six-frame translation of the entire genome is estimated at 6 billion residues (Tanner et al. 2007). However application of the approach for the *Arabidopsis thaliana* and human genomes was demonstrated by Kuster et al. (2001) and Choudhary et al. (2001) respectively. Techniques such as those described in Tanner et al. (2007) and Sevinsky

et al. (2008) have been developed to reduce the computational load of applying LC-MS/MS to human genomes.

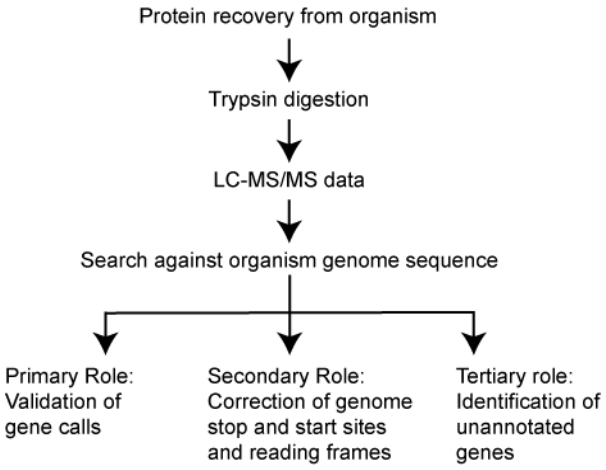

Figure 1.5 Flowchart of a generalised approach to validating a genome annotation using proteogenomics. Following protein extraction from the organism under study, trypic digestion breaks the protein mixture down into a digest mixture, which is then analysed by liquid chromatography-tandem mass spectrometry. The MS/MS peptide spectra produced are then searched against the genome sequence of the organism. Gene annotations are validated and corrected, and unannotated protein-coding genes are identified. Adapted from Ansong et al. (2008a).

The term "proteogenomics", coined by Jaffe et al. in 2004 (2004a) refers to the use of proteomics approaches to improve genome annotation. These approaches have been used to validate protein-coding genes in a growing number of prokaryotic species including *Mycoplasma pneumoniae* (Jaffe et al. 2004a), *Mycobacterium smegmatis* (Wang et al. 2005; Gallien et al. 2009), *Salmonellosis typhimerium* (Adkins et al. 2006), *Yersinia pestis* (Hixson et al. 2006) among others (Jungblut et al. 2001; Lipton et al. 2002; Jaffe et al. 2004b; Elias et al. 2005; Kolker et al. 2005; Ansong et al. 2008b), and a smaller number of eukaryotic species such as *S. cerevisiae* (Oshiro et al. 2002), *Drosophila melanogaster* (Brunner et al. 2007) *Anopheles gambiae* (Kalume et al. 2005) and human (Tanner et al. 2007). The technique, outlined in Figure 1.5, can be used to tackle a number of the problems previously described (Section 1.7). Proteogenomic analysis of *Arabidopsis* indicated that 13% of its proteome was incomplete or incorrectly annotated in the genome sequence, and identified 498 instances of real genes misannotated as pseudogenes as well as 280 previously unpredicted genes and 498 genes that were in the wrong reading frame or had missing/incomplete exons (Castellana et al. 2008). Proteogenomic annotation of *Deinococcus deserti* identified 11 gene predictions with reversed orientation (de Groot et al. 2009). Erroneously predicted start codons have been corrected (Kalume et al. 2005; Gupta et al. 2007; Rison et al. 2007; Gallien et al.

2009) and the expression of hypothetical and conserved hypothetical proteins was confirmed in a number of species (Kolker et al. 2004; Elias et al. 2005; Kolker et al. 2005; Adkins et al. 2006; Hixson et al. 2006; Ansong et al. 2008b). While the standard "shotgun" proteomic approach described above is not designed for the characterisation of protein N terminals (and thus start sites), techniques are being developed to specifically identifiy N- and C-terminal peptides and residues (Gevaert et al. 2003; Aivaliotis et al. 2007; Nakazawa et al. 2008). Proteogenomic "updating" of the *Yersinia pestis* KIM genome led to the removal of spurious gene annotations, the identification of the "true" gene in instances of overlapping ORFs, and the correction of gene models in neighbouring genomes (Payne et al. 2010). Tanner et al. identified or confirmed over 40 alternatively spliced genes in the human genome (Tanner et al. 2007). Direct analysis of protein sequence will identify cases of misannotated selenoproteins, where an in-frame TGA codes for a selenocysteine and is read through (Ansong et al. 2008a).

Proteogenomic approaches, while an invaluable tool for improving genome annotations complementing automatic gene annotation tools, do not remove the need to continue developing and refining bioinformatics gene prediction techniques. Current proteogenomic annotation procedures use the proteomics stage to refine an initial genome annotation created by automated annotation programs, and the more accurate these are the more straightforward and unambiguous the process will be. Proteomics can only detect the proteins that are present under a specific environmental condition, or set of conditions. Thus, automated prediction programs may identify genes expressed only under very specific conditions that are not tested in the proteomic analysis. Payne et al. note "As proteogenomics relies on underprediction to find corrections (meaning that we identify a region of the genome which is not predicted to be coding but should be), this tendency diminishes our power for annotation improvement" (Payne et al. 2010).

Lastly, while proteomic analysis is less costly and labour-intensive than previous low-throughput protein analysis techniques, it will not be financially viable to provide

a proteomic annotation with each and every one of the countless genomes currently being sequenced. As an example of cost, the Proteomics International company lists a single 2 dimensional LC-MALDI-TOF proteome mapping experiment (hundreds to thousands of protein IDs) at $4,000; several such experiments are likely to be necessary (http://www.proteomics.com.au/priceList.aspx).

Despite these caveats, as the completeness and accuracy of gene annotations in the databases improves and the experimental support for protein-coding genes increases in proteogenomically annotated or revised genomes, it will allow for an increase in the accuracy of inferences of gene complement, coordinates and functions drawn from related species when annotating a genome using bioinformatics tools. Furthermore, it should become possible to "purge" databases of many of the spurious annotations that may have spread systemically (Salzberg 2007). Therefore, automated annotation approaches are likely to remain frontline tools in genome annotation, refined and complemented wherever viable by proteomic, expression-based, and manual curation-based approaches.

1.9 Pseudogenes in bacteria and yeast

Pseudogenes, the first example of which was described in 1977 (Jacq et al. 1977) are defined as non-functional relatives of known genes that have lost some or all of the functional repertoire of their homolog or are no longer expressed in the cell (Mighell et al. 2000). They can broadly be categorised as processed or non-processed.

Processed (retroposed) pseudogenes are created through the action of retrotransposons, which can spontaneously reverse transcribe the mRNA produced by a functional gene and insert the cDNA produced at random into the genome (Kaessmann et al. 2009). While the protein-coding sequence of these genes is intact, they lack the necessary promoters for expression, and are thus usually 'dead on arrival' (Podlaha and Zhang 2009). although, some examples exist of these

retroposed copies producing functional proteins (Betran et al. 2002), "donating" coding material to existing genes (Baertsch et al. 2008) or producing chimeric genes (Wang et al. 2000).

Non-processed pseudogenes arise when mutation leads to the inactivation of a functional gene. This can often follow the duplication of a gene, where selective constraint is relaxed on one of the initially identical gene copies leading to the accumulation of mutations and eventual disruption of the gene copy (Bailey et al. 1978; Nei and Roychoudhury 1973). Indeed, it is plausible that dosage issues may result in evolutionary pressure to inactivate unwanted functional duplicates. Nonessential genes can also become pseudogenised due to genetic drift or a change in lifestyle (Mira et al. 2001; Nilsson et al. 2005; Kuo et al. 2009).

With no evolutionary pressure to maintain or remove them, pseudogenes have long been viewed as a "paradigm of neutral evolution" (Li et al. 1981). Most pseudogenes can be expected to be degraded beyond recognition over time (Andersson and Andersson 2001), although examples of longlived gene remnants shared between lineages as diverged as human and rodent exist (Zhang et al. 2004).

Recognised pseudogenes are relatively rare in yeast genomes, despite the widespread recent gene loss in any species to have undergone WGD, a process predicted to have been caused primarily by widespread small pseudogenisation events (Byrnes et al. 2006; Sankoff et al. 2011). Two studies identified 221 "disabled ORFs" (ORFs disrupted by a frameshift or premature stop codon) (Harrison et al. 2002) and 124 "gene relics" (defined by the authors as "more highly degenerate remnants of genes") (Lafontaine et al. 2004) in the *S. cerevisiae* genome. The combined number of disabled ORFs and gene relics still corresponds to only 6% of the *S. cerevisiae* protein-coding gene count and could be regarded as surprisingly low given the loss of approximately 5,000 genes over the ~100 million years since WGD. Analyses of the human genome have resulted in estimates ranging from 8,000 to 20,000 potential pseudogenes and "pseudogenic fragments" in a genome containing ~26,000

functional genes (Ohshima et al. 2003; Torrents et al. 2003; Zhang et al. 2003b), and 14,000 potential mouse pseudogenes have been reported (Waterston et al. 2002). Retention of recognisable pseudogenes is dependent on genome mutation rates; higher nucleotide substitution, insertion and deletion rates in the mouse genome relative to human result in greater decay of processed pseudogenes in mouse (Graur et al. 1989; Waterston et al. 2002; Zhang and Gerstein 2004).

Due to the low percentage of non-functional DNA and paucity of gene duplicates (Mira et al. 2001; Rogozin et al. 2002), it was expected that pseudogenes would be very rare in bacterial genomes (Lawrence et al. 2001). Whereas the genome size of *S. cerevisiae* and *Homo sapiens* differ by a factor of 300 despite only a sixfold difference in gene content, genome sizes and gene content are far more tightly linked in bacteria, with a tenfold difference in genome size corresponding to a similar difference in gene number (Mira et al. 2001). This tight linkage has been seen to indicate greater selective pressure to "streamline" bacterial genomes and remove non-functional DNA (Maniloff 1996; Andersson and Kurland 1998), making it unlikely that "excess baggage" such as pseudogenes would be allowed to persist (Lawrence et al. 2001). However, nearly all bacterial genomes studied have been found to contain gene fragments corresponding to full-length genes in related genomes (Andersson and Andersson 2001; Lerat and Ochman 2004; Liu et al. 2004; Lerat and Ochman 2005; Kuo and Ochman 2010), and are massively prevalent in the genomes of bacteria that have recently transitioned from free-living to intracellular (Cole et al. 2001; Toh et al. 2006).

In contrast to eukaryotic genomes, where retrotransposition and duplication of genomic DNA are the two major sources of pseudogenes, major contributions to pseudogene formation in bacteria come from a high frequency of failed horizontal gene transfer events (Liu et al. 2004), and in some species the widespread loss of native single-copy genes. The latter is specific to pathogens; due to the much narrower functional requirements for a pathogenic lifestyle, former free-living prokaryotes that become facultative or obligate pathogens undergo widespread

pseudogenisation of genes (Mira et al. 2001) as part of an overall eventual genome reduction (Andersson and Kurland 1998). An apparent bias towards genome deletions in bacteria (Mira et al. 2001) results in most of these pseudogenes being lost quite rapidly (Kuo and Ochman 2010); this deletional bias has also been proposed as the driving force behind the compaction of bacterial genomes (Mira et al. 2001; Kuo and Ochman 2009). However, a recent study in *Salmonella* suggests that there may be further positive selection driving pseudogene removal, contradicting the traditional view that these apparently functionless regions are subject to purely neutral evolution in bacteria (Kuo and Ochman 2010).

Pseudogenes can be of great use in evolutionary analysis, particularly in the case of processed pseudogenes, which have been described as "fossils" that can be used to shed light on ancestral gene expression (Podlaha and Zhang 2009) . They have also been used to determine the age of gene duplications, and as a model for studying nucleotide substitutions, insertions and deletions (under the assumption that they evolve neutrally) (Petrov et al. 2000; Zhang and Gerstein 2003). However, they represent a major problem for accurate genome annotation.

A significant problem in itself is the accurate definition of pseudogenes. The identification of processed pseudogenes is straightforward due to the lack of introns and frequent addition of a polyadenylated tail to these sequences (Zhang et al. 2003b), but it is risky to dismiss all such sequences as having no biological function. Some processed pseudogenes play a regulatory role through the production of antisense RNA transcripts (Zhou et al. 1992; Weil et al. 1997; Tam et al. 2008; Watanabe et al. 2008) and there are several cases of processed pseudogenes that appear to have acquired a novel protein-coding function (Betran et al. 2002; Shao et al. 2007). Sakai et al have estimated that 1% of all processed pseudogenes in the human genome have been "reinvigorated by post-retrotransposition transcription", many having retained their coding regions intact, and suggest that they may have been an "indispensable resource" in "driving" mammalian evolution (Sakai et al. 2007).

Non-processed pseudogenes pose more of a problem, and are often defined by somewhat arbitrary cutoffs such as length compared to homologs in the absence of definitive *in vitro* evidence that no functional protein product is produced (Kuo and Ochman 2010; Lafontaine and Dujon 2010; Ochman and Davalos 2006). However, there have been several documented cases of genes that have suffered truncation events but still code for proteins retaining the domains required for functionality (Fikes et al. 1987; Struhl et al. 1993; Brayman and Hausinger 1996; Wang et al. 1997; Ruiz i Altaba 1999; Green et al. 2009) or may acquire alternative functions (Lonnerberg and Ibanez 1999; Merrill et al. 1999). Proteomic analysis of *Yersinia Pestis* KIM identified the translation of the gene coding for the ABC transporter protein y3734, despite the fact that this gene is extensively disrupted by insertion elements, frameshifts and nonsense mutations (Payne et al. 2010). Of course, it remains possible that such a translation product has no function in the cell.

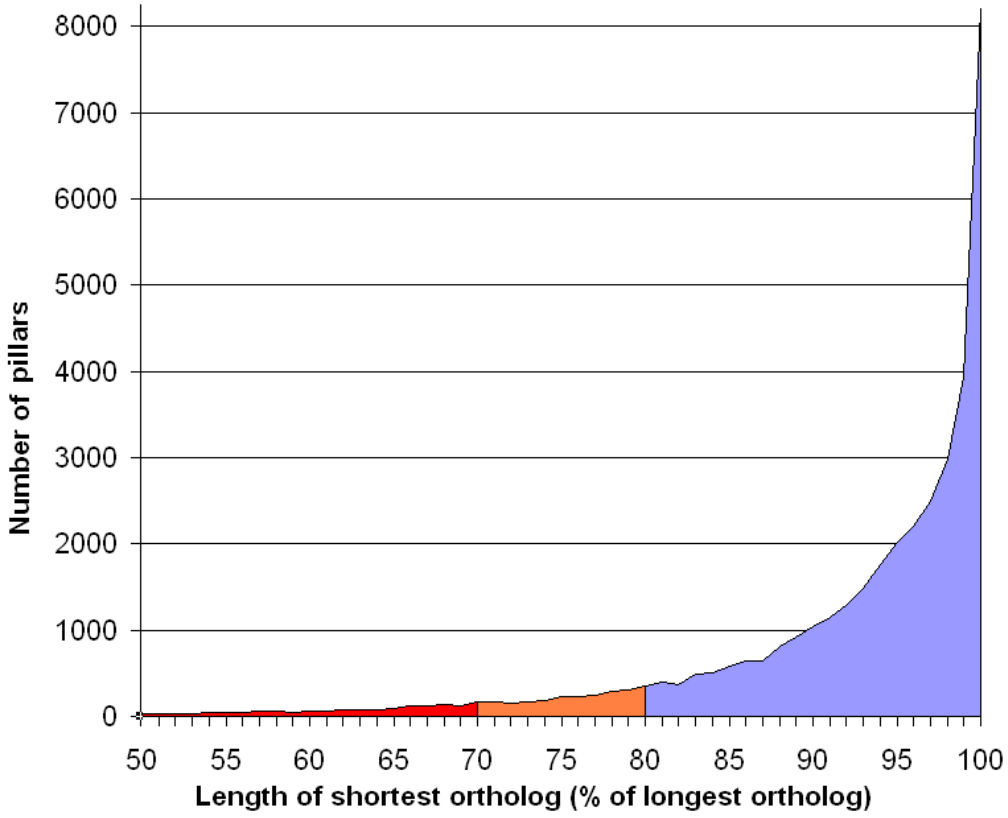

Figure 1.6 Distribution of YGOB homology pillars by homolog length range within each pillar, measured by the percentage ratio of the shortest pillar member to the longest pillar member. Species included are *S. cerevisiae, S. bayanus, C. glabrata, Z. rouxii, L. kluyveri, L. thermotolerans, L. waltii, E. gossypii* and *K. lactis*. The fraction of pillars below 80% and 70% are highlighted in orange and red respectively.

Furthermore, the lengths of homologous functional proteins corresponding to a given locus can vary considerably between related species. This may be due to slight species-specific differences in protein function or a lack of tight evolutionary constraint on protein length at some loci. An analysis of 9 of the 11 yeast species used in the initial SearchDOGS study indicated that the lengths of orthologous proteins in these species separated by ~100 million years vary considerably; the ratio of the smallest ortholog to the largest ortholog ranged from 1 (indicating identical lengths) to just 0.13 for the locus corresponding to *YNL195C* (Figure 1.6) (*V. polyspora* and the original *N castellii* annotation was excluded from this analysis due to sequencing and annotation inaccuracies) . The average ratio over 4,657 ortholog pillars was 0.83, indicating that the smallest annotated ortholog is on average 83% of

the length of the largest ortholog. 9.1% of *S. cerevisiae* genes are less than 80% of the length of their longest annotated orthologs in these species, and might be classed as pseudogenes using the 80% cutoff employed by Kuo and Ochman in their pseudogene surveys (Kuo and Ochman 2010; Ochman and Davalos 2006) (Figure 1.7). Lafontaine et al chose a 70% cutoff based on their observation that "among the functional members of a given protein coding gene family, the length variation does not exceed 30% in the majority of cases"; yet 867 YGOB ortholog pillars show a length variation greater than this between annotated orthologs at a locus (Figure 1.6) (Lafontaine and Dujon 2010).These observations further highlight that classifying ORFs as pseudogenes based on their being under 70% of the length of an annotated homolog may lead to misannotation of some *bona fide* genes.

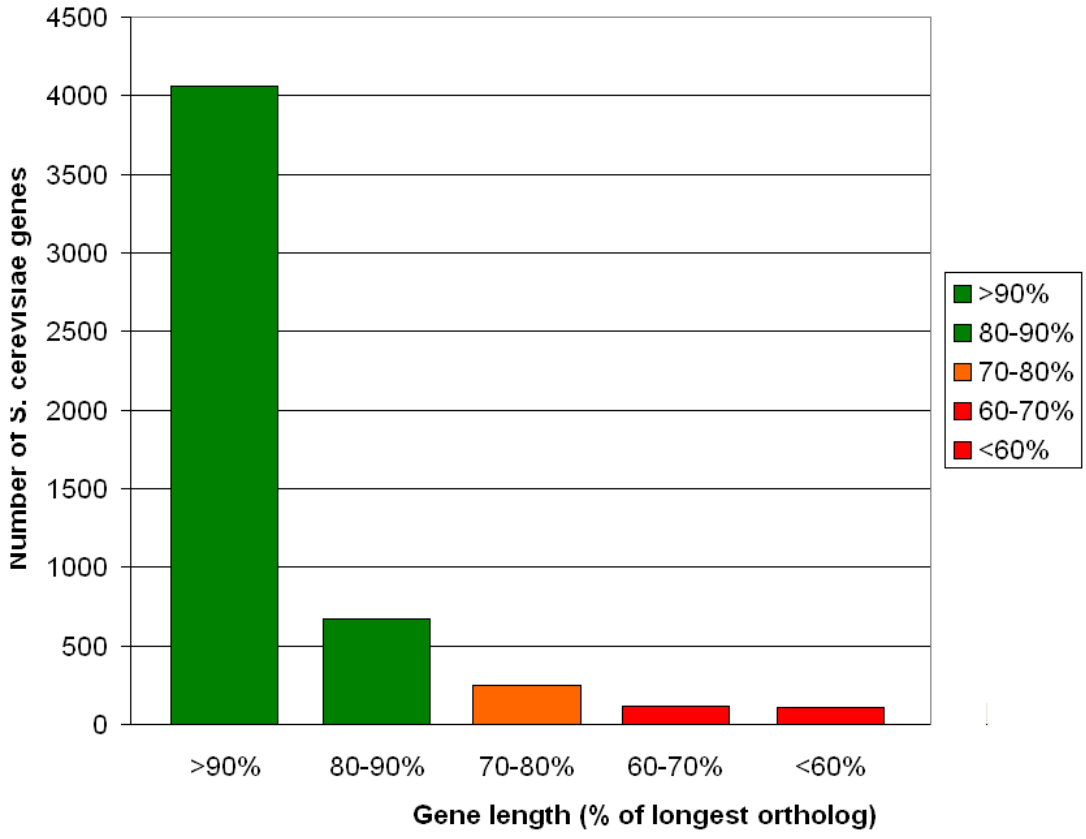

Figure 1.7 Distribution of *S. cerevisiae* genes by the percentage ratio of their lengths compared to the longest homolog in the corresponding YGOB homology pillar. Species included are the same as for Figure 1.6.

Another approach has been to identify pseudogenes based on their neutral rate of evolution using *Ka/Ks* comparisons (Torrents et al. 2003; Ochman and Davalos 2006), where *Ka* represents the rate of non-synonymous substitutions (nucleotide changes that alter the amino acid sequence) per non-synonymous site and *Ks* represents the rate of synonymous substitutions (nucleotide changes that leave the amino acid sequence unaltered) per synonymous site. This test makes the assumption that synonymous substitutions are always neutral, whereas nonsynonymous changes will be selected against in a conserved protein where there is selective pressure to maintain the amino acid sequence (Yang and Bielawski 2000). Thus, a *Ka/Ks* value of significantly less than 1 indicates region subject to protein sequence conservation, whereas a *Ka/Ks* value ~1 indicates that a region is evolving neutrally with respect to protein sequence, and a *Ka/Ks* value significantly greater than 1 indicates selection for adaptive evolution. This approach is also not without problems. The assumption that synonymous substitutions are always neutral whereas nonsynonymous substitutions are deleterious may not always hold (Torrents et al. 2003), particularly in instances where a protein, or parts of a protein, are under positive selection. The accuracy of the *Ka/Ks* calculation can also be affected by the quality of alignment between comparison sequences, the genomic context of the sequences, and the selection of comparison sequences (Bustamante et al. 2002; Torrents et al. 2003). Sequences that are insufficiently diverged from each other will not provide informative *Ka/Ks* results, nor will sequences that are saturated with substitutions. Furthermore, very young pseudogenes which have not had time to degenerate significantly will still bear the hallmarks of protein conservation associated with functional genes (Ochman and Davalos 2006).

The final problematic assumption is that pseudogenes truly are evolving neutrally in all cases. Kuo and Ochman reported several lines of evidence indicating that newly formed pseudogenes were eliminated from *Salmonella* genomes faster than neutral expectation, suggesting that they confer deleterious effects (Kuo and Ochman 2010). They suggested that this may reflect selection to remove the energetic cost of "useless" transcription and translation of young pseudogenes that are non-functional

but still retain the ability to be expressed. However, the few pseudogenes that have escaped large-scale deletion were found to correspond to genes with few interacting partners, indicating that the deletion of young pseudogenes may be due to selection against the expression of "toxic" damaged proteins that interfere with the function of protein networks.

A consequence of the difficulty in differentiating pseudogenes from *bona fide* genes is that in both prokaryotes and eukaryotes they are frequently misannotated and included in the databases as real genes (Lander et al. 2001; Mounsey et al. 2002; Payne et al. 2010; Waterston et al. 2002); Mounsey et al found that up to a fifth of genes annotated in the genome of the model organism *C. elegans* were potentially pseudogenic. The reverse is also true; a proteogenomic analysis of the *Arabidopsis thaliana* genome identified 498 true genes previously annotated as pseudogenes (Castellana et al. 2008).

As large-scale proteomic analysis becomes faster and more affordable, this will become a powerful tool in rooting out many of the pseudogene misannotations that are systemic within databases. However, while identification of gene expression and translation using proteomics resolves cases of real genes wrongly annotated as pseudogenes, the absence of evidence of translation is only supportive of the hypothesis that a particular locus is a pseudogene and not evidence of the absence of a function. One cannot rule out that the sequence is a real gene expressed under untested conditions. Bacterial SearchDOGS (described in Chapter 3) aims to provide the user with as much information as possible to differentiate real genes from pseudogenes, displaying the sequences and lengths of annotated homologs at a locus compared to the putative gene and providing *Ka/Ks* values for each putative gene against its homologs. Combining these approaches with a biological knowledge of the species tested in order to know what genes can and cannot reasonably be expected to become pseudogenised, and backing up bioinformatics data with proteomic data wherever possible represents the most comprehensive strategy for tackling the "pseudogene problem".

1.10 Hunting "Elves" and other wily features – annotation of short and highly diverged genes.

There is ample and growing evidence that small protein-coding genes play a multitude of crucial roles in eukaryotes and prokaryotes. In yeasts, *MFA* genes coding for the **a**-factor pheromone, which is secreted to direct the mating process between haploid cells of complementary mating types, range from 32-38 codons in length (Chen et al. 1997; Dignard et al. 2007). The *Saccharomyces cerevisiae* DASH complex, an essential microtubule-binding component of the kinetochore, contains three proteins of length 69-94 amino acids (Miranda et al. 2005). In the Gram positive bacterium *Bacillus subtilis*, the 46 amino acid Sda protein has a role in controlling sporulation by inhibiting the histidine kinases that initiate it (Rowland et al. 2004). In the Gram negative *E. coli* K12, *rpmJ* and *rpmH* code for components of the 50S ribosomal subunit of length 39 and 47 amino acids respectively (Blattner et al. 1997), and recent studies have shown the expression of small proteins in response to various stresses (heat shock, oxygen limitation, zinc limitation, envelope stress, acid stress, oxidative stress), many of which appear to have membrane functions (Hemm et al. 2008; Hobbs et al. 2010).

However the successful annotation of short protein-coding genes (15-50 amino acids) remains one of the biggest problems in genome annotation. Due to their small molecular masses, the proteins produced are difficult to identify using standard biochemical techniques such as 2D gel electrophoresis or mass spectrometry (Link et al. 1997b; Rudd et al. 1998; Fountoulakis et al. 1999; Han and Lee 2006; Hemm et al. 2010), and the length of the genes results in their frequent failure to be disrupted in genetic screens (Basrai et al. 1997; Kastenmayer et al. 2006). Identifying them through bioinformatics methods is even more difficult earning them the nickname ELFs, or "evil little fellows" (Ochman 2002).

Ab initio composition-based gene-calling methods use domain recognition and statistics such as codon bias to infer the existence of a gene; however many of these genes are too short to contain identifiable domains and motifs (Basrai et al. 1997; Blattner et al. 1997; Rudd et al. 1998; Cliften et al. 2001), or to reliably discriminate between coding and non-coding DNA based on codon usage (Skovgaard et al. 2001). This led to an overestimation of the number of small genes in the initial annotation of *E. coli* K12 and the annotation of a large number of spurious short ORFs that were subsequently removed.

Database homology methods rely on sequence similarity to annotated homologs in public databases. However small genes produce weak hits using sequence similarity programs such as BLAST (Altschul et al. 1990; Altschul et al. 1997) and are difficult to distinguish from hits to random stretches of DNA that happen to contain an ORF (Basrai et al. 1997; Blattner et al. 1997; Cliften et al. 2001). Furthermore in prokaryotic genomes the existence of spurious noncoding ORF annotations in the databases (Skovgaard et al. 2001; Salzberg 2007) leads to propagation of these short meaningless ORFs when used in sequence homology searches (Hemm et al. 2008).

As a result, many annotation procedures adopt a conservative strategy and exclude genes below a certain size (Salzberg 2007). For example, in the original annotation of *S. cerevisiae* strain S288c, genes shorter than 100 codons in length were excluded entirely (Goffeau et al. 1996; Fisk et al. 2006). In instances where a larger and a smaller gene overlap, composition-based automated annotation methods will usually annotate the larger and discard the smaller. However a proteogenomics study of the *Yersinia pestis* KIM genome identified that in 6 out of 24 instances of a larger ORF entirely overlapping a smaller ORF, the smaller ORF represented the true gene (Payne et al. 2010).

Therefore many genomes are nearly certainly both over- and under-annotated with respect to short genes, containing spurious short ORFs misannotated as real genes,

while missing many *bona fide* short genes with important cellular functions (Skovgaard et al. 2001; Hemm et al. 2008).

Hit Name	Status	Length (aa)	HSP Length	HSP Score	HSP E-value
YPL096C-A (ERI1)	ON	68	68	353	4e-34
Sbay_632.37	ON	68	68	298	8e-28
Kpol_1074.2	ON	68	66	199	2e-16
Cgla_YGOB_Anc_8.576	ON	69	70	154	4e-11
Zrou_YGOB_Anc_8.576	ON	67	66	147	3e-10
ABR192CA	ON	71	66	121	3e-07
Kthe_YGOB_Anc_8.576	ON	68	66	107	1e-05
Scas_YGOB_Anc_8.576	ON	77	66	80	0.017
CAGL0I10747g	ON	478	24	64	1.2
Kpol_526.7	ON	1640	45	63	1.4
Scas_550.7	ON	421	47	63	1.5
Scas_297.1	ON	798	34	62	2.0
Sbay_527.2	ON	1062	71	60	3.6
Sbay_656.5	ON	241	44	58	5.8
ZYRO0B04400g	ON	497	24	57	6.6
Kpol_1050.101	ON	487	24	56	8.6

Figure 1.8 BLASTP results for the 68 codon *S. cerevisiae* gene *ERI1* searched against the YGOB protein database. Pink shading indicates genes that are in the same pillar as *ERI1*. The locus is highly divergent, and the weak hit between *ERI1* and the newly discovered *N. castellii* homolog *Scas_YGOB_Anc_8.576* is highlighted.

A similar problem exists for highly diverged genes. These genes, when real, may be of major biological importance as they can represent rapidly evolving loci under adaptive selection in a species. However the extent of their divergence can lead to these genes "falling through the cracks" when using homology-based gene identification methods, especially if these genes are small. It becomes very difficult to tell spurious similarities from hits to real homologs. For example, *S. cerevisiae ERI1*, coding for an endoplasmic reticulum membrane protein only 68 amino acids long, hits homologs in *L. thermotolerans* and *N. castellii* with Expect (E) values of only 1e-5 and 0.17 respectively (Figure 1.8). Prior to the SearchDOGS study described in Chapter 2, the homologs in these species (as well as the *Z. rouxii* and *C. glabrata* homologs) had not been annotated. A TBLASTN hit between *S. cerevisiae ERI1* and the nucleotide region representing these unannotated homologs is sufficiently weak that it is difficult to distinguish from a random hit to a noncoding region.

The SearchDOGS software was designed with these difficult cases in mind. By identifying the syntenic context of both the query and the hit and showing unambiguously that they represent syntenic genomic regions, SearchDOGS provides strong additional evidence that a bona-fide homologous gene is being detected. In the case of *ERI1*, we found that the TBLASTN hits in *L. thermotolerans* and *N. castellii*

were located in a region between the homologs of *MSY1* and *PNG1*, the genes that flank *ERI1* in *S. cerevisiae*. SearchDOGS is described in detail in Chapter 2.

Chapter 2

Systematic discovery of unannotated genes in 11 yeast species using a database of orthologous genomic segments

Note – This chapter has been published as OhEigeartaigh et al. (2011).

Abstract

Background: In standard BLAST searches, no information other than the sequences of the query and the database entries is considered. However, in situations where two genes from different species have only borderline similarity in a BLAST search, the discovery that the genes are located within a region of conserved gene order (synteny) can provide additional evidence that they are orthologs. Thus, for interpreting borderline search results, it would be useful to know whether the syntenic context of a database hit is similar to that of the query. This principle has often been used in investigations of particular genes or genomic regions, but to our knowledge it has never been implemented systematically.

Results: We made use of the synteny information contained in the Yeast Gene Order Browser database for 11 yeast species to carry out a systematic search for protein-coding genes that were overlooked in the original annotations of one or more yeast genomes but which are syntenic with their orthologs. Such genes tend to have been overlooked because they are short, highly divergent, or contain introns. The key features of our software – called SearchDOGS – are that the database entries are classified into sets of genomic segments that are already known to be orthologous, and that very weak BLAST hits are retained for further analysis if their genomic location is similar to that of the query. Using SearchDOGS we identified 595 additional protein-coding genes among the 11 yeast species, including two new genes in *Saccharomyces cerevisiae.* We found additional genes for the mating pheromone a-factor in six species including *Kluyveromyces lactis.*

Conclusions: SearchDOGS has proven highly successful for identifying overlooked genes in the yeast genomes. We anticipate that our approach can be adapted for study of further groups of species, such as bacterial genomes. More generally, the concept of doing sequence similarity searches against databases to which external information has been added may prove useful in other settings.

Background

Yeast species have many features that make them an attractive model system for eukaryotic comparative genomics. These features include a high level of synteny conservation and small genome sizes (9–21 Mb) due to a low content of repetitive DNA and few introns (Wolfe 2006; Dujon 2010). We previously developed an online tool – the Yeast Gene Order Browser (YGOB) – for comparing local gene order relationships among species in genera such as *Saccharomyces*, *Kluyveromyces* and *Lachancea* (Byrne and Wolfe 2005). YGOB now contains genomic data from 11 species (Figure 2.1). Among these species, some form a clade of descendants from an ancestral whole-genome duplication (WGD) that changed the basal chromosome number from 8 to 16 (Wolfe and Shields 1997), whereas others diverged before the WGD occurred. We are unsure what depth of evolutionary time is represented by the species in YGOB, but when measured in terms of average protein sequence divergence this group of yeasts is approximately as diverse as the whole phylum Chordata (Dujon 2006).

YGOB contains 'pillars' of homology assignments across the 11 species. Each pillar can contain up to one gene from each non-WGD species and up to two genes from each post-WGD species (Byrne and Wolfe 2005). The genes in a pillar are therefore orthologs or (in the case of a post-WGD species retaining two genes) paralogs resulting from the WGD. The pillars have undergone several years of manual editing to make them as accurate as possible. YGOB also contains an 'Ancestral Genome',

which is the inferred gene content and gene order of the extinct ancestor that existed immediately prior to WGD (Gordon et al. 2009).

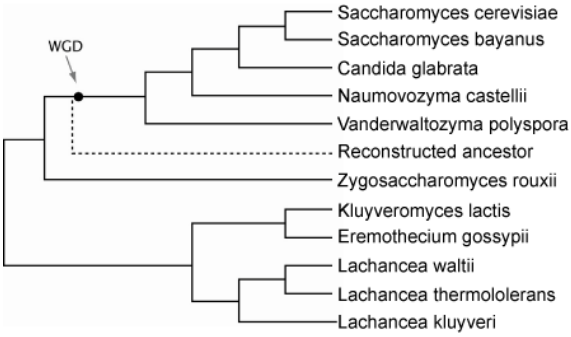

Figure 2.1 Phylogenetic relationship among the 11 yeast species used in this study. WGD indicates the position of the whole-genome duplication. The position of the inferred Ancestral genome (Gordon et al. 2009) is indicated. Tree topology is from (Hedtke et al. 2006).

The gene annotations in YGOB are derived from the original authors' annotations of the genome sequence of each species. In some cases we have 'switched off' genes in the original annotation that we believe to be spurious, but until now we have not added any genes to the original annotation sets (or to the current Saccharomyces Genome Database (Cherry et al. 1997) annotation for *S. cerevisiae*). However, while using YGOB we noticed many instances in which a particular gene appears to be missing in a particular species, in a genomic region that otherwise shows conserved synteny among all the species. Such loci appear as gaps in the YGOB display. For the post-WGD species it is quite common for one of the two paralogs formed by WGD to have been deleted, but it is more surprising to find genes that are completely absent (zero copies) in either a non-WGD or a post-WGD species.

Upon further examination we found that many of these apparently zero-copy loci are artefacts. When we examine the relevant DNA region, we find *bona fide* genes that were not annotated or were mistakenly labeled as pseudogenes, even in the case of highly curated genomes. This is particularly a problem with short genes of less than 100 codons, highly diverged genes, and genes containing introns. In some cases, all genes <100 codons were excluded entirely from the original curators' annotations due to the difficulty in telling these apart from spurious ORFs (Goffeau et al. 1996; Fisk et al. 2006). However, current estimates according to the Saccharomyces Genome Database (SGD) (Nash et al. 2007) are that the *S. cerevisiae* nuclear genome contains

131 verified ORFs of <100 codons and even among these, 28 contain introns. Detecting these 'missing genes' is important for many reasons, but our particular interest in this topic is that it would allow the correct identification of genuine lineage-specific gene gains and losses which may have evolutionary significance.

The primary reason why short genes are difficult to annotate is that they do not generate sufficiently strong hits (low *E*-values) in BLAST searches (Altschul et al. 1997). For instance the amino acid sequence of ribosomal protein L41 is nearly identical among all the species in YGOB, but because this protein is only 25 residues long the BLASTP E-value between any two Rpl41 sequences is only of the order of e-07 to e-06. Many annotation pipelines would regard such a hit as insignificant, because *E*-values of this magnitude are often obtained purely by chance when longer query sequences are used. More generally, any gene whose predicted protein product cannot generate a significantly strong BLAST score against its orthologs will tend to remain unannotated. Weak BLAST scores can be caused by very rapid sequence divergence (Kellis et al. 2003; Wolfe 2004), or a high content of repetitive sequence that is masked by sequence-filter software (Wootton and Federhen 1996), as well as by short sequence length.

In this chapter we describe SearchDOGS, a piece of software that works in conjunction with BLAST (Altschul et al. 1997) to identify unannotated genes. It is particularly designed to find genes that generate only weak BLAST hits, but whose syntenic context indicates that they are genuine orthologs to known genes. The major feature of SearchDOGS is that the genomes in the nucleotide sequence database used in the BLAST search have been pre-processed to subdivide them into sets of genomic regions that are already known to be orthologous. DOGS is an acronym for Database of Orthologous Genomic Segments. The BLAST results can then be post-processed to identify cases, even with very high *E*-values, where (i) the query protein hits genomic regions from multiple species in the database, and these regions are orthologous; or (ii) the syntenic context of the query protein is known, and it matches that of one or more of the database entries it hits.

SearchDOGS was initially developed as a standalone tool for displaying the syntenic contexts of the genomic hits obtained in a TBLASTN search using a single protein query. We then adapted it to carry out an automated and systematic search for unannotated genes in the genomes of all 11 yeast species in YGOB. Because the detection of a small or highly-diverged gene in one species may in turn make it possible to detect further orthologs of this gene in other species when the first gene is used as a query, we re-ran successive iterations of SearchDOGS on the yeast genomes until the program failed to find any more new genes.

Results

Orthologous genomic segments

The key concept behind SearchDOGS is that the nucleotide database that is searched by BLAST is organized into sets of sequences called Orthologous Genomic Segments (OGSs). We split up each of the 11 yeast genome sequences (Figure 2.1) used in YGOB into overlapping segments. Each segment consists of two adjacent annotated genes and the intergenic sequence between them (Figure 2.2). A BLAST nucleotide database ('DOGS') containing all the segments from all 11 species was then constructed. Separately, we mapped the two genes contained on each segment to the Ancestral yeast genome, which represents the gene order that existed just prior to the WGD event (Gordon et al. 2009). For each interval between two adjacent genes in the Ancestral genome, we were then able to identify genomic segments in the 11 modern species that are orthologous to this interval. A segment in a modern species can be orthologous to several consecutive intervals of the Ancestral genome due to gene deletions (Figure 2.2). Segments that are orthologous to the same Ancestral interval constitute an OGS group.

SearchDOGS was initially developed as a standalone program with a web interface (http://wolfe.gen.tcd.ie/searchDOGS), designed to search a single query protein against the DOGS database using TBLASTN (Figure 2.3A). Genomic segments hit in the search are identified in terms of their OGS groups. A typical protein query will hit the genomic segments that contain the annotated coding sequences of its orthologs in different species, which will constitute an OGS group. The BLAST HSPs (high-scoring pair alignments) of these hits will occur within the parts of the genomic segment that are already annotated as protein-coding, rather than the intergenic DNA between them. However, if in a particular species an ortholog of the query protein exists but has not been annotated, the DNA coding for it will have been annotated as intergenic DNA but the genomic segment containing this DNA will be part of the same OGS group as the orthologs in other species. For example, in Figure 2.2, if gene D exists in species 2 but has not been annotated, a TBLASTN search using gene D from another species as a query will hit an apparently noncoding region of segment 2 from species 2, as well as hitting coding regions of segments 2 and 3 from species 1. These three segments will all be in the same OGS group. So, by highlighting TBLASTN hits that occur in regions of database entries that are supposedly noncoding, we can identify possible unannotated orthologs of the query. We can consider even very weak hits between the query and noncoding regions of database entries, because we can reject any database hits that are not in the relevant OGS group.

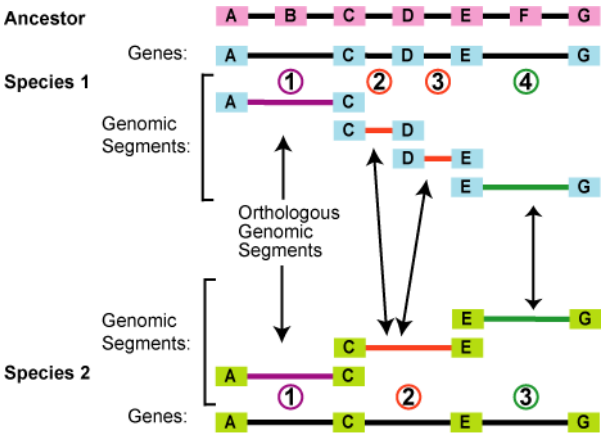

Figure 2.2 Definition of orthologous genomic segments. The genome sequences of species 1 and 2 are subdivided into overlapping regions, each containing two annotated genes and the intergenic DNA between them. The segments from species 1 and 2 in this example are classified into three orthologous genomic segment (OGS) groups, as indicated by coloring. Letters A-G represent genes in the Ancestral genome, some of which are retained in species 1 and 2.

As an example of results from the standalone SearchDOGS application, we found orthologs of the small (70 codons) *L. kluyveri* gene *SAKL0B06622g* in eight other yeast species in which it had not previously been annotated: *S. cerevisiae, S. bayanus, C. glabrata, Z. rouxii, K. lactis, E. gossypii, L. thermotolerans* and *L. waltii*, with *E* values ranging between 4e-08 and 0.049 (Figure 2.3A, 2.3B). In each of these noncoding regions an intact open reading frame was found, ranging in length from 61–88 codons, and showing significant amino acid sequence conservation (Figure 2.3C). When used as a BLASTP query the *S. cerevisiae* ORF, which we named *YBL026W-A,* hits *SAKL0B06622g* with an *E* value of 6e-04, and hits the other ORFs with *E* values ranging from 4e-21 to 0.009.

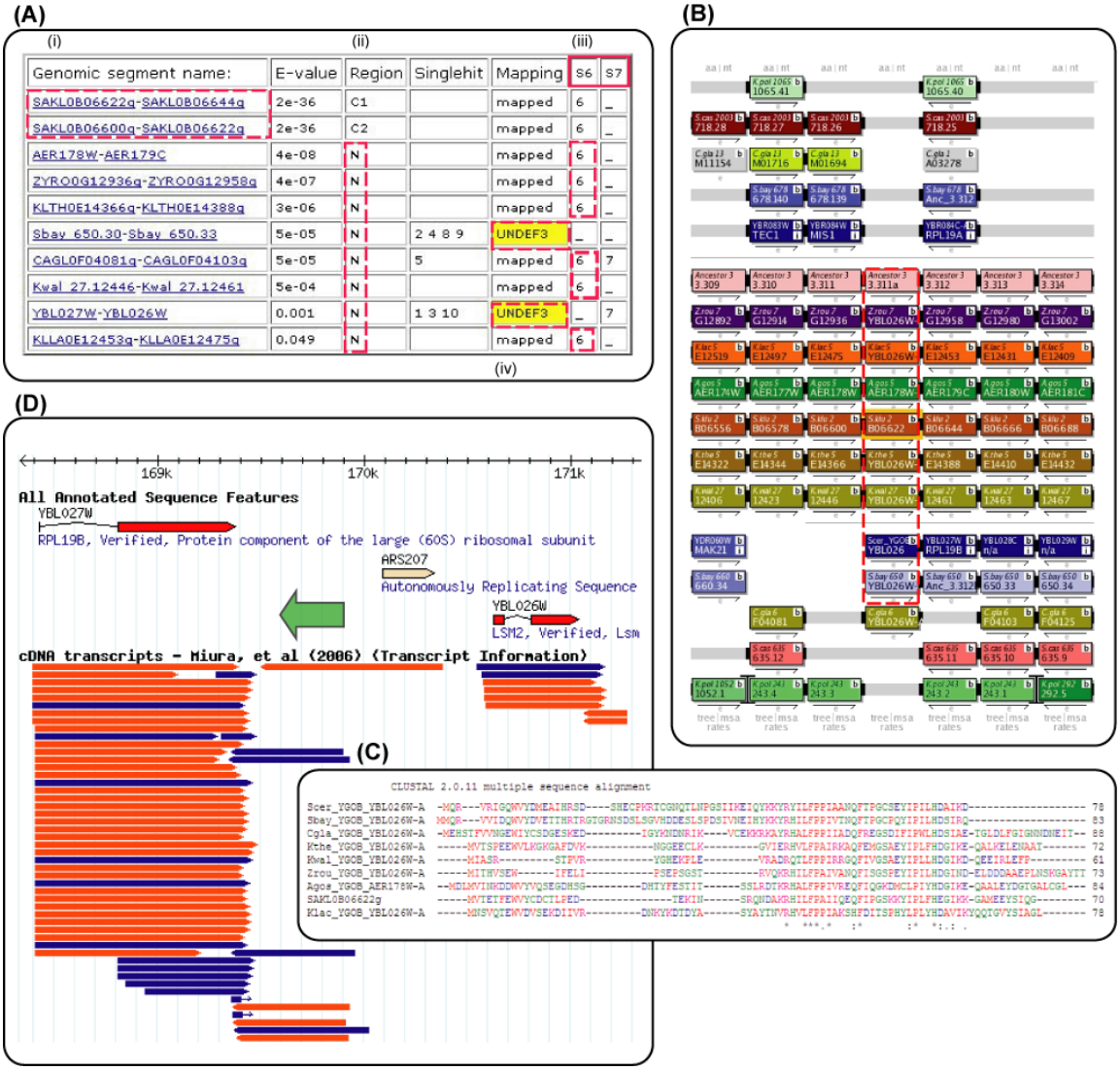

Figure 2.3 Orthologs of the *L. kluyveri* gene *SAKL0B06622g* discovered in eight species.
(A) Output for web SearchDOGS with *SAKL0B06622g* used as a query. (i) The dashed box highlights the genomic segment containing the query. (ii) The letter N indicates the hits to noncoding regions in other species; C1 and C2 indicate coding regions. (iii) S6 and S7 refer to segments of the Ancestral genome. Genomic segments in other species that map to the same Ancestral segments constitute an OGS group. Six noncoding hits map to the same ancestral region (S6) as the query (dashed red boxes) and the seventh, located between *YBL027W* and *YBL026W* in *S. cerevisiae*, can be mapped to an adjacent ancestral region. (iv) 'UNDEF3' for the *S. bayanus* and *S. cerevisiae* genomic segments indicates that they have undergone some rearrangement relative to the Ancestor. The Ancestral segments to which they map are listed as "singlehit" if they are not shared with any other species. (B) YGOB screenshot after addition of the new genes, which are indicated by the dashed box. An ortholog of this gene is also inferred to have existed in the Ancestral genome because it is present in both non-WGD and post-WGD species. (C) ClustalW multiple sequence alignment (Larkin et al. 2007) of the inferred protein sequences from nine species. (D) Location of the newly inferred *S. cerevisiae* gene *YBL026W-A* (green arrow), superimposed on a screenshot of the relevant region of chromosome II from SGD (Nash et al. 2007). Eight expressed sequence tags from Miura et al. (2006) indicate transcription of the gene.

These ORFs have been added to the YGOB database as new genes. Analysis of expressed sequence tag data (Miura et al. 2006) confirms expression of the newly identified *S. cerevisiae YBL026W-A* on the correct strand (Figure 2.3D). Prior to this study *SAKL0B06622g* appeared to be a species-specific gene in *L. kluyveri*, with no homologs annotated in the ten other YGOB species. These discoveries mean that orthologs of the gene are now known to exist, at a conserved location, in 9 of the 11 yeast species. We have not been able to find orthologs in the remaining two species (*V. polyspora, N. castellii).*

Automation and cycling

As we began to use the standalone SearchDOGS program, it became clear that due to the large number of hits and prospective genes being identified it would be necessary to automate the program to run over entire genomes. The automated version of SearchDOGS uses a modification of the original approach for increased speed and a slight increase in accuracy of synteny identification. The intergenic sequence between the two annotated genes in each genomic segment is used as a BLASTX query against a small database consisting only of the protein sequences of the genes that are syntenic with the query genomic segment. If a syntenic gene is hit, the region is retained for further processing. We retain all BLASTX hits with an *E*-value lower than 10, a very liberal cutoff.

We then test whether an intact gene structure with a protein sequence showing homology to the syntenic proteins can be identified within the intergenic region of the genomic segment. We use the program GetORF from the EMBOSS package (Rice et al. 2000) to generate a list of open reading frames located in the intergenic region. We use the protein translation of each ORF in the list as a BLASTP query against the syntenic YGOB pillar, and retain ORFs that hit the expected pillar. As well as this verification of synteny conservation we also require several other criteria to be met before an ORF is proposed as a genuine gene (Appendix I: Figure S2.1), such as a comparison of the length of the HSP generated using the protein translation of the

ORF as a BLASTP query against the syntenic pillar relative to the median length of the genes in that pillar. Finally, all proposed new genes are inspected by eye, considering their BLASTP results and a T-COFFEE multiple sequence alignment with other proteins in the pillar, for manual acceptance or rejection.

We ran a total of six cycles of the automated SearchDOGS program. In each cycle, genes discovered in the previous cycle were added to the query set. We also made modifications to the program between the cycles, to extend the range of situations it could deal with. The modifications included steps to automatically annotate intron-containing genes (see Methods), and modification of the synteny filter to allow unannotated genes to be detected in regions of genomes that have undergone rearrangement relative to other species. We also developed a method for dealing with pseudogenes. Pseudogenes are relatively rare in yeast genomes, but a few dozen have been described in *S. cerevisiae* and it is likely that similar numbers exist in other species (Lafontaine et al. 2004; Lafontaine and Dujon 2010). In addition, there are many degenerated fragments of mobile elements such as Ty retroelements in yeast genomes. These pseudogenes are detected by SearchDOGS but it is not possible to annotate a corresponding intact gene. To prevent these loci being rediscovered in each cycle, we flagged them as pseudogene-containing regions and excluded them from the results of subsequent SearchDOGS runs.

Automated SearchDOGS results

After six cycles of SearchDOGS we reached the point where no additional candidate genes were detected. The cumulative results of the six cycles are summarized in Table 2.1.

We added 595 new genes to the YGOB database, which can be viewed at http://wolfe.gen.tcd.ie/ygob (version 5: January 2011). A complete list of new genes is available at http://wolfe.gen.tcd.ie/searchDOGS. Of these, the largest proportion (43%) was in *S. bayanus*, which is still relatively poorly studied and annotated

(Cliften et al. 2003; Kellis et al. 2003). However, new genes were discovered in every species included. We were surprised to find two new genes in *S. cerevisiae* and 17 new genes in *E. gossypii*, given that these genomes have already been annotated to a very high standard (Brachat et al. 2003; Nash et al. 2007). The two new genes in *S. cerevisiae* are *YBL026W-A* (Figure 2.3) and *Scer_YGOB_Anc_7.495*. The latter gene, located between *YJR107W* and *YJR108W*, contains a frameshift in the 'reference' *S. cerevisiae* genome sequence of strain S288c, but not in alternative sequences of S288c obtained by Liti et al. (2009) and Miura et al. (2006), nor in sequences from other *S. cerevisiae* strains. In *E. gossypii* our results are supported by a recent resequencing and reannotation project that independently identified 15 of the 17 genes we discovered using SearchDOGS (Dietrich, F.S. et al, unpublished data. GenBank: AE016819, GenBank: AE016899-AE016904). During the course of the study we also identified a large number of genes across all species that were in need of modification or removal, due to errors such as a failure to annotate an intron, or partition of a single gene into multiple fragments due to frameshifts (Table 2.1). For some loci we found that a new gene could only be annotated in a particular species if apparent sequencing errors were overcome. We took a pragmatic approach to these loci: if a gene appeared to be intact except for one frameshift site or one internal stop codon, we annotated it and assumed that the problem was a sequencing error; if a gene contained more than one such site, we assumed that it is a pseudogene.

The list of new genes identified using SearchDOGS is heavily enriched for short genes: 64% of them are <200 codons long, and 38% are <100 codons. Most of them have yet to be assigned probable functions. By comparison, in the YGOB genome annotations of *S. cerevisiae* and *C. glabrata*, 16–17% of genes are <200 codons, and 3% are <100 codons. The large number of short genes discovered by SearchDOGS indicates not only that our approach is highly effective at detecting short genes, but also that a significant proportion of short genes has remained undiscovered to date.

For each new gene that we added, we calculated the ratio of nonsynonymous-to-synonymous nucleotide substitutions (*Ka/Ks*) between it and the other genes in the same

Table 2.1 New genes identified in 11 yeast species after six iterations of SearchDOGS.

Species	Updated number of genes	New genes added			Existing genes modified			Genes removed
		Total genes added	Intron-containing genes added	Frameshift / internal stop corrected	Total genes modified	Intron modified	Frameshift / internal stop corrected	
V. polyspora	5510	16	0	1	10	8	1	0
N. castellii	5688	18	1	1	13	9	1	1
C. glabrata	5224	16	0	1	6	3	2	1
S. bayanus	5223	258	142	3	17	8	7	3
S. cerevisiae	5606	2	0	1	2	0	2	1
Z. rouxii	5039	35	2	4	5	3	1	0
K. lactis	5120	40	4	9	4	2	1	1
E. gossypii	4742	17	2	0	3	0	1	0
L. kluyveri	5393	54	3	10	4	3	1	1
L. thermotolerans	5158	51	5	6	4	2	1	2
L. waltii	5275	88	56	1	38	19	13	4
Total		595	216	37	105	57	31	14

YGOB pillar using PAML (Yang 2007). In all cases we found that the ratio was less than 1, indicating natural selection to preserve the amino acid sequence of the encoded protein.

Examples of genes discovered by SearchDOGS

Highly divergent genes

The *S. cerevisiae* gene *NTC20*, coding for a protein required for pre-mRNA splicing, originally had orthologs annotated in all species except *E. gossypii*. SearchDOGS found a syntenic ortholog in *E. gossypii*, but the protein sequence divergence between it and the *S. cerevisiae* ortholog is so large that they do not hit one another in a BLASTP search ($E > 10$), which is probably the reason that the gene was overlooked in the original *E. gossypii* annotation (Dietrich et al. 2004) even though it is relatively long (171 codons). SearchDOGS initially found a hit to this genomic region in *E. gossypii* by using the *Z. rouxii* ortholog (*ZYRO0A13266g*) as a BLAST query. The *E. gossypii* ORF is confirmed as an *NTC20* ortholog because it hits six other proteins from the *NTC20* YGOB pillar when used as a BLASTP query. All six of these hits have very high *E*-values (ranging from 0.11 to 8.5), and the other four proteins in the pillar must have *E*-values greater than 10. *NTC20* is an exceptionally divergent gene: none of the 11 *NTC20* orthologs in the YGOB pillar is able to detect all the other members of the pillar with a BLASTP *E*-value below 10.

We also found new orthologs of *S. cerevisiae REC107*, an intron-containing gene that is involved in the early stages of meiotic recombination. At the start of this study orthologs were only known in the other post-WGD species and in *E. gossypii*. SearchDOGS identified divergent orthologs of *REC107* in all the other non-WGD species (*Z. rouxii, K. lactis, L. kluyveri, L. thermotolerans, L. waltii*), with BLASTP *E*-values to the *S. cerevisiae* protein ranging from 5e-19 to 6e-11.

Genes for a-factor

MFA genes coding for the **a**-factor mating pheromone in budding yeasts are known to be difficult to identify due to their short size (32–38 residues), high sequence

divergence and an apparently high rate of gene duplication or transposition (Dignard et al. 2007). Using a combination of SearchDOGS and standard TBLASTN searches we identified ten unannotated *MFA* genes: three in *N. castellii*, two in *K. lactis* and *L. kluyveri*, and one each in *Z. rouxii*, *L. waltii*, and *L. thermotolerans.* A previous study by Ongay-Larios et al. (Ongay-Larios et al. 2007) identified and knocked out one of the *K. lactis MFA* genes but did not notice the second gene.

An analysis of *MFA* gene locations reveals a complex history of gene duplication and relocation (Figure 2.4). For example *N. castellii* has five *MFA* genes, two of which are a pair formed by WGD and located at a site that is conserved with most other post-WGD species and *Z. rouxii* (Site 1 in Figure 2.4), but the other three *N. castellii* genes are at locations that are not shared with any other species. Among the 11 species, all but three of the ten separate genomic sites where *MFA* genes are currently located represent new sites to which *MFA* moved in the time since the WGD occurred (Figure 2.4).

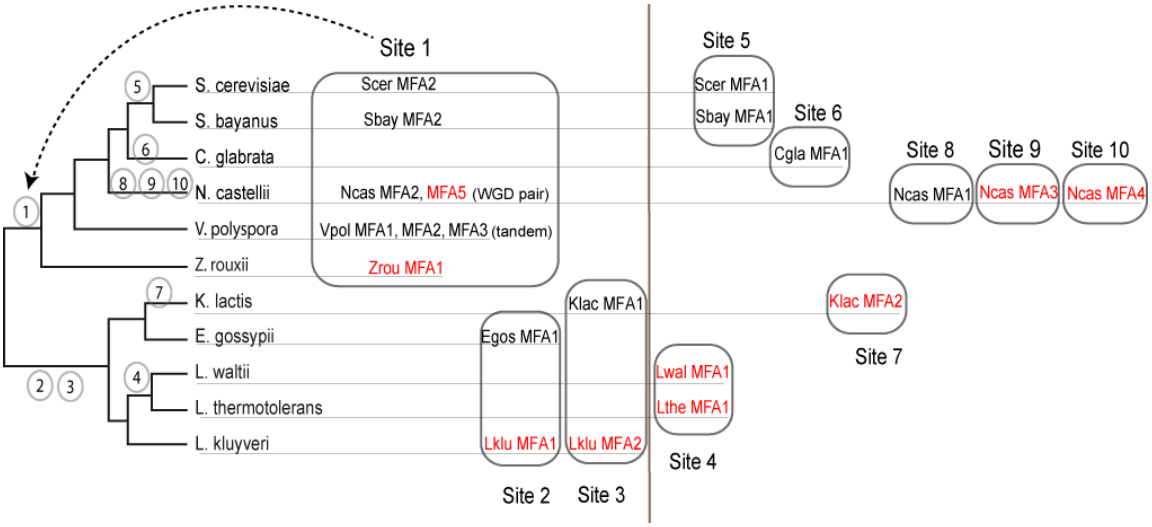

Figure 2.4 Summary of *MFA* (a-factor) gene locations in 11 yeast species. Sites 1-10 indicate the ten different genomic locations at which *MFA* genes are found. *MFA* genes newly discovered by SearchDOGS are highlighted in red. Numbers on the phylogenetic tree indicate the earliest branches to which each location maps. Sites to the right of the vertical line indicate recent species- or genus-specific gene movements.

Discovery of new divergent ohnolog pairs in S. cerevisiae

In three instances, the discovery of orthologs of *S. cerevisiae* genes in non-WGD species led us to realize that a pair of *S. cerevisiae* genes are ohnologs (paralogs produced by the WGD). The first pair is the *S. cerevisiae* genes *HOR7* and *DDR2* (59 and 61 codons, respectively). We initially discovered that *HOR7* has unannotated syntenic orthologs in the non-WGD species *K. lactis, L. kluyveri* and *L. waltii*, and then found that these non-WGD genes were also syntenic with, and had weak similarity to, *S. cerevisiae DDR2*. There is no direct BLASTP hit between the two *S. cerevisiae* proteins. The second pair is similar: *S. cerevisiae YDR524C-B* (66 codons) and *YCL048W-A* (79 codons) were found to be an ohnolog pair, because they are both syntenic with, and have weak similarity to, a newly discovered gene in each of *L. kluyveri*, *L. thermotolerans* and *L. waltii.* PSI-BLAST searches show that these four small *S. cerevisiae* proteins are members of a single family, but their precise function is ill-defined. *HOR7* and *DDR2* are known to be upregulated in response to stress. *HOR7* is responsive to hyperosmolarity (Lisman et al. 2004) and *DDR2* is a member of a family of multistress-responsive genes (Kobayashi et al. 1996). Both *HOR7* and *YDR524C-B* are expressed across a range of conditions, although whereas *HOR7* is upregulated in response to heat shock, *YDR524C-B* is downregulated. Both *DDR2* and *YCL048W-A* have low expression in rich media but are upregulated in response to ethanol or heat shock (Xu et al. 2009; Yassour et al. 2009) .

The third pair of newly-recognized ohnologs are *S. cerevisiae ABC1* and *YBR230W-A*, a small gene previously identified by McCutcheon and Eddy (2003). SearchDOGS identified that *YBR230W-A* (which was originally 'switched off' in YGOB's *S. cerevisiae* annotation) and *ABC1* hit the same genes in non-WGD species. *ABC1* and *YBR230W-A* have an unusual history because they no longer retain any homologous sequence. After WGD, the two ohnologs retained only different, non-overlapping parts of the original gene. Consequently, *ABC1* and *YBR230W-A* cannot be aligned to one another, but they both align to parts of a longer gene in non-WGD species that is orthologous to both of them (Figure 2.5). *S. cerevisiae ABC1* (501 codons) is a large

single-exon gene that corresponds to exon 2 of its orthologs in non-WGD species. *YBR230W-A* (66 codons) shows high similarity to exon 1 of the gene in non-WGD species (Figure 2.5), and is conserved within the genus *Saccharomyces* (McCutcheon and Eddy 2003). It appears that after WGD, one *S. cerevisiae* gene *(ABC1)* lost exon 1 and the other *(YBR230W-A)* lost exon 2 in a reciprocal fashion. Thus, these two genes that show no sequence similarity to each other share a common ancestor.

The origin of *ABC1* and *YBR230W-A* by fission of an ancestral gene raises a puzzle about the origin of *ABC1*'s mitochondrial import signal. *S. cerevisiae* Abc1 is a mitochondrial protein that is involved in activation of the cytochrome bc1 complex and is required for coenzyme Q biosynthesis(Bousquet et al. 1991; Won-Ki Huh 2003; Johnson et al. 2005). It is imported into the mitochondrion by means of a signal sequence at its amino terminus. Ybr230w-a and the proteins from non-WGD species are also predicted bioinformatically to be targeted to mitochondria (Claros and Vincens 1996; Emanuelsson et al. 2000; Guda et al. 2004; Small et al. 2004). Since *ABC1* did not retain the 5' end of the ancestral gene, it must have gained a new signal sequence upstream of the former exon 2. It is interesting to note that the *N. castellii* ortholog of *ABC1* also appears to have lost exon 1, but there is no evidence that exon 1 exists in the form of a separate gene in that species. The function of *YBR230W-A* is unknown, but transcriptome data indicates that both *ABC1* and *YBR230W-A* are expressed (Xu et al. 2009), and that expression of *YBR230W-A* is upregulated under heatshock conditions (Yassour et al. 2009).

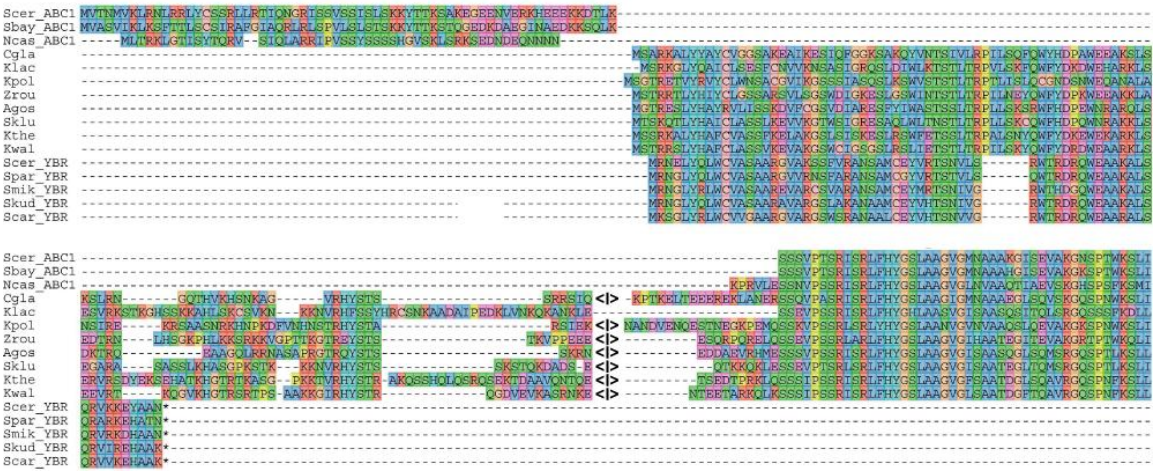

Figure 2.5 Alignment of Abc1, Ybr230w-a and orthologous proteins. Only the N-terminal region of Abc1 is shown, and the positions of introns are marked by <I>. The alignment was made using MUSCLE as implemented in Seaview (Gouy et al. 2010). The *YBR230W-A* genes in the *Saccharomyces* species (*S. cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii* and *S. carlsbergenesis*) align to the first exon of the two-exon *ABC1* gene in V. polyspora, Z. rouxii, E. gossypii, L. kluyveri, L. thermotolerans and L. waltii whereas *ABC1* of *S. cerevisiae, S. bayanus* and *N. castellii* align only to the second exon.

Application of SearchDOGS to CTG group yeasts and integration into the YGAP pipeline

After the publication of SearchDOGS (Oheigeartaigh et al. 2011), I subsequently adapted it to run on a set of 11 CTG group yeasts that make up the CGOB (*C*andida *G*ene *O*rder *B*rowser) database (Fitzpatrick et al. 2010), which is hosted by Professor Geraldine Butler's laboratory at University College Dublin. The initial candidate gene list generated by SearchDOGS is currently being analysed, and 201 new genes have to date been added to the CGOB database (Maguire S., personal communication). A SearchDOGS search step is also included in the recently developed YGAP (*Y*east *G*enome *A*nnotation *P*ipeline) annotation pipeline (Proux-Wéra et al., in preparation), following an initial genome annotation stage based on sequency homology and synteny information provided by the YGOB browser.

Discussion

The principle underlying SearchDOGS is one that is familiar and intuitive – that orthologous genes should be located in orthologous genomic regions. For the yeast species considered here, this principle turns out to be useful for gene discovery, because their genomes have undergone relatively little gene order change while accumulating extensive gene sequence divergence (Fischer et al. 2006). The idea that two orthologous genes can diverge so much in sequence that they fail to hit each other in a BLAST search is somewhat unsettling, and we were surprised when we encountered the first examples of this phenomenon (Wolfe 2004). We can now quantify the phenomenon more precisely as follows. In our YGOB database there are 5108 pillars that contain at least two genes. Among these, 135 pillars (2.6%) contain at least two genes that do not hit each other at all, despite being orthologous (BLASTP search, $E > 10$, Blosum62 matrix, merging hits seen both with and without the SEG low-complexity filter). The orthology of these genes has been confirmed via hits to a third sequence in the same pillar, or via longer chains of hits (Park et al. 1997).

Most annotation pipelines will only annotate a putative gene if it has significant similarity to another gene in a database, or if it has an ORF above a certain length threshold. Therefore it is not surprising that short, intron-containing, and highly-divergent genes tend to have been overlooked by the annotation process. SearchDOGS provides a method for finding these genes. In the near future it will probably also become possible to detect them using high-throughput transcription data such as RNA-seq (Yassour et al. 2009), but at the moment we have many genome sequences from species whose transcriptomes remain unstudied. Also, RNA-seq data establishes that a locus is transcribed, but does not identify its orthologs in other species.

Although the principle behind our method is simple, to our knowledge SearchDOGS is the first attempt to apply this principle in a systematic and automated way. As well

as the obvious advantage of speed, the automated approach has the additional advantage of robustness because it often finds multiple lines of evidence for the existence of the same gene. For example a *Z. rouxii* ortholog of *YPR036W-A* was detected in the intergenic region between *ZYRO0G17248g* and *ZYRO0G17270g* when it hit the *S. cerevisiae* protein in a search, but the same intergenic region also hit the orthologs of *YPR036W-A* from *V. polyspora, N. castellii,* and *E. gossypii*. In this way, searches using different members of the same pillar can back each other up, lending confidence to the predictions.

The only substantial problem we encountered using SearchDOGS was the difficulty of differentiating pseudogenes from unannotated but genuine genes. This is a particular problem for sequences that contain large ORFs but are nevertheless truncated relative to their orthologs in other species. Without experimental verification it will be difficult to distinguish between a functional gene that is shorter than its orthologs and a truncated pseudogene. Furthermore, the sequenced strain of a species may contain null alleles at some loci. These are loci at which the population is polymorphic with a mixture of functional and nonfunctional alleles. For instance, the *CRS5* gene contains an in-frame stop codon in *S. cerevisiae* strain S288c but not in other strains (Dujon et al. 1997). Without information from other strains it is impossible to distinguish between a null allele and a pseudogene that is fixed in the population.

One limitation of SearchDOGS is that, to find an unannotated gene, it must use a gene currently annotated in another species as a query. Therefore it cannot find completely novel genes that are not annotated in any species. We tried to overcome this limitation by using TBLASTX searches (six-frame translations of a query DNA sequence compared to six-frame translations of the database) after the six cycles of automated TBLASTN/BLASTX searches were finished. However, this approach generated a very large number of spurious hits (attributable to translations of sequences such as retrotransposon LTRs and RNA-coding genes), and we did not find any genuine additional genes in the 11 yeast species using it.

Conclusions:

We have successfully used SearchDOGS to identify a large number of genes previously overlooked in the genomes included in YGOB. The principle of using local gene order information to inform searches for unannotated genes is completely generic so in principle the SearchDOGS method could be applied to many groups of organisms, although in its current implementation – without sophisticated gene structure modeling – it is best suited to species with few or no introns. The broad requirements for a SearchDOGS approach to be viable are as follows:
(i) The species must already be reasonably well annotated. SearchDOGS will find missing genes, but if the majority of genes in a species are missing SearchDOGS will have difficulty pinpointing the locations of new genes relative to those already identified.
(ii) The species in the dataset must not be too rearranged. SearchDOGS can only make predictions in regions of the genome where it can establish local synteny relationships.
(iii) A pillar structure (*i.e.*, homology assignments for the genes) must exist or be generated. In our implementation we classified the genomic segments from each yeast species into orthologous groups (OGSs) by mapping them onto an Ancestral yeast gene order that we had previously determined (Gordon et al. 2009). For SearchDOGS to be applied to other systems, the user would need to nominate one genome as a reference onto which the OGS groups would be mapped.

Based on these requirements, we anticipate that SearchDOGS may prove useful in the future for finding unannotated genes in bacterial genomes, but it may be less useful in genomes with many introns and large noncoding regions, such as mammals, or in species that lack close relatives with well-annotated genomes.

Methods

Database and search method

The DOGS database was constructed using the genome sequence and gene annotations in the YGOB browser, data release 4 (May 2010) (Byrne and Wolfe 2005), which includes 11 species (Table 2.2). The Ancestral yeast gene order is from Gordon et al. (Gordon et al. 2009). For the standalone version of SearchDOGS we constructed a single nucleotide database containing all the genomic segments. This database can be searched using either TBLASTN or TBLASTX. For the early automated cycles of the program, the amino acid sequence of each protein in the YGOB database was used as a TBLASTN query against syntenic genomic segments. To reduce computation time we constructed a specific small database for use with each pillar's queries, containing only the genomic segments that are syntenic to it. The TBLASTN searches used cutoffs of $E < 10$ and 100 results listed, with the low-complexity SEG filter turned off. Hits to noncoding regions syntenic with the query protein were retained.

As subsequent iterations of SearchDOGS were run, modifications were made to improve the initial synteny-determining method, and to improve speed by using BLASTX instead of TBLASTN. The final method for establishing synteny is as follows: For each genomic segment the pillars containing the flanking genes are retrieved. This information is used to map the intergenic region against the other species in YGOB. If no rearrangement has occurred between a given species and the species of the query, all the genes in that species between the flanking pillars are retrieved, making up a database against which the intergenic region of the genomic segment is searched (BLASTX) (Appendix I: Figure S2.2). Otherwise we ‘step out’ from one flanking pillar towards the other, retaining each gene until we reach a gene for which pillar information shows that synteny with the intergenic sequence has been lost, or up to a maximum of 10 genes from the pillar (Appendix I: Figure S2.2).

Table 2.2 Genome sequences and annotations used in the SearchDOGS database.

Species	Coverage	Sequence	Gene annotation
V. polyspora	7.8x	(Scannell et al. 2007)	(Scannell et al. 2007)
N. castellii	4x	(Cliften et al. 2003)	Wolfe laboratory, based on Cliften et al. (2003)
C. glabrata	Complete	(Dujon et al. 2004)	(Dujon et al. 2004)
S. bayanus	6.4x	(Kellis et al. 2003)	(Kellis et al. 2003)
S. cerevisiae	Complete	(Goffeau et al. 1996)	Wolfe laboratory, based on SGD 2009 release
Z. rouxii	Complete	(Souciet et al. 2009)	(Souciet et al. 2009)
K. lactis	Complete	(Dujon et al. 2004)	(Dujon et al. 2004)
E. gossypii	Complete	(Dietrich et al. 2004)	(Dietrich et al. 2004)
L. kluyveri	Complete	(Souciet et al. 2009)	(Souciet et al. 2009)
L. thermotolerans	Complete	(Souciet et al. 2009)	(Souciet et al. 2009)
L. waltii	8x	(Kellis et al. 2004)	(Kellis et al. 2004)

Intron and frameshift prediction

Open reading frames within the regions of interest identified by SearchDOGS are obtained using GetORF with default parameters (Rice et al. 2000) except that the minimum ORF size is 60 nucleotides (start to stop). The set of ORFs generated by GetORF are subjected to a first step of analysis using BLASTP, as described in the Results. This step identifies coding regions that are free of frameshifts and consist of a single exon (Appendix I: Figure S2.1). Next, results are subjected to a second step of analysis designed to identify genes containing frameshifts or introns. In this step we look for pillars of genes that map to the intergenic region of the genomic fragment in which a BLAST hit has been found. If a potential ORF within the intergenic region contains a single frameshift that can be corrected to translate to a protein similar to other genes in the homologous pillar, it is considered real and is corrected. The location of the frameshift is an estimate, and therefore the ORF is flagged for manual verification.

We anticipate that a newly discovered gene might contain an intron if one or more of the genes in the YGOB pillar that hits it contains an annotated intron. In the case of pillars of genes containing introns, TBLASTN is used to search the protein sequence of each of the exons of the genes in these pillars against the intergenic region of the fragment to define potential exons within the fragment. If two or more potential ORFs have the same order as the exons of any genes in the syntenic pillar, and if the lengths of the ORFs are within 10 amino acids of the lengths of the exons in the pillar, an intron is predicted and we search for splice sites (GT-AG) associated with the boundaries of the intron. No frameshifts are allowed when an intron is predicted. If the TBLASTN hits do not include start and stop codons, an enlargement of the coding region of up to 40 amino acids is allowed until start and stop codons are reached. The final protein length is tested against the median protein length of the homologous pillar for a measure of prediction confidence. Exons smaller than 20 codons are difficult to identify by BLAST, so if a pillar that generates a hit contains a small exon, only the larger exon(s) are usually detected in a TBLASTN search, and therefore the hit is flagged for manual annotation.

New genes identified using SearchDOGS were added to the YGOB database and given temporary names containing the tag 'YGOB' such as *Zrou_YGOB_Anc_5.606* to indicate a *Z. rouxii* ortholog of the gene at ancestral position Anc_5.606 (Gordon et al. 2009). We will communicate lists of these loci to the relevant databases so that permanent names can be assigned to them.

Criteria for rejection of hits

A BLASTX hit between a genomic segment and a protein from an orthologous YGOB pillar could be rejected, either automatically (Appendix I: Figure S2.1) or after manual inspection. The most common reasons why hits between a genomic segment and an orthologous protein were rejected were:

- Segments did not contain an intact ORF, due to multiple stop codons and/or frameshifts. These were classed as pseudogenes.
- BLAST relationship was not reciprocal: the intergenic sequence of a genomic segment had a BLASTX hit to a protein in an orthologous YGOB pillar, but when the translated ORF from the genomic segment is used as a BLASTP query it failed to hit any of the proteins in the same pillar.
- The length of the HSP generated by the BLASTP search was not sufficiently long compared to the median length of the genes in the corresponding YGOB pillar.
- The translated ORF showed too little sequence similarity to existing genes in the pillar in a subjective inspection of a T-coffee alignment (Notredame et al. 2000), and were therefore considered unlikely to be real.
- Segments syntenic to intron-containing pillars, for which we were unable to construct a convincing gene model.

Chapter 3

Bacterial SearchDOGS: Automated identification of potentially missed genes in annotated bacterial genomes

3.1 Abstract

We report the development of Bacterial SearchDOGS, software to automatically detect missing genes in annotated bacterial genomes by combining BLAST searches with comparative genomics. Having successfully applied the approach to yeast genomes, we redeveloped SearchDOGS to function as a standalone, downloadable package, requiring only a set of GenBank annotation files as input. The software automatically generates a homology structure using reciprocal BLAST and a synteny-based method; this is followed by a scan of the entire genome of each species for unannotated genes. Results are provided in a HTML interface, providing coordinates, BLAST results, syntenic location, *Ka/Ks* protein conservation estimates and other information for each candidate gene. Using Bacterial SearchDOGS we identified 171 gene candidates in the *Shigella boydii* sb227 genome, including 62 candidates of length <60 codons. Bacterial SearchDOGS has two major advantages over currently available annotation software. Firstly, it outperforms current methods in terms of sensitivity, and is highly effective at identifying small or highly diverged genes. Secondly, as a freely downloadable package it can be used with unpublished or confidential data.

3.2. Introduction

With the rapidly decreasing cost and increasing speed of genome sequencing, a wealth of genome sequence information is becoming available to the scientific community. As of September 2011, Entrez Genome (http://www.ncbi.nlm.nih.gov)

reports 1750 complete bacterial genomes and another 5230 in progress. However, the speed at which these new genomes can be accurately annotated is quickly becoming a bottleneck. In the vast majority of cases, protein-coding genes are annotated using automated programs (Stothard and Wishart 2006), which for the most part can be divided into two classes: "composition-based" gene prediction programs that use characteristics of sequence composition to predict where protein-coding open reading frames exist (Frishman et al. 1998; Delcher et al. 1999; Bocs et al. 2003; Larsen and Krogh 2003; Besemer and Borodovsky 2005), and "sequence similarity" programs that predict protein-coding ORFs based on the identification of sequence similarity to annotated homologs in other species (Samayoa et al. 2011).

However both these approaches run into difficulties when annotating short or highly diverged genes. Statistical methods such as codon bias have less power in discriminating coding from non-coding DNA for short genes (Skovgaard et al. 2001) using the "composition-based" approach, and small or highly diverged genes genes return weak hits to homologs in BLAST searches and therefore cannot be accurately differentiated from ORFs that occur by chance (Skovgaard et al. 2001). This uncertainty has resulted in overestimation of the number of protein-coding genes in annotated bacterial species (Skovgaard et al. 2001; Nielsen and Krogh 2005), but also in many *bona fide* short genes being overlooked (Hemm et al. 2008; Kucerova et al. 2010; Hemm et al. 2010 ; Samayoa et al. 2011)

One way to overcome this problem is to ascertain the syntenic context of the potential gene. If a potential unannotated gene is found to lie in the same local genomic region as its orthologs in other species, then there is a good likelihood that it is a *bona fide* gene even if it produces relatively weak BLAST results to its orthologs. Identification of regions of conserved synteny can also be used to detect gene duplications, fusions, and paralogy relations when comparing multiple genomes (Vallenet et al. 2006), as well as make functional association predictions (Enault et al. 2005; Friedberg 2006).

In the previous study, I described the development of the software SearchDOGS (standing for searches against a Database of Orthologous Genomic Segments), which uses conserved local synteny across species combined with BLASTX sequence similarity searches to identify genes that may have been missed from published annotations due to small size or a high level of divergence (Oheigeartaigh et al. 2011). Using this approach, we identified 594 previously undetected genes in 11 published yeast genomes, including a number of new genes in well-studied model organisms such as *Saccharomyces cerevisiae* and *Eremothecium gossypii*. Many of the genes identified are very highly diverged, and 36% are less than 100 amino acids in length.

Having applied the method successfully in yeast, we wished to extend the scope of SearchDOGS to allow it to be applicable to any set of suitable species where extensive local synteny can be established. Bacteria are an ideal candidate for the SearchDOGS approach due to their simple genomic architecture, low noncoding DNA content, and the large number of species to choose from for comparison. In this chapter I report the application of SearchDOGS to sets of strains and species from the γ-proteobacteria. This clade includes gram negative bacteria, many of which are common human pathogens, and contains the model organism *E. coli* K12 (Lukjancenko et al. 2010).

A host of powerful and comprehensive annotation pipelines have been developed for bacterial genomes in recent years, such as Microscope (Vallenet et al. 2009), AGeS (Kumar et al. 2011) and AGMIAL (Bryson et al. 2006). Bacterial SearchDOGS complements these platforms by providing an exceptionally sensitive tool for identifying the genes that prove most tricky for automated annotation programs. In contrast to the yeast implementation of SearchDOGS, which imported information about homologous sequence pillars from the YGOB database, Bacterial SearchDOGS was designed as a standalone piece of software. Thus the only input data required are GenBank files of the relevant bacterial genomes. It allows the user to choose a number of annotated genomes to compare, and automatically generates pillars

containing genes that are orthologous between species. For a given species, it generates a list of candidate loci where a missing gene may exist and where an open reading frame showing sequence homology to orthologous genes has been identified. A detailed HTML output page is generated, providing syntenic location, coordinates, BLASTP results and omega (*Ka/Ks*) values as an estimate of protein conservation (Yang and Bielawski 2000). The aim is to provide the user with sufficient information to accept or reject each candidate with confidence. The software is designed to be freely downloaded from http://wolfe.gen.tcd.ie and used locally, and thus is suitable for use with confidential or unpublished genomic data. It will be released shortly.

3.3. Results

3.3.1 Generation of results

After data input, the first stage of the Bacterial SearchDOGS program is to automatically develop a homology pillar structure analogous to the pillars in YGOB. Multiple rounds of reciprocal BLAST and syntenoblast (see Methods) searches are used to generate a set of pillars, each of which contains any annotated orthologs of a given locus in each genome in the dataset. To find unannotated genes, each genome is then sliced into overlapping genomic segments containing two genes and the intergenic sequence between them. Each of these genomic segments is tested against the array of ortholog pillars to identify pillars that share synteny with it. The intergenic sequence between the two annotated genes in the genomic segment is then used as a BLASTX query against a small database consisting of the protein sequences of the genes in the syntenic ortholog pillars. In order to identify weak hits in an orthologous position, all hits are considered regardless of E value.

(A)

adjacent left	Dist (nt)	pillar hit (4739)	Dist (nt)	adjacent right
ECK1_NP_418210.1	23	ECK1_YP_026245.1	5801	ECK1_YP_026246.1
ECO1_NP_312723.1	23	ECO1_NP_312724.2	5801	ECO1_NP_312725.1
ECS8_YP_002393742.1	23	ECS8_YP_002393743.1	5801	ECS8_YP_002393744.1
SBOY_YP_410065.1	23	SBOY_YP_410065.1-SBOY_YP_410066.1	5779	SBOY_YP_410066.1
SETY_NP_807269.1-SETY_NP_807270.1	9	SETY_NP_807269.1	5738	SETY_NP_807110.1
				VCHO_YP_001215061.1
YPAN_YP_649927.1	80	YPAN_YP_649928.1	6685	YPAN_YP_650032.1
PSYR_NP_790256.1				

ORF in intergenic region between SBOY_YP_410065.1 (*yieO*) and SBOY_YP_410066.1 (*yifDA*)

(B)

YPAN	ECK1	SBOY	PSYR	SETY	ECO1	ECS8	VCHO	XCAM
K 230	K 230	P 230		K 234	K 230	K 230		

(C)

```
>SBOY_YP_410065.1-SBOY_YP_410066.1
*kekgrlvMPLSAQQLAAQKNLSYVLAEKLAQRILKGEYEPGTILPGEIELGEQFGVSRT
AVREAVKTLTAKGMVLPRPRIGTRVMPQSNWNFLDQELLTWWMTEENFHQVIDHFLVMRI
CLEPQACLLAATVGTAEQKAHLNTLMAEMAALKKNFRRERWIEVDMAWHEHIYEMSANPF
LTSFASLFHSVYHTYFTSITSDTVIKLDLHQAIVDAIIQSDGDAAFKTCQALLRSPDK*
>YPAN_YP_649928.1
MQLNTQQQEAAQRNLSYLLAEKIGQRILAGEYEAGSILPGEIELGEQFGVSRTAVREAVK
MLAAKGMLLPRPRIGTRVMPQTNWNYLDQELLTWWMTKENFDQVMQHFLILRTSLEPQAC
YLAATHANEKQRELLASLVAEMRALHSNFNREQWIQVDTQFHKLIYEASGNPFLISFANL
FSSVYYSYFRAITGDEVIKLQHHQNIVDTILAGDNQGALFACQVLLKTVD*
>SETY_NP_807269.1
MPLSAQQLAAQKNLSYVLAEKLAQRILKGDYAPGTILPGEIELGEQYGVSRTAVREAVKT
LTAKGMVLPRPRIGTRVMPQGNWNFLDQELLTWWMTEENFHQVVDHFLVMRISLEPQACL
LAATLGTPEQKARLNALMEEMVALKKHFKRERWIEVDMAWHEHIYEMSANPFLISFATLF
HSVYHTYFTSITYNEVVKLDLHQAIVDAIADGDGERAFQACQALLIAPNERPDN*
>ECK1_YP_026245.1
MPLSAQQLAAQKNLSYVLAEKLAQRILKGEYEPGTILPGEIELGEQFGVSRTAVREAVKT
LTAKGMVLPRPRIGTRVMPQSNWNFLDQELLTWWMTEENFHQVIDHFLVMRICLEPQACL
LAATVGTAEQKAHLNTLMAEMAALKENFRRERWIEVDMAWHEHIYEMSANPFLTSFASLF
HSVYHTYFTSITSDTVIKLDLHQAIVDAIIQSDGDAAFKACQALLRSPDK*
>ECO1_NP_312724.2
MPLSAQQLAAQKNLSYVLAEKLAQRILKGEYEPGTILPGEIELGEQFGVSRTAVREAVKT
LTAKGMVLPRPRIGTRVMPQSNWNFLDQELLTWWMTEENFHQVIDHFLVMRICLEPQACL
LAATVGTAEQKAHLNTLMAEMAALKENFRRERWIEVDMAWHEHIYEMSANPFLTSFASLF
HSVYHTYFTSITSDTVIKLDLHQAIVDAIIQSDGDAAFKACQALLRSPDK*
>ECS8_YP_002393743.1
MPLSAQQLAAQKNLSYVLAEKLAQRILKGEYEPGTILPGEIELGEQFGVSRTAVREAVKT
LTAKGMVLPRPRIGTRVMPQSNWNFLDQELLTWWMTEENFHQVIDHFLVMRICLEPQACL
LAAKVGTAEQKAHLNTLMAEMAALKENFRRERWIEVDMAWHEHIYEMSANPFLTSFASLF
HSVYHTYFTSITSDTVIKLDLHQAIVDAIIQSDGDAAFKACQALLRSPDK*
```

Figure 3.1 Identification of a *yieP* ortholog in *S. boydii*. (A) SearchDOGS output showing the syntenic neighbourhood of the *S. boydii* hit. The intergenic regions of the *S. boydii* genomic segments named in red hit proteins in a syntenic ortholog pillar (pillar 4739) in a BLASTX search and contain a candidate gene. Each row names the genes flanking this pillar in each species, if their syntenic context is conserved. (B) Amino acid lengths of proteins coded by predicted (P) and known (K) genes at the yieP locus. (C) Amino acid sequences coded for by the stop-to-stop ORF identified by SearchDOGS in *S. boydii* (highlighted in yellow) and the annotated genes in the ortholog pillar corresponding to the *yieP* locus. Start codons are highlighted in green, stop codons in red.

Where a hit occurs, SearchDOGS then tries to identify an intact gene structure corresponding to it. A subroutine called ORFmaker identifies a stop to stop ORF (i.e., from one stop codon to the next stop codon in the same frame) corresponding to the HSP. In order to cope with potential sequence errors, ORFmaker's default setting allows the readthrough of a single stop codon or frameshift where HSP evidence exists to indicate the existence of a longer ORF, although such ORFs are flagged as dubious. The set of ORFs predicted in a genome is then filtered based on criteria such as HSP length and position of start and stop codons relative to the HSP (see Methods). HTML output pages are then produced showing details of the ORFs that pass these criteria (Figure 3.1), and providing links to the BLAST results and nucleotide sequences. Conservation of protein sequence across wide evolutionary distances, corresponding to nonsynonymous-to-synonymous rate ratios (*Ka/Ks*) significantly less than 1, is a strong indicator of the authenticity of

a gene (Yang and Bielawski 2000). Bacterial SearchDOGS integrates PAML software (Yang 2007) to perform *Ka/Ks* tests in a pairwise fashion between annotated and potential genes (Figure 3.2A). *Ka/Ks* values significantly smaller than 1 are seen as a strong indicator of protein sequence conservation. Individual *Ka/Ks* values are calculated as well as an average value of *Ka/Ks* between the potential gene and its corresponding ortholog pillar. A 95% confidence interval test of the difference between *Ks* and *Ka* (*Ks*-*Ka*) is calculated for each pairwise comparison as a measure of statistical significance.

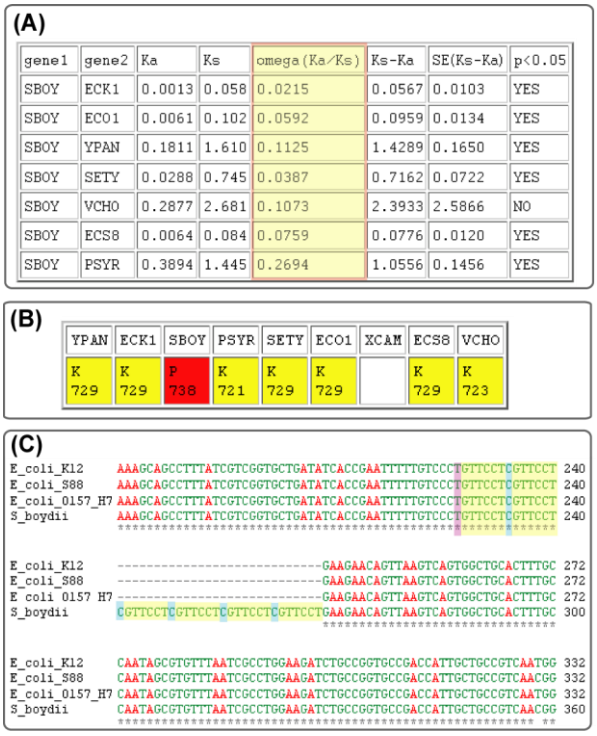

Figure 3.2 Insertion in the genomic sequence of *S. boydii FadB*. (A) By reading through a frameshift, SearchDOGS's ORFmaker program was able to create a full-length ORF (highlighted in red) at the locus corresponding to highly conserved gene *FadB*. (B) Screenshot of the SearchDOGS *Ka/Ks* output for the reconstructed *FadB* candidate against the annotated FadB orthologs. Omega (*Ka/Ks*) values are highlighted in red. 95% confidence interval tests carried out on the value of *Ks-Ka* indicate that in all cases except the *S. boydii/V. cholerae* pairwise comparison *Ks-Ka* is greater than 0 with statistical significance, indicating protein sequence conservation. (C) ClustalW alignment of the nucleotide sequence surrounding an apparent expansion of a 7 bp repeat in *S. boydii FadB*. *E. coli* K12, *E. coli* S88, *E. coli* O157:H7 and *S. boydii* are shown, and the repeat sequence is highlighted. The entire length of the insertion is 28bp, and thus causes a frameshift at this location in *S. boydii FadB*.

genome within this dataset.

3.3.2 γ-proteobacterial species used in the study

To test the software, we chose 9 species from the γ-proteobacterial clade (Table 3.1). These include the model organism *E. coli* K12 strain MG1655, some closely related strains (*E. coli*, *Shigella*), and some species of increasing evolutionary distance (*Vibrio*, *Pseudomonas*, *Xanthomonas*). The genomes range in size from 4.1 Mb and 3875 protein-coding genes (*Vibrio cholerae*) to 6.5 Mb and 5481 protein-coding genes (*Pseudomonas syringiae*). All consist of a single circular chromosome, other than *Vibrio cholerae* which has two chromosomes. Bacterial SearchDOGS was successfully able to generate an extensive ortholog pillar structure for these species (Table 3.2). Using these pillars, we predicted candidate genes that remained unannotated in each

Table 3.1 Genomes used in the SearchDOGS species comparison, including model organism *E. coli* K12 MG1655. All species are from the gammaproteobacterial clade.

Species:	GenBank accession no	Genome size (MB)	Protein-coding genes	Acronym
Escherichia coli K12 substr. MG1655	U00096.2	4.6	4148	ECK1
Escherichia coli O157:H7 str. Sakai	BA000007.2	5.6	5229	ECO1
Escherichia coli S88	CU928161.2	5.2	4692	ECS8
Shigella boydii Sb227	CP000036.1	4.9	4133	SBOY
Salmonella enterica subsp. enterica serovar Typhi str. Ty2	AE014613.1	4.8	4314	SETY
Yersinia pestis antiqua	CP000308.1	4.9	4164	YPAN
Pseudomonas syringae pv. tomato str. DC3000	CP000075.1	6.5	5481	PSYR
Vibrio cholerae O395	CP001235.1 CP000626.1	4.1	3875	VCHO
Xanthomonas campestris pv. campestris str. ATCC 33913	AE008922.1	5.1	4179	XCAM

Table 3.2 Breakdown of ortholog pillar structure for each species in the comparison set, where pillars containing 9 orthologs represent loci with an ortholog present in every species, and single ortholog pillars represent species-specific singletons relative to the other species in the set. For example, *E. coli* K12 has 187 singletons and 836 genes with orthologs in all species. Species with a greater percentage of genes in higher number pillars are mapped more successfully against the species set. This is reflected in the identification of more candidate ORFs in species with few singletons (*E. coli, Shigella*) as opposed to species in which a large percentage of the genome lacks identifiable orthologs (*Xanthomonas, Pseudomonas, Vibrio*).

Number of genes in pillar	*E. coli* K12	*E. coli* O157:H7	*E. coli* S88	*Shigella boydii*	*Salmonella enterica*	*Yersinia pestis*	*Peudomonas syringiae*	*Vibrio cholerae*	*Xanthomonas campestris*
9	836	836	836	836	836	836	836	836	836
8	483	483	481	461	477	464	401	379	235
7	639	635	641	595	622	556	228	422	156
6	526	529	526	487	447	370	154	108	93
5	539	555	547	478	448	135	123	93	82
4	443	458	457	281	226	118	123	78	96
3	290	391	352	175	215	195	305	207	234
2	205	446	398	131	344	367	865	395	637
1	187	896	454	689	699	1123	2446	1357	1810
Total genes in species	4148	5229	4692	4133	4314	4164	5481	3875	4179
% of genes in pillars containing 5+ orthologs									
	73%	58%	65%	69%	66%	57%	32%	47%	34%
Number of candidate genes predicted per species									
	272	230	177	487	243	89	85	34	115

3.3.3 Missing genes in the *Shigella boydii* genome annotation

We analysed in detail the results set for *Shigella boydii* strain Sb227 (Yang et al. 2005) (Tables 3.3 and 3.4). While *Shigella* has historically been treated as a different species to *E. coli*, the two genera are actually part of the same, diverse species (Lan and Reeves 2002). SearchDOGS generated an initial candidate list of 487 additional ORFs in this species, including some very short predictions that are unlikely to be real genes. From this initial list Table 3.3 contains 138 candidate unannotated genes listed in decreasing order of length. In this analysis candidates under 90% of the median length of their orthologs were excluded and are not listed, although some of these truncated genes may well be functional; this is a user-adjustable parameter. In most cases, the "low-hanging fruit" – large, intact genes with strong sequence identity to genes in related species – were already correctly identified in the initial annotations. However, we identified 7 gene candidates of length >200 codons, each conserved across a number of species and showing protein sequence conservation, as well as 69 other candidates of length 60-200 amino acids.

For example, *E. coli* K12 *yieP*, coding for a 230 amino acid predicted transcriptional regulator (Riley et al. 2006), has a well-conserved ortholog annotated in each of the *E. coli* strains included (Hayashi et al. 2001; Touchon et al. 2009) as well as in *Salmonella enterica* and *Yersinia pestis* (Deng et al. 2003; Chain et al. 2006). We identified a 230 codon candidate in *S. boydii* in a conserved location (Figure 3.1A) showing very high similarity in length and sequence to the annotated orthologs (Figure 3.1B, 3.1C). Two additional stains of *Shigella flexneri* examined were found to have an annotated ortholog of *yieP*.

We identified 62 candidates of length <60 codons (Table 3.3). 40 of these correspond to unannotated homologs of short genes in *E. coli* K12 and 17 correspond to homologs of genes annotated in other species and predicted to exist in *E. coli* K12 (i.e. *E. coli* K12 also contains a suitable intact ORF). A further 5 are predicted to exist in *S. boydii* but not *E. coli* K12.

21 of the *S. boydii* candidates we identified contained short overlaps (20bp or less) with an adjacent annotated gene. This may have led to the rejection of these ORFs by *ab initio* gene-finding programs despite homology and conserved protein sequence with genes in related species (Aggarwal et al. 2003; Poptsova and Gogarten 2010). It is likely that some of these annotated neighbours that overlap conserved, unannotated genes are spurious, or have incorrectly annotated start coordinates (Palleja et al. 2008; Bakke et al. 2009). We also identified many instances of candidates with non-consensus start codons (9% of annotated genes *in E. coli* K12 are currently annotated as having a TTG or GTG start (Riley et al. 2006) and there are rare instances of bacterial genes starting with TTC, CTG and ATC (Zhu et al. 2004; Riley et al. 2006; Poptsova and Gogarten 2010).

In 33 cases a very highly conserved protein-coding gene of identical length to its homologs could be produced by allowing readthrough of a single stop codon or correction of a single frameshift (Table 3.4). The possibility must be considered that these are not genuine truncation events, but instead are the result of a sequencing or assembly error, particularly in the case of very long genes, or genes encoding proteins that appear to be highly conserved in many related species. An example of an unexpected stop codon occurs in the *S. boydii* orthologs of *E. coli nemA*. This gene codes for *N*-Ethylmaleimide Reductase, one of the "Old Yellow Enzyme" family (Williams and Bruce 2002). It plays a role in the beta-oxidation of fatty acids by being involved in reductive degradation of toxic nitrous compounds (Miura et al. 1997; Umezawa et al. 2008), and is regulated by *nemR*. The *nemA*-*nemR* operon is present in every species in this study except *S. boydii*, where *nemR* is present but *nemA* is annotated as a pseudogene truncated by an in frame stop codon at position 101 of 366 (Yang et al. 2005). By reading through the stop codon, SearchDOGS was able to produce a full length NemA protein showing very high sequence similarity to its annotated orthologs. An intact *nemA* gene is also present in several strains of *Shigella flexneri* we examined (Jin et al. 2002; Nie et al. 2006), indicating that the CAG to TAG transition leading to the premature stop codon in *S. boydii* may be due

to a point sequence error. It is also possible that a readthrough of a genuine stop codon may occur. *E. coli* K12 gene *fdhF*, coding for a subunit of a formate dehydrogenase complex involved in anaerobic respiration, has an annotated homolog in each species studied except *Pseudomonas syringiae* and *Shigella boydii*. The UGA stop codon at position 140 in the *E. coli* K12 protein sequence is translated *in vivo* as selenocysteine under anaerobic conditions (Chen et al. 1992), allowing readthrough of the entire protein, and is indicated as a "U" in the protein sequences of the *E. coli* strains. However the in frame stop codon has led to the rejection of the *S. boydii* homolog in the original annotation (Yang et al. 2005) despite protein sequence conservation over the entire 716 amino acid length of the protein. It appears that an A-C transversion has resulted in a TGC codon coding for a cysteine at that location in *Y. pestis*, *V. cholerae* and *X. campestris* (da Silva et al. 2002; Chain et al. 2006; Feng et al. 2008).

3.3.4 Identification of short bacterial proteins using SearchDOGS

Very small genes are notoriously difficult to accurately identify and annotate by experimental, *ab initio* and homology-based approaches (Basrai et al. 1997; Blattner et al. 1997; Rudd et al. 1998; Ochman 2002; Hemm et al. 2008). Using the February 2010 release of the *E. coli* K12 MG1655 genome (GenBank accession number U00096.2;) as a gold standard and the same set of test genomes (Table 3.1), we tested the ability of SearchDOGS to identify unannotated homologs of short genes showing both conserved sequence similarity and synteny with their annotated counterparts. Among the 113 genes smaller than 60 amino acids annotated in *E. coli* K12, we found that 81 (72%) were not correctly annotated across all of the 8 other genomes in our test dataset (Appendix I: Table S3.1). Many of these are small toxic or membrane-associated proteins recently added to the *E. coli* K12 annotation (Fozo et al. 2008; Hemm et al. 2008) but we also identified genes coding for the 50S ribosomal subunit protein L36 in *E. coli* S88 and *Yersinia pestis*, as well as the gene coding for ribosome-associated protein Sra in *E. coli* S88. The sensitivity of the SearchDOGS method allowed us to accurately predict the location of homologs of the

smallest leader peptide genes annotated in *E. coli* K12, such as a homolog of the 17 codon *hisL* in *Yersinia pestis*, and the 15 codon *pheM* in *Shigella boydii*.

3.3.5 Potential missing genes in the *E. coli* K12 MG1655 annotation?

As a model organism, the *E. coli* K12 MG1655 genome is annotated to a very high standard and is frequently updated (Riley et al. 2006). However, we identified a number of unannotated ORFs showing retention in each of the *E. coli*/*Shigella* strains and a high degree of conservation between the *E. coli* strains, *S. boydii* and often *S. enterica* (Appendix I: Table S2.2). Again the majority of the candidates are short ORFs, ORFs featuring short coding sequence overlaps and ORFs beginning with nonstandard start codons. Due to how closely the *E. coli* and *Shigella* strains are related, it is likely that some of these ORFs are noncoding and happen to be retained by chance, but many are likely to be genuine and warrant study.

One such example is *Shigella boydii SBO_4385*, an 85 codon gene encoding a hypothetical protein which has orthologs of length 88 codons annotated in *E. coli* S88 and *E.* coli O157:H7 (Hayashi et al. 2001; Yang et al. 2005; Touchon et al. 2009). An 85 codon ORF corresponding to a protein of near identical sequence was found by SearchDOGS at an orthologous position in *E. coli* K12 MG1655, and shows statistically significant *Ka/Ks* values of 0.31 and 0.28 against the other *E. coli* orthologs. A 4 bp overlap between the *E. coli* K12 ortholog and the neighbouring gene *yjiL* may have led to it being overlooked.

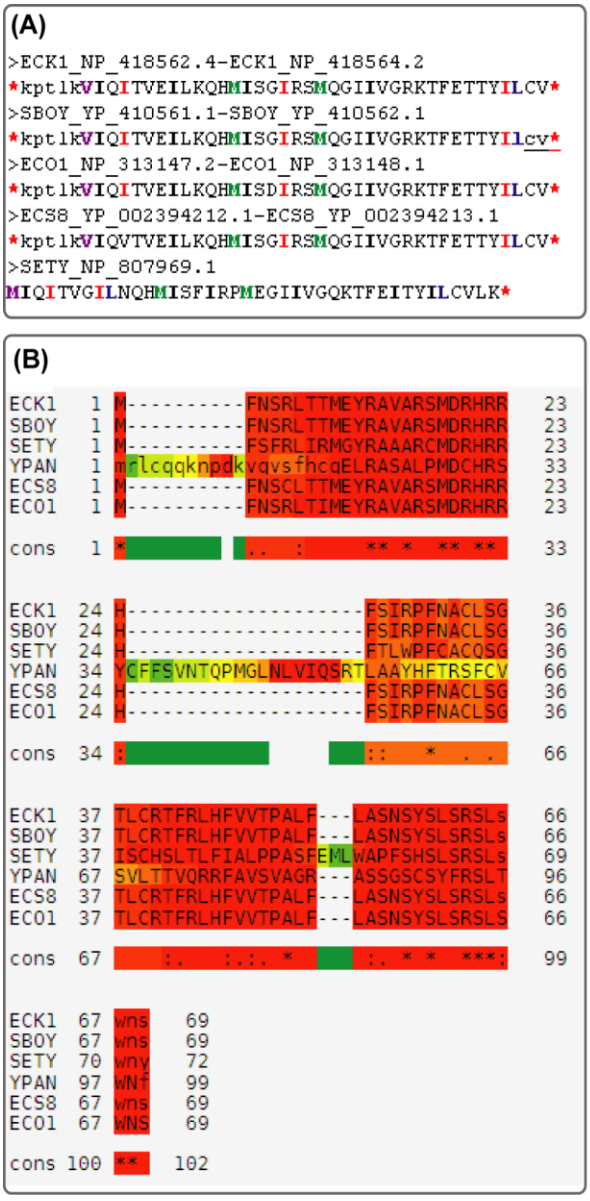

Figure 3.3 Further examples of candidate genes identified by SearchDOGS. (A) *Salmonella enterica subsp. enterica serovar Typhi* str. Ty2 gene *t4378* coding for hypothetical protein NP_807969 (highlighted in yellow) hits ORFs coding for proteins with near-identical protein sequences at a syntenic location in *E. coli* K12, *E. coli* O157:H7, *E. coli* S88 and *S. boydii*. (B) TCOFFEE protein sequence alignment corresponding to an annotated genes in E. coli O157:H7 (bottom sequence) and highly conserved unannotated ORFs at a syntenic location in *E. coli* K12, *E. coli* S88, *S. boydii*, *S. enterica* and *Y. pestis* (Notredame et al. 2000). SearchDOGS also hits a more highly diverged ORF with sequence similarity in *P. syringae* (not shown).

Conservation in many species can be a good indicator that an ORF represents a *bona fide* gene. A 40 codon gene annotated in *S. enterica*, *t4366* (protein id: NP_807959.1), hits near identical unannoted ORFs in *E. coli* K12, *E. coli* 0157:H7, *E. coli* S88 and *S. boydii*, each beginning with the GTG nonconsensus start codon (Deng et al. 2003) (Figure 3.3A). A 69 codon gene (*ECs2526*) annotated in *E. coli* O157:H7, hits highly similar unannotated ORFs in *E. coli* K12, *E. coli* S88, *S. boydii*, *S. enterica* and *Y. pestis* (Hayashi et al. 2001) (Figure 3.3B). In five species the ORF overlaps by 13 bp with the neighbouring gene.

3.3.6 Improving the annotation of the *E. coli* S88 genome

The genome of *E. coli* S88 (Touchon et al. 2009) was annotated using MaGe, a sophisticated annotation pipeline that employs the estimation of synteny conservation both to identify genes and to resolve gene duplicate/fusion/paralog relations (Vallenet et al. 2006). In order to test whether SearchDOGS improves on

the sensitivity of this method, we performed a run testing the *E. coli* S88 annotation against the same version of the *E. coli* K12 MG1655 genome that was available to Touchon et al. (2009) at the time they annotated the S88 genome. This version of the *E. coli* K12 MG1655 genome dates from February 2006 and was obtained from GenBank (accession number U00096.2; version GI:48994873). It lacks several newly-discovered genes (Riley et al. 2006). Using this input, SearchDOGS identified 14 likely gene candidates, 5 of which are less than 60 codons in length. It also identified 7 cases in which frameshift correction or stop codon readthrough creates a highly conserved full-length gene in strain S88 (Table 3.5).

Table 3.5 List of loci at which orthologs of *E. coli* K12 MG1655 genes have been annotated in *E. coli* S88 using the February 2006 annotation of *E. coli* K12 MG1655. Cases in which a single frameshift correction/stop codon readthrough produces a full length gene candidate are also presented.

Neighbouring genes in *E. coli* S88 (protein ids)	length (codons)	*E. coli* K12 ortholog name	*E. coli* K12 protein product (Riley et al. 2006)
YP_002393521.1 YP_002393522.1	242	*yhjH*	EAL domain containing protein involved in flagellar function
YP_002391500.1 YP_002391501.1	180	*infC*	protein chain initiation factor IF-3
YP_002390879.1 YP_002390880.1	124	*yceQ*	predicted protein
YP_002390118.1 YP_002390119.1	109	*ykgJ*	predicted ferredoxin
YP_002392243.1 YP_002392244.1	91	*ypdI*	predicted lipoprotein involved in colanic acid biosynthesis
YP_002391540.1 YP_002391541.1	90	*ynjH*	predicted protein
YP_002392845.1 YP_002392846.1	83	*yqgD*	predicted inner membrane protein
YP_002390794.1 YP_002390795.1	76	*ymcE*	cold shock gene
YP_002391115.1 YP_002391116.1	71	*hokD*	Qin prophage; small toxic polypeptide
YP_002390124.1 YP_002390125.1	46	*ykgO*	rplJ (L36) paralog
YP_002394547.1 YP_002394548.1	46	*yjjY*	predicted protein
YP_002390959.1 YP_002390960.1	46	*ylcG*	DLP12 prophage; predicted protein
YP_002393273.1 YP_002393274.1	37	*rpmJ*	50S ribosomal subunit protein L36
YP_002390560.1 YP_002390561.1	37	*ybgT*	conserved protein
Frameshift correction/Stop readthrough allowed			
YP_002390825.1 YP_002390826.1	807	*putP*	proline:sodium symporter
YP_002393391.1 YP_002393392.1	532	*rtcR*	sigma 54-dependent transcriptional regulator of rtcBA expression
YP_002390588.1 YP_002390589.1	477	*ybhI*	predicted transporter
YP_002392075.1 YP_002392076.1	443	*yfaV*	predicted transporter
YP_002390454.1 YP_002390455.1	386	*ybdL*	methionine aminotransferase, PLP-dependent
YP_002391213.1 YP_002391214.1	350	*ycjQ*	predicted oxidoreductase, Zn-dependent and NAD(P)-binding
YP_002390681.1 YP_002390682.1	171	*ybjP*	predicted lipoprotein

Automated annotation programs can miss or misannotate certain genes due to unusual start codons or nucleotide composition in these genes. The essential translation initiation factor IF3 is encoded by *infC,* one of only two *E. coli* genes known to start with the rare start codon ATT (Binns and Masters 2002). IF3 plays a role in the fidelity of translation initiation, and has been shown to regulate the frequency of initiation from non-canonical start codons (Meinnel et al. 1999; Maar et al. 2008). In this fashion it functions in a negative feedback loop, repressing its own translation when in abundance (Butler et al. 1986; Binns and Masters 2002). We identified a full-length ORF corresponding to *infC* in *E. coli* S88. A truncated *infC,* beginning at a downstream GTG is annotated as a pseudogene in the *E. coli* S88 genome, presumably because MaGe does not consider ATT as a possible start codon. An incorrect start was also annotated for this locus in the existing S. *boydii, Y. pestis, V.cholerae, P. syringiae* and *X. campestris* annotations; these genes all appear to begin with an ATT start codon, although they are annotated with different start codons. This highlights the need for manual curation or more rigorous comparative genomic approaches to correctly annotate loci with unusual features.

3.3.7 Identification of a possible gene fusion in *Xanthomonas campestris*

We identified a possible case of gene fusion within a biosynthetic operon in *Xanthomonas campestris*. The *Xanthomonas campestris* gene *cobC* (which is called *cobD* in *Salmonella* (Brushaber et al. 1998)) codes for a 327 amino acid enzyme in the cobalamin (vitamin B12) biosynthesis pathway. This gene has no annotated homolog in *X. campestris* pathovar campestris strain ATCC 33913 but is annotated in two other *X. campestris* pathovars, campestris strain B100 and vesicatoria strain 85-10 where it overlaps by 4 bases with the downstream *cobQ* (Thieme et al. 2005; Vorholter et al. 2008). The apparent gene fusion occurs in *X. campestris* pathovar campestris strain ATCC 33913. Here, a single base pair insertion near the end of the *cobC* ORF has led to the loss of the stop codon, and has also brought it into frame with *cobQ* creating what appears to be an 817 codon CobCQ fusion protein (Figure

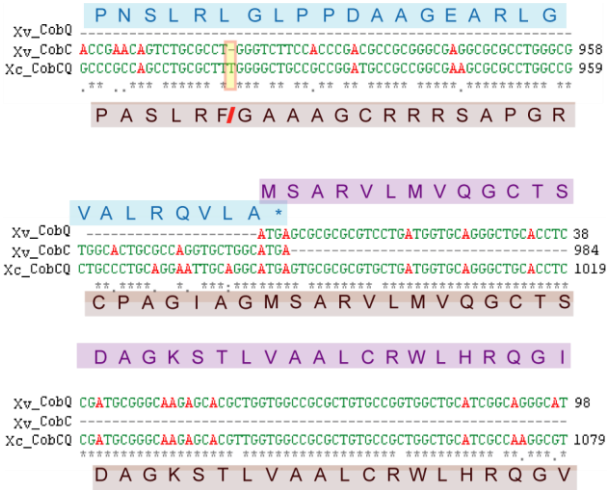

Figure 3.4 Possible gene fusion in the *X. campetris* cobalamin synthesis pathway. Nucleotide alignment of the point of overlap between the adjacent *Xanthomonas campestris pv. vesicatoria* genes *cobC* and *cobQ* (Xv_CobC, Xv_CobQ) against the *Xanthomonas campestris pv. campestris str. ATCC 33913* fused *CobCQ* ORF (Xv_CobCQ) (Larkin et al. 2007). Protein sequences of Xv_CobQ shown above in blue, Xv_CobC in purple, and the Xv_CobCQ fusion below in brown. A single base pair insertion towards the end of the *X. campestris pv campestris cobC* gene (highlighted in yellow) has abolished the stop codon at the end of this ORF and has brought it into frame with the downstream *CobQ* creating a gene fusion.

3.4). *De novo* cobalamin production from uroporphyrinogen III is a complex pathway involving 30 or so enzymes; however some bacteria are able to produce cobalamin from pathway intermediates. The *Salmonella* and *Pseudomonas* strains can produce cobalamin *de novo* (Zhang et al. 2009), whereas *E. coli* and *Shigella* strains can only produce cobalamin from the intermediate molecule cobinamide and are missing the early genes in the pathway (a set of genes known as the CobI genes) (Lawrence and Roth 1995). *Xanthomonas* is widely reported to produce cobalamin and is used in industrial production (Martens et al. 2002) but also appears to lack the full pathway (Zhang et al. 2009). However, it appears to have fully retained the 9 genes of the cobA-S operon, and thus appears capable of producing cobalamin from the intermediate hydrogenobyrinic acid (Raux et al. 1996). The cobalamin pathway differs extensively even between *de novo* cobalamin-producing bacteria (Raux et al. 1996), so it is possible that *Xanthomonas* can perform other steps of the pathway using non-homologous proteins.

3.3.8 Identification of pseudogenes

A significant problem associated with homology-based automated annotation methods is the difficulty in differentiating *bona fide* unannotated genes from pseudogenes that share both sequence similarity and location with their intact

orthologs in other species. SearchDOGS tries to overcome this by providing as much information as possible in order for the user to make an informed choice on whether to accept, reject, or further study a candidate. In theory, low *Ka/Ks* ratios between a candidate gene and its orthologs across a range of evolutionary distances provide strong evidence of protein conservation (Yang and Bielawski 2000). However, even a *Ka/Ks* <1 does not guarantee the existence of a functional gene, because recently-formed pseudogenes will still bear the hallmarks of sequence constraint to code for protein (Ochman and Davalos 2006). Furthermore, *bona fide* unannotated genes will often not return statistically significant values in these tests if they are short or are too highly similar in nucleotide sequence to their orthologs. One pragmatic approach is to ignore any potential candidates that are shorter than 80% of the median length of annotated orthologs in closely related species (Lerat and Ochman 2004); however there is no guarantee that such a truncation will render a gene's protein product nonfunctional.

For illustration, *ilvG* is a recently formed pseudogene in *E. coli* K12 MG1655. The functional gene codes for a catalyst of 2-acetolactate synthesis from pyruvate (Hayashi et al. 2001), and exists as a 549-574 codon gene in *E. coli* S88, *E. coli* 0157:H7, *S. boydii*, *S. enterica*, *Y. pestis*, *V. cholera* and *X. campestris* (Hayashi et al. 2001; da Silva et al. 2002; Deng et al. 2003; Yang et al. 2005; Chain et al. 2006; Feng et al. 2008; Touchon et al. 2009). In *E. coli* K12 MG1655 a frameshift at codon 328 (Ecogene; (Rudd 2000)) has resulted in a truncated 329 codon pseudogene which was identified as a candidate gene by SearchDOGS. *Ka/Ks* ratios between the annotated orthologs and the hypothetical translation of the pseudogene are low (0.04 to 0.18 for the statistically significant values), and there is very high protein sequence similarity between the annotated orthologs and the pseudogene. It appears that *ilvG* is the victim of a recent pseudogenization event in *E. coli* K12, and has not degenerated enough to lose the evolutionary characteristics of a bona-fide protein-coding gene.

We identified one interesting example in which an apparent expansion of a 7bp sequence appears to have led to the pseudogenisation of a highly conserved gene in

Shigella boydii. Degradation of fatty acids under aerobic conditions in *E. coli* is carried out by the FadBA enzyme complex, a tetramer made up of proteins encoded by *fadB* and *fadA* (Binstock et al. 1977; Pramanik et al. 1979), as part of the *fadR* regulon (Cho et al. 2006). *fadB* and *fadA* are highly conserved throughout both Gram positive and Gram negative bacteria (Fujita et al. 2007), and both genes are present in all of the 8 species in our study set with the exception of *S. boydii*, where *fadB* is annotated as a truncated pseudogene containing a frameshift (Yang et al. 2005). By correcting the frameshift, SearchDOGS was able to automatically predict a full length (738 amino acid) FadB protein showing very high sequence similarity to its annotated orthologs (Figure 3.2B). This full-length construct shows *Ka/Ks* values ranging from 0.02 to 0.19 when compared to its orthologs (Figure 3.2A). Closer examination of the *S. boydii fadB* nucleotide sequence shows that the frameshift is caused by a 28 bp insertion consisting of four tandem repeats of a 7bp repeat (Figure 3.2C). This 7bp repeat matches a 7bp segment that is imperfectly repeated in *E. coli* K12 (6 out of 7 bases match); in *S. boydii* 2 copies have expanded to 6. As well as being intact in all the *E. coli* strains studied, *fadB* is also intact in several strains of *Shigella flexneri* we examined (Jin et al. 2002; Nie et al. 2006). Without experimental evidence it is unclear whether the *fadB* truncation in *S. boydii* is due to an error in sequencing or assembly, or a genuine pseudogenisation event.

Obligate and facultative pathogens have been shown to harbour a relatively high number of pseudogenes for reasons that may be due to ongoing genome reduction as an adaptation to the host environment (Mira et al. 2001). The inclusion of *E. coli* K12 MG1655, a reference sequence with a high degree of annotation accuracy (Riley et al. 2006) allowed us to identify many cases of what appear to be pseudogenes incorrectly annotated as real genes in the genomes studied. Table 3.6 is a list of likely pseudogenes identified in the genomes studied based on a high degree of similarity in length and sequence to known *E. coli* K12 pseudogenes. These genes are automated predictions, and most represent a truncated form of an ortholog present in one of the other species. (Table 3.6 is likely to represent only a subset of all incorrectly annotated pseudogenes in these genomes. It does not include those automatic gene

predictions that appear to be truncated orthologs of full-length genes in other species where there is no *E. coli* K12 feature at the locus to compare against.)

Table 3.6 Set of genes annotated in the species studied that are likely to be pseudogenic based on length and sequence similarity to known pseudogenes in *E. coli* K12 MG1655. Each gene in this list hit a syntenic region in *E. coli* K12 containing a pseudogene. Pseudogene descriptions provided from the GenBank reference file (Blattner et al. 1997; Riley et al. 2006) or Ecogene (Rudd 2000)

E. coli K12 pseudogene name	Annotated homologs likely to be pseudogenic	Description of *E. coli* K12 pseudogene (Rudd 2000; Riley et al. 2006)
yedS	*S. boydii* YP_407522.1	*Salmonella* OmpS1 homolog.
yhiL	*S. boydii* YP_409799.1	An intact version of YhiL is present in *E. coli* O157:H7 as Z4888. The yhiL gene can be transcribed in vitro with sigma28 (FliA) holoenzyme (Yu 2006)
yaiT	*S. boydii* YP_406812.1	First 27 aa predicted to be a signal peptide.
insZ	*E. coli 0157:H7* NP_313293.1 *S. boydii* YP_408257.1	Two frameshifts (at codons 62 and 111) and an internal deletion of about 150 codons have mutated this homolog of IS4 transposase InsG (442 aa)
ygeQ	*E. coli 0157:H7* NP_311764.1	Remnant of the type three secretion system (T3SS) pathogenicity island ETT2.
bscQ	*E. coli 0157:H7* NP_312441.1 *S. boydii* YP_409845.1	Stop codon 6 is translated as an X in the reconstructed protein sequence; other *E. coli* strains have a Leu codon at this position. bscQ(yhjQ) is a member of the minD superfamily.
yghE	*P. syringiae* NP_794443.1 *E. coli S88* YP_002394339.1 *V. cholerae* YP_001215282.1	The yghFED operon appears to have suffered a deletion of the gspDEFGHIJK homologs (7403 bp) between the gspC-like (yghF) and the gspLM-like (yghED) genes. The stop codon of yghF was removed, fusing 12 C-terminal residues out-of-frame but overlapping part of the fused yghE gene. The N-terminal 74 residues of yghE were removed by the deletion event.
yejO	*S. boydii* YP_408523.1 *E. coli 0157:H7* NP_311108.1	IS5K inserted at codon 21 and made a 4 bp target site duplication TTAT. The first 29 aa are predicted to be a signal peptide
yjbI	*S. boydii* YP_410341.1	YjbI' and YjcF belong to COG1357. Apparent frameshifts at codons 62 and 86 were repaired to make a hypothetical reconstruction.
ydfJ	*E. coli 0157:H7* NP_310179.1	The first 28 codons of ydfJ were separated by the insertion of 20,460 bp of the Qin prophage; 28 aa (translated from 1650862 to 1650779 bp) have been added back to the YdfJ protein sequence presented. An intact version is present in *E. coli* 536 (UniProtKB: Q0THP5).
mdtQ	*E. coli S88* YP_002391970.1	First 21 aa are predicted type II signal peptide. An apparent frameshift at codon 51 has been reconstructed.
yfdL	*E. coli 0157:H7* NP_308310.1 *E. coli S88* YP_002392505.1	"pseudogene, CPS-53 (KpLE1) prophage; Phage or Prophage Related" (.gbk)
ylbH	*E. coli 0157:H7* NP_308590.2	pseudogene, rhs-like (.gbk)
cybC	*S. boydii* YP_410464.1	pseudogene, truncated cytochrome b562 (.gbk)
pinH	*S. boydii* YP_409208.1	pseudogene, predicted invertase fragment (.gbk)
yeeW	*E. coli 0157:H7* NP_310834.1 *E. coli S88* YP_002391799.1	CP4-44 prophage; predicted protein;Phage or Prophage Related (.gbk)
ydeT	*E. coli 0157:H7* NP_310137.1	Outer membrane fimbrial subunit export usher protein FimD family.
yneO	*E. coli 0157:H7* NP_310144.1	pseudogene, AidA homolog
ycgH	*E. coli S88* YP_002391008.1	Probable pseudogene; putative ATP-binding component of a transport system (.gbk)
yddK	*E. coli 0157:H7* NP_310102.1	A deletion has apparently removed the 5' end of yddK and the 3' 273 codons of yddL.
lfhA	*E. coli 0157:H7* NP_308283.1 *E. coli S88* YP_002390098.1	Intact *E. coli* O42 allele: SP Q5DY37. The *E. coli* K-12 lfhA pseudogene is missing the first 127 codons.

yghO	*E. coli S88* YP_002392963.1	pseudogene, DNA-binding transcriptional regulator homology
yegZ	*E. coli O157:H7* NP_310917.1	yegZ is adjacent to the ogrK copy of the P2 ogr gene, indicating the presence of a P2-like prophage remnant. Intact alleles are present in several *E. coli* strains and *Yersinia pestis* phage L-413C (UniProtKB: Q858U5).
ydfE	*E. coli S88* YP_002391362.1	Qin prophage; pseudogene;Phage or Prophage Related
yibS	*E. coli O157:H7* NP_312499.1	Four stop codons (3, 11, 25, 27)
arpB	*E. coli S88* YP_002391502.1 *E. coli O157:H7* NP_310454.1	A frameshift at codon 142 is translated as an X in the reconstructed protein sequence. An intact allele is present in O157:H7 EDL933 as Z2749, which has K142.
yjhZ	*E. coli S88* YP_002394396.1	An inframe stop codon at position 44 was translated as an X for the reconstruction. An intact version of YjhZ is present in Escherichia sp. 3_2_53FAA as ESAG_039().
yhdW	*E. coli S88* YP_002393250.1 *S. boydii* YP_409586.1	An apparent frameshift mutation at codon 23, as compared to other alleles and homologs of this gene, is translated as H23 in the reconstructed protein sequence since this position is a His residue in all the intact *E. coli* alleles.
ybfQ	*E. coli O157:H7* NP_308758.1	N-terminal domain fragment, matches first 79 residues of paralogs YhhI, YdcC, YbfD, pseudogene YbfL, and the more distant pseudogene paralog YncI
bcsQ	*E. coli O157:H7* NP_312441.1	Stop codon 6 is translated as an X in the reconstructed protein sequence; other *E. coli* strains have a Leu codon at this position
rhsE	*S. boydii* YP_406937.1	seudogene, rhsE element core protein RhsE
ybfG	*S. boydii* YP_407071.2	An in-frame stop at codon 70 is replaced with an X in the reconstruction. An intact allele is found in *E. coli* 53638 as Ecol5_01004515 (GenBank gi:75511145).
yhiS	*S. boydii* YP_409817.1	IS5T inserted at codon 249 and made a 4 bp target site duplication TTAG. *E. coli* O157:H7 YhiS (Z4907) has no IS5 and has a frameshift near the C-terminus relative to K-12. the *S. flexneri* version (SF3539) has a similar C-terminus to the K-12 version.
ybeM	*S. boydii* YP_407019.1 *E. coli S88* 002390478.1	1bp deletion at codon 66
yqfE	*E. coli S88* YP_002392819.1	An inframe stop codon at position 19
ykiA	*S. boydii* YP_406830.1 *E. coli 0157:H7* NP_308469.1 *E. coli S88* YP_002390215.1	An intact 759 aa version of YkiA is present in *E. coli* B185 (UniProt:D6I6K1)
ymdE	*E. coli O157:H7* NP_309324.1	Pseudogene

3.4 Discussion

A crucial cog in the SearchDOGS method is the ascertainment of orthology of genomic segments in multiple species accurately and unambiguously. For this to be possible for a given set of species, several conditions must be met. Species included cannot have diverged too far from each other, otherwise rearrangements and gene gain/loss will obscure attempts to identify conserved synteny. In the SearchDOGS analysis, the five species with the most candidate hits were also the species with the highest numbers of genes in ortholog pillars of 5 or more members; *i.e.* the species that could be mapped most effectively against the majority of other species (Table 3.2) However if a set of species with too narrow a divergence range is chosen, it is often difficult to get a clear signal of protein sequence conservation (Ochman and Davalos 2006).

In addition to a suitable evolutionary range, several other factors must be considered when choosing a comparison species set. SearchDOGS only identifies unannotated genes for which an ortholog exists and is annotated in another species. Hence we suggest that at least one closely related species with a very high quality of annotation is used for reference. Gene content in a genome is also affected by lifestyle and environment. Furthermore, genes may be annotated in certain strains that have been missed in others due to different annotation methods, differing degrees of sequencing and annotation rigor, or studies carried out over a different range of conditions.

One of the greatest challenges when using sequence homology-based techniques to predict genes is differentiating *bona fide* genes from pseudogenes. A pseudogene can be incorrectly annotated as a functional gene for two reasons. Firstly, a genuine gene can hit a pseudogene showing sequence similarity to the query. If the pseudogenisation event is sufficiently recent, the remaining gene fragment may still bear many of the hallmarks of a genuine gene, including protein conservation (Ochman and Davalos 2006). Secondly, it is likely that many pseudogenes in sequenced genomes are currently incorrectly annotated as if they were functional.

This is evidenced by the number of "genes" in other strains/species we identified that are identical in sequence to known *E. coli* K12 pseudogene features (Table 3.6). As many automatic annotation procedures including SearchDOGS rely heavily on sequence similarity to a reference set of annotated features, these "false genes" in the reference set can lead to the spurious misannotation of other pseudogenes, spreading false annotations through databases if unchecked (Hemm et al. 2008). Brown and Sjolander estimated in 2006 that only 3% of those proteins in the UniProt database not labeled as "hypothetical" or "unknown" had experimental support (Brown and Sjolander 2006), the remainder having inferred by bioinformatics means, and this percentage is surely considerably lower by now.

A third problem lies in correctly differentiating genuine frameshift mutations (creating pseudogenes) from sequencing errors. Current next-generation sequencing methods such as Roche/454 and Illumina have a higher background rate of error than previous methods such as Sanger sequencing (Margulies et al. 2005; Quinlan et al. 2008; Farrer et al. 2009; Kircher and Kelso 2010), particularly in genomes sequenced to low coverage, and these errors mainly take the form of single nucleotide insertions and deletions (indels). Bacterial SearchDOGS is designed to correct a single frameshift, if HSP evidence indicates that a sensible and conserved full-length protein can thereby be created. However, as with other gene disruption events a gene in which a genuine frameshift mutation has occurred recently will have many of the same characteristics, such as a low *Ka/Ks* ratio, as a genuine gene containing a sequencing error.

For each species in the input set, SearchDOGS generates a list of automatic predictions based on sequence similarity and conserved genomic location. Our aim was to include as much information as possible in order for the user to make an informed decision as to whether to reject a candidate gene or accept it for further study. Users are encouraged to look at the length, sequence similarity and level of protein conservation of a candidate relative to its annotated orthologs. However, full

proof that a candidate gene genuinely codes for a functional protein requires detection of the protein translation product, for example by mass spectrometry.

As the tools of proteomics come of age (Ansong et al. 2008a; Armengaud 2009) and bioinformatics methods become ever more sophisticated, these approaches combined should result in fast, accurate and complete genome annotations to complement the accelerating pace of genome sequencing.

3.5 Methods

3.5.1 Generation of ortholog pillars.

Pillars of orthologs are generated using a two-step process. For each genome studied, pairwise BLASTP searches are performed using the protein sequence of each gene as a query against the protein set of each other species. Pillars of reciprocal best BLASTP hits are generated. In this first stage, a reciprocal best BLASTP hit is required for each ortholog against every other ortholog in the pillar, and a cutoff value of e-5 is used in the BLASTP searches.

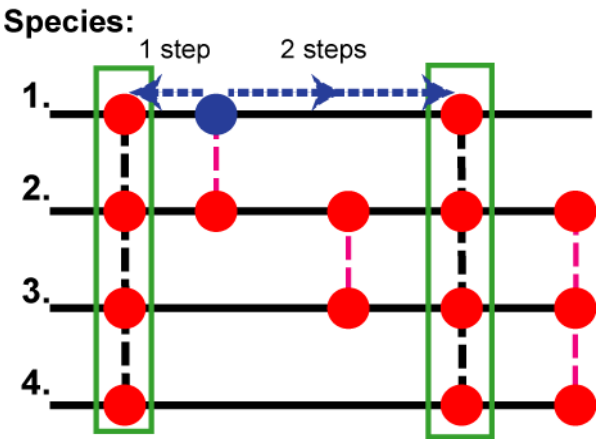

Figure 3.5 Bacterial SearchDOGS's pillar generation method. A set of initial pillars (highlighted in green) is created using reciprocal best BLASTP results. The orthology of genes producing one way BLASTP hits is confirmed by automatically searching for shared synteny. In the example above, the gene from species 1 highlighted in blue has neighbouring genes that are in ortholog pillars with neighbouring genes of the gene from species 2, confirming that they both belong in a single ortholog pillar.

The initial stage of pillar generation is followed by a "Syntenoblast" approach (Byrne and Wolfe 2005) to try to place the remaining genes in pillars (Figure 3.5). A second round of BLASTP searches is carried out using a permissive cutoff value of *E*=10. For each gene that hits a potential ortholog in another species, the syntenic context of the gene is evaluated. Five "steps out" are performed along the genome of the query and the genome of the hit. If the query and the hit are found to share an ortholog pillar, then they are deemed to be supported by shared synteny and are put in a single ortholog pillars. In this stage full reciprocality of BLASTP hits with all of the genes in the pillar hit is not required. In cases where there are multiple candidates for a single pillar position, the pillar member is chosen by strength of shared synteny, and then by the *E* value of the BLASTP hit. This process is run in an iterative fashion, until each gene has been labeled a singleton or has been placed in an ortholog pillar with synteny and BLASTP support.

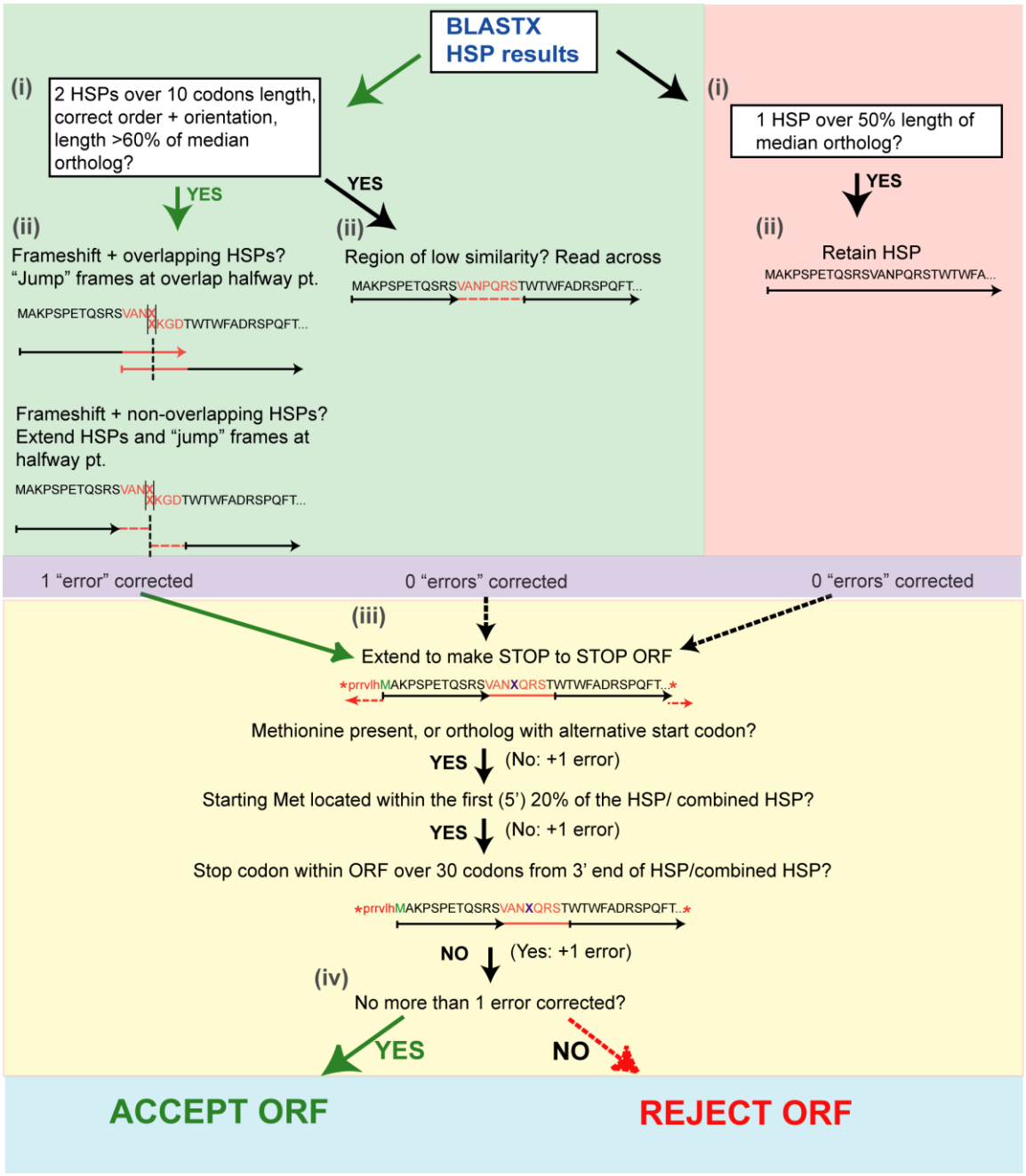

Figure 3.6 Flowchart illustrating the process by which ORFmaker identifies and accepts a stop to stop ORF. (i), (ii) If the BLASTX search produces two HSPs or sufficient cumulative length that are in the correct orientation and order, ORFmaker attempts to create a single ORF spanning both by reading across between the HSPs or by correcting a frameshift. If a single HSP of sufficient length is produced by the BLASTX search, this is retained. (iii) The construct is extended outwards from each direction until a stop codon is reached, creating a stop to stop ORF. The ORF is tested for errors such as the location of the starting methionine and presence of premature stop codons. (iv) ORFmaker accepts ORFs with a single error, whether it is a frameshift, premature stop or lack of start, to allow for the possibility of a sequence error. ORFs with two or more errors are rejected. In the instance where an ORF passing criteria can be made using a single HSP or a two-HSP construct, the ORF producing the highest BLASTP bit score against its orthologs when translated is retained.

3.5.2 Generation of database and SearchDOGS search procedure.

The procedures involved in establishing synteny between genomic segments across species, generating the SearchDOGS database, and searching against this database are as described for the yeast implementation of SearchDOGS in Chapter 2.

3.5.3 Generation of candidate open reading frames.

For each intergenic region that hits one or more genes from a syntenic homology pillar in the BLASTX search, the high similarity pairs (HSPs) corresponding to the hit with the lowest *E* value are retained. Open reading are obtained using ORFmaker, an inbuilt ORF-finder specifically designed for use with Bacterial SearchDOGS (Figure 3.6).

3.5.4 Evidence for protein conservation

Omega (*Ka/Ks*, where Ks is the number of synonymous substitutions per synonymous site and Ka is the number of nonsynonymous substitutions per nonsynonymous site (Li et al. 1985)) values are provided for each candidate ORF as a measure of protein conservation. The largest standard start-to-stop ORF within the stop-to-stop ORF is used in this test provided no genes with non-consensus start codons exist within the corresponding ortholog pillar. Due to the difficulty in predicting non-consensus start codons with certainty, in these cases only the length of ORF from the start of the HSP to the end of the ORF is used. Pairwise comparisons are performed between each ORF and every gene in the corresponding ortholog pillar using the program yn00 from the PAML package (Yang 2007). As calculating an accurate standard error measurement for *Ka/Ks* is problematic, the value and standard error of the difference between *Ks* and *Ka* (*Ks-Ka*) are calculated in order to test the statistical significance of these results. Assuming neutral evolution, a *Ks-Ka* value of approximately 0 is expected, and a *Ks-Ka* value significantly greater than 0 indicate

constrained protein evolution (Yang and Bielawski 2000). A 95% confidence interval for (*Ks*-*Ka*) is calculated using the following formula:

$Ks\text{-}Ka \pm 1.96(\mathrm{SE}(Ks\text{-}Ka))$

where SE(*Ks*-*Ka*) is the standard error of (*Ks*-*Ka*) and is calculated as follows:

$\mathrm{SE}(Ks\text{-}Ka) = \sqrt{[(\mathrm{SE}(Ka)^2) + (\mathrm{SE}(Ks)^2)]}$.

Two problems associated with the *Ka/Ks* test must be noted:

- *Ka/Ks* values are often not statistically significant for genes which are short or have very similar nucleotide sequence due to an insufficiency of informative sequence.
- *Ka/Ks* values for candidates with a potential non-consensus start codon are only approximate and may not be entirely accurate if a significant length of ORF upstream of the HSP is excluded from the calculation.

Chapter 4

Comparative analysis of programmed ribosomal frameshifting sites in yeast chromosomal genes

4.1 Abstract

Programmed ribosomal frameshifting refers to the phenomenon by which sequences within certain genes have become optimised to stimulate what normally represents an error by the translational machinery: changing of the reading frame during translation of mRNA by the ribosome skipping forward over a base (+1 frameshift) or repositioning backwards (-1 frameshifting). Examples of +1, -1 and -2 programmed frameshifting have been identified.

In this chapter I examine three examples of programmed ribosomal frameshifting in yeast chromosomal genes. These three genes, *EST3*, *OAZ1* and *ABP140*, were previously known to contain frameshifting signals; we extend previous comparative analyses to examine the extent of frameshift signal conservation at these loci and to identify the phylogenetic point at which the frameshift is introduced in the Saccharomycetaceae. In the case of *ABP140*, I identify previously unidentified cases of ohnolog retention following whole genome duplication. I also describe a fourth example of unusual gene evolution that may be explained either by a gene split or the introduction of a programmed ribosomal frameshift. The *URA6* locus is unusual in that, in either scenario, the distribution of intact and split genes we see in the species studied can only be explained by number of separate gene-splitting/frameshift-introduction events.

With species from 11 of the 12 currently defined Saccharomycetaceae clades included, this represents the most comprehensive comparative study of programmed ribosomal frameshifting in yeast to date.

4.2 Introduction

The genetic code, first deciphered in the 1960s by Nirenberg, Leder, Khorana and others (Nirenberg et al. 1966), is conserved to a remarkable degree across life on earth. However it features slight variations in different organisms and organelles. Mitochondria and chloroplasts employ a number of variations on the standard code including variant stop codons (Osawa et al. 1989; Jukes and Osawa 1990); bacteria use three different start codons (ATG, TTG and GTG) at high frequency and a further three (CTG, ATT, ATC) in rare circumstances (Zhu et al. 2004; Riley et al. 2006; Poptsova and Gogarten 2010); and there are instances of codon reassignment in certain species groups, such as the CTG group of yeasts (including *Candida albicans*) in which CTG has been reassigned to code for serine rather than the standard leucine (Ohama et al. 1993; Sugita and Nakase 1999; Fitzpatrick et al. 2006). In addition to this, the way in which the genetic code is read has evolved variations.

Recoding refers to the phenomenon of dynamic genetic signals that have evolved in order to stimulate the decoding of mRNA in a non-standard way at specific sites in genes (Atkins and Baranov 2010). These alternative decoding events are in competition with the standard decoding system in the organism. An example of one type of recoding is the redefinition of UGA stop codons to specify selenocysteine in selenoproteins. This is distinct from codon reassignment in that it allows readthrough of stop signals and the creation of selenocysteine-containing proteins from only a specific subset of genes in the cell (Atkins and Baranov 2010). A different type of recoding, only one example of which has to date been identified, involves the ribosome being induced to disassociate and reattach to the mRNA at a downstream triplet, causing a length of mRNA sequence to be bypassed in translation. In the bacteriophage T4 gene *60*, a 50 nucleotide region is skipped by about 50% of ribosomes during translation, with the other 50% translating the mRNA as normal (Maldonado and Herr 1998; Herr et al. 2000). Recoding can also occur at the level of transcription, with a mechanism known as transcription slippage resulting in

transcripts containing inserts of 1 to 15 additional nucleotides or short deletions (Atkins 2010).

The focus of this chapter is on a type of recoding at the level of translation, in which an alternate protein product is produced by the ribosome shifting from the standard 0 frame into the +1 or -1 frame at a specific point in translation, and translating the rest of the product in the new frame. Examples of -2 frameshifts have also been documented (Xu et al. 2004); these are identical to +1 frameshifts in terms of frame change, but result in the inclusion of an extra amino acid and necessitate a different mechanism (Ivanov et al. 1998b).

Translation of messenger RNA by the ribosome is highly efficient but results in a low frequency of errors, such as missense errors (where an incorrect amino acid is incorporated), processivity (where the ribosome stalls and "falls off", resulting in premature termination of mRNA translation) (Vimaladithan and Farabaugh 1994) and spontaneous frameshifting.

Translational accuracy is maintained by multiple mechanisms. Kinetic proofreading selects against incorrect incoming aminoacyl tRNAs based on their rapid dissociation in steps before and after GTP hydrolysis by EF-1A (Thompson 1988). An "induced fit" mechanism involves cognate aminoacyl-tRNAs (aa-tRNAs) inducing a change in the structure of the ribosome resulting in an acceleration of the rate of their acceptance relative to non-cognates (Rodnina et al. 2005). However, there is a tradeoff that occurs between translational accuracy and speed, and to maintain the necessary rate of elongation the rate of GTP hydrolysis becomes fast enough that some translational accuracy based on differences in stability between correct and incorrect decoding events is sacrificed (Gromadski and Rodnina 2004) .

Spontaneous frameshifting is very rare. The rate at which it occurs in bacteria has been estimated at under 3×10^{-5} per codon (Kurland 1992). However, certain genes contain sequences that manipulate the translational machinery to increase this rate up

to four orders of magnitude (Farabaugh 1996; Namy et al. 2004). The majority of programmed frameshifting events identified to date have been in RNA viruses (Brierley 1995; Dinman 1995; Plant and Dinman 2008), including -1 frameshifts in the Rous sarcoma virus (Jacks and Varmus 1985) and the HIV-1 retrovirus (Jacks et al. 1988b). In many of these cases frameshifting occurs in a gene where an N-terminal domain and a C-terminal domain are encoded by two overlapping open reading frames in different frames (Hammell et al. 1999). Frameshifting allows a fusion protein to be made including both domains, and the ratio of N-terminal-only protein product to full-length product is controlled by the frequency of successful frameshifting.

One of the earliest examples of frameshifting identified was in the Gag-Pol gene of the Ty1 retrotransposon in *S. cerevisiae* (Clare and Farabaugh 1985). This transposable element transposes through an RNA intermediate using a reverse transcriptase encoded by the *pol* gene (Boeke et al. 1985). The reaction occurs inside a virus-like particle made up of structural protein components encoded by the *gag* gene. A single mRNA containing the adjacent *gag* and *pol* genes is transcribed. The *gag* and *pol* ORFs overlap, and the *pol* ORF lacks a starting AUG (Clare and Farabaugh 1985). Using mutagenesis a minimal frameshift sequence in the 38 nt *gag*/*pol* overlap region, CUU-A-GGC, was identified, where the A is skipped as the ribosome shifts to the +1 frame (Belcourt and Farabaugh 1990). The percentage frequency at which frameshifting occurs allows a 50:1 stochiometric ratio of Gag to full-length Gag-Pol to be maintained (Dinman and Wickner 1992), which is crucial to efficient transposition (Xu and Boeke 1990; Kawakami et al. 1993). The Ty2, Ty3 and Ty4 elements also use +1 frameshifting to produce a Gag-Pol fusion protein (Belcourt and Farabaugh 1990; Stucka et al. 1992; Farabaugh et al. 1993), while the yeast L-A double-stranded RNA virus uses a -1 frameshift to produce a fusion of proteins that can be considered functional analogs of Gag and Pol (Icho and Wickner 1989; Dinman et al. 1991).

While the end product is the same, several different frameshift-stimulating sequences have been identified, and different mechanisms have been proposed for how frameshifting is brought about in these genes. In Ty1, the frameshift occurs when the ribosomal P site is occupied by the CUU triplet (the first three bases of the CUU-A-GGC heptamer). At the A site, the ribosome selects a tRNA recognising the GGC triplet in the +1 frame instead of the AGG 0-frame triplet, and then continues to translate thereafter in the +1 frame (Belcourt and Farabaugh 1990). Several features of the tRNAs involved in decoding this heptamer appear to play a role in stimulating the ribosome to switch frames. The CUU P-site codon is decoded in *S. cerevisiae* by $tRNA^{Leu}_{UAG}$, which is unusual in that its wobble uridine is unmodified. This allows it to decode all six Leu codons, but the weak U-U pair that is formed with the third base in CUU result in a weaker-than normal interaction in this case. The tRNA decoding the 0 frame AGG ($tRNA^{Arg}_{CCU}$) at the A site is relatively rare, and competition for the A site is provided by the tRNA decoding GGC. Experiments by Belcourt and Farabaugh show that overexpression of $tRNA^{Arg}_{CCU}$ drastically reduced frameshifting, as did expression of a novel tRNA making a normal Watson-Crick pairing with CUU at the P site (Belcourt and Farabaugh 1990). These observations are consistent with a model proposed by Belcourt and Farabaugh in which a pause in translation is induced by the slow decoding of the AGG triplet by $tRNA^{Arg}_{CCU}$ at the A site, allowing the $tRNA^{Leu}_{UAG}$ at the A site to "slip" from CUU to UUA in the +1 frame at the P site. Simultaneously, competition from the tRNA decoding GGC results in the ribosome shifting into the +1 frame at the A site. Translation is then continued in the +1 frame.

Ty2 and Ty4 contain an identical heptamer in roughly the same location resulting in production of the Gag-Pol fusion protein (Clare et al. 1988; Janetzky and Lehle 1992; Stucka et al. 1992). Ty3 has a similar *gag/pol* gene structure but utilises a GCG-A-GUU frameshift site (Farabaugh et al. 1993). This sequence is similar to the Ty1/Ty2/Ty4 sequence in that it features a pause-inducing codon corresponding to a rare tRNA in the 0 position of the A site. However the GCG codon at the P site is unlikely to allow peptidyl-tRNA slippage to the CGA codon in the +1 frame

(Farabaugh et al. 1993). Farabaugh et al. proposed a different mechanism, in which frameshifting occurs by out-of-frame binding of the incoming aminoacyl-tRNA, rather than slippage of the peptidyl tRNA. Hansen et al. (2003) and Baranov et al. (2004) present alternative explanations, predicting that out-of-frame A-site binding is unlikely to occur without a preceding event at the P-site.

While frameshifting may occur by different mechanisms, the heptamers share some common factors. In Ty3 and Ty1/2/4 a non-legal wobble base pairing occurs at the ribosomal P-site in the 0 frame. In Ty1/2/4 this occurs between the CUU bases and the unusually deconstrained $tRNA^{Leu}_{UAG}$. The cognate tRNA for the Ty3 GCG triplet, $tRNA^{Ala}_{CGC}$, is missing in *S. cerevisiae*, and is decoded by $tRNA^{Ala}_{UGC}$, a tRNA with a 5-carbomoylmethyluridine (ncm^5U) wobble base that recognises G- and A- ending codons (Johansson et al. 2008). Normally yeast tRNA families also include a tRNA with a C in the wobble position dedicated to recognising the G-ending codon, and Farabaugh et al. suggest that this may indicate that ncm^5U-pairing may be inefficient. A study by Sundarajan and colleagues, in which they replaced the P site 0-frame codon in the Ty3 frameshift heptamer with each of the 64 possible triplets, found that the codons decoded by near-cognate tRNAs were those that stimulated frameshifting (Sundararajan et al. 1999). Stahl et al. (2002) suggest that a near-cognate interaction in the P site may interfere with proper ribosomal function at the A site as the ribosome is in contact with the codon:anticodon complexes at both the A and P sites simultaneously (Yusupova et al. 2001; Stahl et al. 2002).

A second common factor is a translational pause induced by decoding of a triplet by a rare cognate tRNA at the ribosomal "A" site in the 0 frame. The codons AGG (Ty3), AGU (Ty1/2/4) and UGG have all been shown to be able to induce translational pausing (Pande et al. 1995). *S. cerevisiae* chromosomal frameshifting genes *OAZ1* and the frameshift candidate *URA6*, discussed below, both feature stop codons, which are also a common feature of frameshifting sites (Vimaladithan and Farabaugh 1994) and also induce a translational pause in their proposed frameshift heptamers. Recent studies indicate that interactions at the ribosomal "E" site plays a role in maintaining

the ribosomal reading frame (Marquez et al. 2004), and may play a synergistic role in stimulating +1 frameshifting (Liao et al. 2008).

Cis-acting elements can also play a stimulatory role in frameshifting. In isolation the Ty3 frameshifting heptamer is less efficient than that of Ty1. For frameshifting to occur at the maximal efficiency of 15%, a 14 nt stimulator sequence downstream of the heptamer, identified by mutagenesis, is required (Farabaugh et al. 1993). Without this stimulator, frameshifting at this heptamer occurs at only 2%, far less than the 40% efficiency in Ty1 (Farabaugh 2010)

In the L-A virus frameshifting between *orf1* and *orf2*, analogs of *gag* and *pol*, occurs in the -1 direction and employs a completely different mechanism (Icho and Wickner 1989; Dinman et al. 1991). Frameshifting occurs by backwards "slippage" at a G-GGU-UUA heptamer, where the first G is in the -1 frame. This is an example of what is called a "slippery heptamer" which are usually of the form X-XX.Y-YY.Z. Here XXX can be a run of any nucleotide (A/C/G/U), Y is usually either U or A, and Z is usually A, C or U (Jacks et al. 1988a). Frameshifting was proposed to occur when two tRNAs bound to XXY and YYZ at the P and A ribosomal sites respectively simultaneously slip in the -1 direction to the XXX and YYY codons in the -1 frame, although Baranov et al. propose a modified model, predicting that simultaneous slippage at both sites is unlikely (Baranov et al. 2004). Disruption of the runs of identical bases strongly reduce frameshifting efficiency in the L-A virus, while other changes such as UUU to AAA at the A site or GGG to AAA/UUU/CCC at the P site increase efficiency (Dinman et al. 1991). The E site may also play a role (Horsfield et al. 1995), and a downstream region forming a pseudoknot is necessary for efficient frameshifting (Dinman et al. 1991; Dinman and Wickner 1992).

The majority of -1 frameshift sites so far identified conform to this X-XX.Y-YY.Z structure (Farabaugh 2010), although there are exceptions: C_CA.A_AA.G_A and G_CG.A_AA.G are used in genes in some phages and insertion sequences (Mejlhede et al. 2004; Xu et al. 2004). In addition to downstream RNA structures such as

pseudoknots, -1 frameshifting is also stimulated by upstream Shine-Dalgarno-like sequences (Baranov et al. 2006).

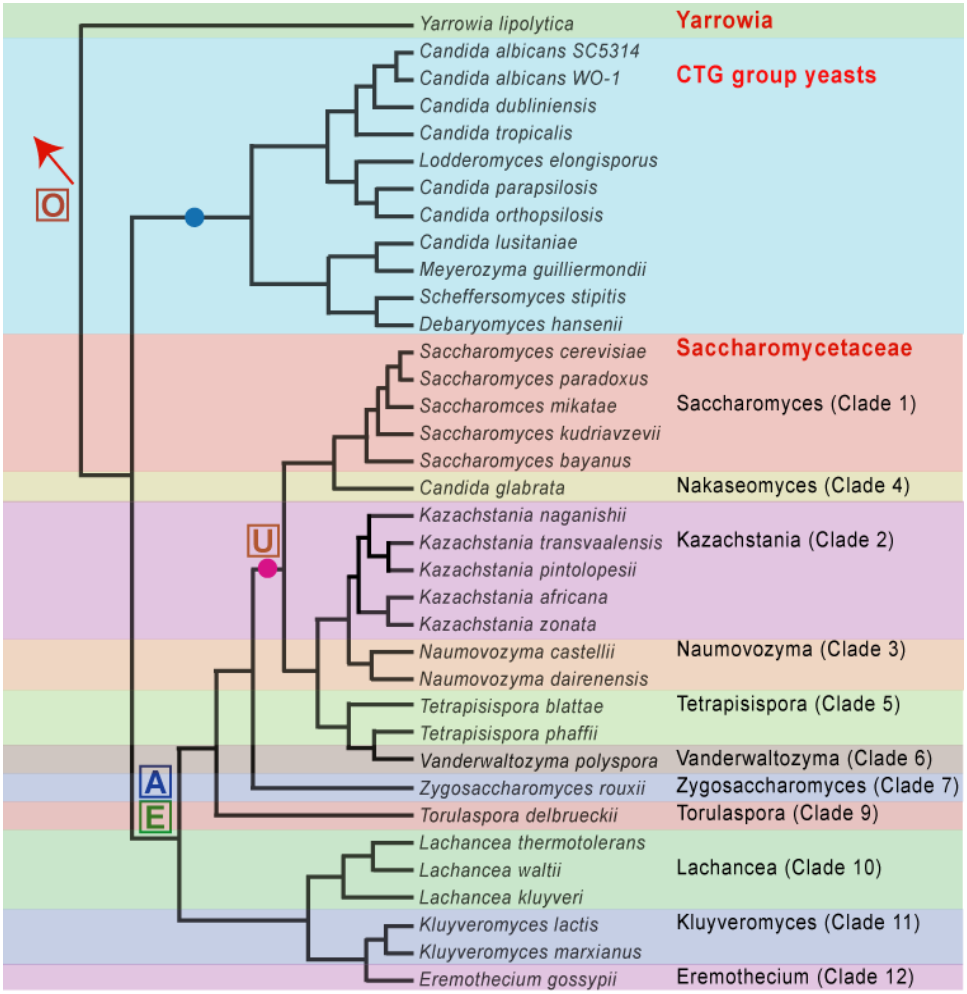

Figure 4.1 Phylogenetic tree of the Saccharomycetaceae, CTG group yeasts and *Yarrowia lipolytica* species used in this analysis, adapted from Kurtzman (2003) and Fitzpatrick et al. (2006). The branch along which the frameshift (or potential gene split in the case of *URA6*) appears is indicated for each of the four genes studied with the following symbols: E (*EST3*), O (*OAZ1)*, A (*ABP140*), U (*URA6*). The arrow beside the O indicates that the *OAZ1* frameshift is of an earlier origin. This tree is not drawn to scale.

During the annotation and analysis of new yeast genomes sequenced in our laboratory, we noticed interspecies variation in the sequences of some genes (*EST3*,

OAZ1, *ABP140*; Figure 4.1) known to undergo ribosomal frameshifting in *S. cerevisiae*. In our SearchDOGS analysis (Chapter 2) we also detected a small ORF conserved upstream of *URA6* in some species. This pair of ORFs appears to be either a previously unknown instance of ribosomal frameshifting, or a gene split in which a tight relationship has been maintained through co-expression of the ORF pair as a bicistronic mRNA. In this chapter I analyse the detailed structures of these genes and propose models for their evolution. My analysis of *EST3 and ABP140* expands on a previous study of the evolution of these genes by Farabaugh et al. (2006), and my analysis of *OAZ1* expands on studies by Ivanov et al. (2000a; 2006).

4.3 Results

4.3.1 *EST3* frameshifting is conserved in all Saccharomycetaceae clades except *Kluyveromyces*

The first identified example of a chromosomal gene in yeast featuring a programmed frameshift was *EST3*, a gene involved in the essential process of telomere maintenance. Telomere maintenance in eukaryotes is mostly carried out by the enzyme telomerase (Morris and Lundblad 1997), and is a highly regulated process (Smogorzewska and de Lange 2004). Telomere lengths are kept within certain length bounds by the preferential elongation of short telomeres by telomerase and the repression of telomerase at overelongated telomeres (Lee et al. 2010). In *S. cerevisiae*, the telomerase complex consists of the TLC1 telomerase RNA in association with three Est (ever shorter telomere) proteins. Initially identified by Lendvay et al. (1996), the role of the Est3 subunit in the complex has not been fully characterised. Experiments by Lee et al. (2010) indicate that Est3 has an essential regulatory function, although they do not rule out a contribution to enzymatic activity. Knocking out *EST3* results in a telomere shortening and senescence phenotype consistent with elimination of telomerase activity (Lendvay et al. 1996). Morris and Lundblad (1997) found that Est3 was encoded by two adjacent ORFs in different frames. Mutagenesis experiments showed that translation of *EST3* included sequence (translated in the +1 frame) upstream of the start codon of the downstream ORF. They pinpointed a CUU-A-GUU heptamer towards the end of the 0 frame upstream ORF at which frameshifting occurs; the introduction of silent mutations in this sequence completely eliminated Est3 function, whereas a deletion of a single base abolishing the need for a +1 frameshift resulted in full-length, fully functional Est3.

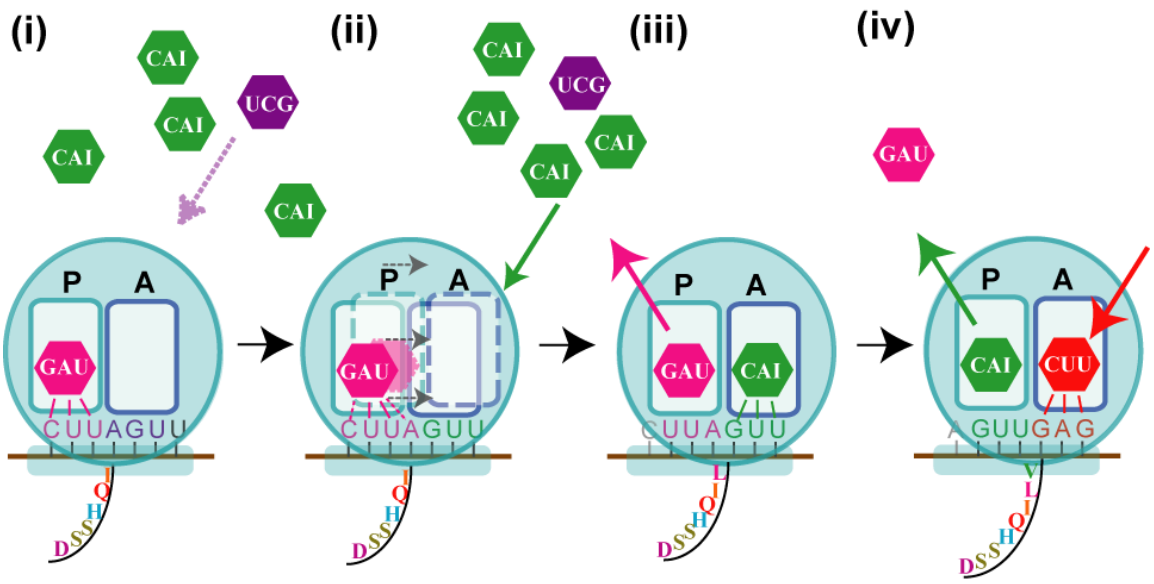

Figure 4.2 Cartoon illustrating the proposed method of +1 frameshifting at the CUU-A-GUU heptamer in *EST3* (Farabaugh et al. 2006). (i) The $tRNA^{Leu}_{UAG}$ (represented here as the pink "GAU" tRNA for illustrative purposes) loaded at the ribosomal "P" site has made a slightly weak "wobble" pairing to the CUU triplet. (ii) The low abundance of $tRNA^{Ser}_{CGU}$ (purple; UGC), the tRNA that recognises the AGU triplet results in a pause at this stage of translation. $tRNA^{Leu}_{IAC}$ (green; CAI), the cognate tRNA for the overlapping GUU codon in the +1 frame, is highly abundant and provides a competing interaction at the ribosomal "A" site. (iii) During the pause, $tRNA^{Leu}_{UAG}$ slips from the CUU triplet to the UUA triplet in the +1 frame. The next codon in the +1 frame, GUU, is translated by $tRNA^{Leu}_{IAC}$, resulting in the "A" base being skipped in the tranlation and a CUU-GUU readthrough occurring. (iv) The ribosome continues to translate in the +1 frame.

The CUU-A-GUU heptamer is highly conserved among the *EST3* genes of yeasts (Morris and Lundblad 1997; Farabaugh et al. 2006). It is also identical over the first 5 bases to the frameshifting heptamer used by the Ty1, Ty2 and Ty4 retrotransposons (Clare et al. 1988; Belcourt and Farabaugh 1990; Stucka et al. 1992), and identical over the last four bases to the Ty3 frameshift heptamer (Farabaugh et al. 1993). Several aspects of this sequence contribute to increase the frequency of frameshifting at this location, as illustrated in Figure 4.2 and described in more detail by Farabaugh et al., (2006). In *S. cerevisiae* the CUU triplet is primarily translated by the near-cognate $tRNA^{Leu}_{UAG}$ which is far more abundant than the cognate $tRNA^{Leu}_{GAG}$ (three gene copies of the former compared to one copy of the latter in the *S. cerevisiae* genome (Goffeau et al. 1996)). An unusual U_{33}->C_{33} substitution in the tRNA

structure may also place the cognate tRNA at a competitive disadvantage; U_{33} helps stabilise the anticodon loop and its absence leads to decreased translational efficiency (Farabaugh et al. 2006). The near-cognate $tRNA^{Leu}_{UAG}$ also has an unusual structure. In most tRNAs, post-transcription of the uridine base in position 34 restricts codon pairing to U:A or U:G. An unmodified uracil wobble base (Weissenbach et al. 1977; Randerath et al. 1979) allows $tRNA^{Leu}_{UAG}$ to recognise A, G or U quite well, and C more weakly, making it possible for it to make a slightly weakened pairing with the CUU triplet in the mRNA. As discussed in the introduction, weak codon:anticodon pairings appear to play a crucial role in frameshifting.

The cognate tRNA for the AGU triplet that follows CUU is of particularly low abundance. Thus, a pause is induced at the ribosomal "A" site, due to the absence of a readily available cognate tRNA. During this pause it is possible for the wobble-paired $tRNA^{Leu}_{UAG}$ to "slip" onto the UUA triplet in the +1 frame which also codes for leucine. At the same time, a competition is occurring at the ribosomal "A" site between the rare $tRNA^{Ser}_{GCU}$ which translates AGU and $tRNA^{Leu}_{IAC}$, the more abundant cognate tRNA for the overlapping GUU triplet in the +1 frame. The combination of these factors results in a high frequency of +1 frameshifting: $tRNA^{Leu}_{UAG}$ slips from CUU to UUA at the P site, and at the A site the tRNA translating AGU is outcompeted by the tRNA translating GUU. The ribosome then continues to translate in the +1 frame. In isolation the heptamer promotes frameshifting with a frequency of 8% (Vimaladithan and Farabaugh 1994). Frameshifting at this locus is further stimulated by a 27 nucleotide stimulator sequence immediately downstream which increases frameshifting approximately 8-fold (Taliaferro and Farabaugh 2007).

Building on a previous study of *EST3* frameshift sequence conservation in yeast (Farabaugh et al. 2006), we studied the sequences of *EST3* orthologs in 18 yeast species representing 11 of the 12 clades currently described for the Saccharomycetaceae family of yeasts (Figure 4.1). We found the CUU-A-GUU frameshift signal to be conserved perfectly in 16 of these 18 species (Figure 4.3). As

reported by Farabaugh et al., *Kluyveromyces lactis* has lost the requirement for a frameshift, and produces a full-length protein from a gene encoded in a single frame (Farabaugh et al. 2006). We found the same to be true for the closely related *Kluyveromyces marxianus*.

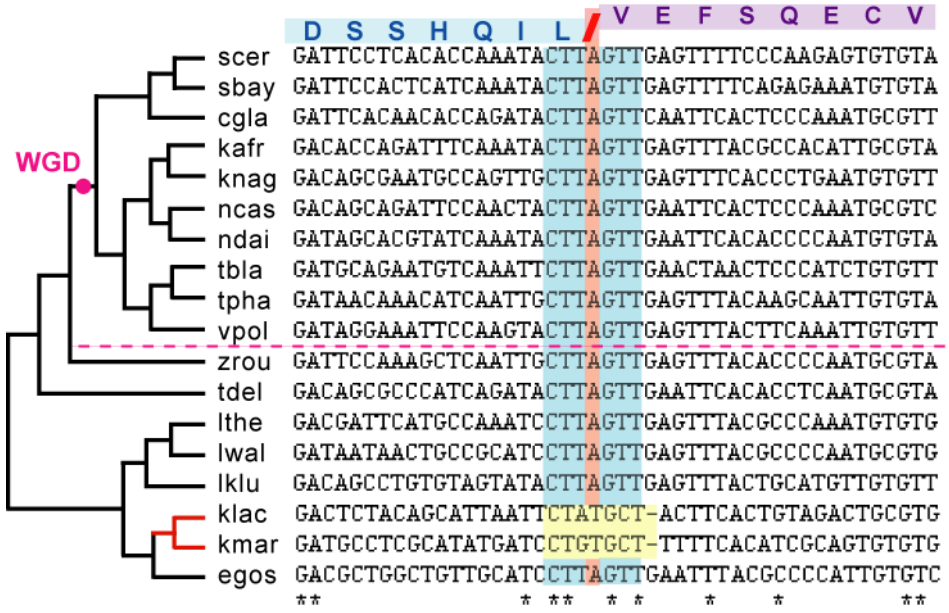

Figure 4.3 ClustalW nucleotide alignment (Larkin et al. 2007) of the region including and surrounding the location of the +1 ribosomal frameshift in *EST3*, which is conserved in all species except *K. lactis* and *K. marxianus*, the only species not to require a frameshift to produce a full-length protein. The species tree is adapted from Kurtzman (2003) and is not drawn to scale. The location of the WGD is indicated. The branch leading to the non-frameshift-containing species is highlighted in red. Post-WGD sequences are separated from non-WGD sequences by the dashed pink line. The protein sequence corresponding to the *S. cerevisiae* ortholog is shown above the nucleotide sequence, with the frameshift indicated by the red dash. Species acronyms used are as follows: scer (*S. cerevisiae*), sbay (*S. bayanus*), cgla (*C. glabrata*) kafr (*K. africana*), knag (*K. naganishii*), ncas (*N. castellii*) ndai (*N. dairenensis*), tbla (*T. blattae*), tpha (*T. phaffii*), vpol (*V. polyspora*), zrou (*Z. rouxii*) tdel (*T. delbrueckii*), lthe (*L. thermotolerans*), lwal (*L. waltii*), lklu (L. *kluyveri*) klac (*K. lactis*), kmar (*K. marxianus*) egos (*E. gossypii*).

No evidence exists for a frameshifting requirement in *EST3* in species studied from the CTG group of yeasts (Figure 4.1). *S. cerevisiae EST3* aligns over its full length with the CTG group homologs studied, all of which are coded in a single frame.

Farabaugh and colleagues suggested that frameshifting may have originated at the divergence of *K. lactis* and *E. gossypii*, or alternatively may have a much deeper origin and was independently lost in *K. lactis*, and in a separate event in the CTG group yeasts *D. hansenii* and *C. albicans*. Our increased sampling of species allows us to reject the first hypothesis. The *Kluyveromyces* and *Eremothecium* clades are closely related and form a monophyletic group relative to the post WGD clade, and all other clades studied within the Saccharomycetaceae have a frameshift-containing *EST3*, indicating that the frameshift is ancestral to the Saccharomycetaceae, and has most likely been lost in a single event either on the branch leading to *K. lactis* and *K. marxianus*, or in the *Kluyveromyces* genus as a whole. The absence of a frameshift in any of the full-length *EST3* orthologs present in 11 CTG group species indicates that either the frameshift has been lost in the ancestor of these species, or that the frameshift originated in the ancestor of the Saccharomycetaceae. We were unable to identify any *EST3* homologs in *Yarrowia lipolytica* or more distantly related fungal species; whether this is due to *EST3* being a yeast-specific protein or due to a somewhat high level of divergence at this locus (Farabaugh et al. 2006) (Figure 4.4) is unclear. Thus, from the data available we hypothesise that the frameshift originated in the ancestor of the Saccharomycetaceae approximately 100-200 million years ago (Figure 4.1) (Taylor and Berbee 2006), although an earlier origin with multiple losses of frameshifting cannot be ruled out.

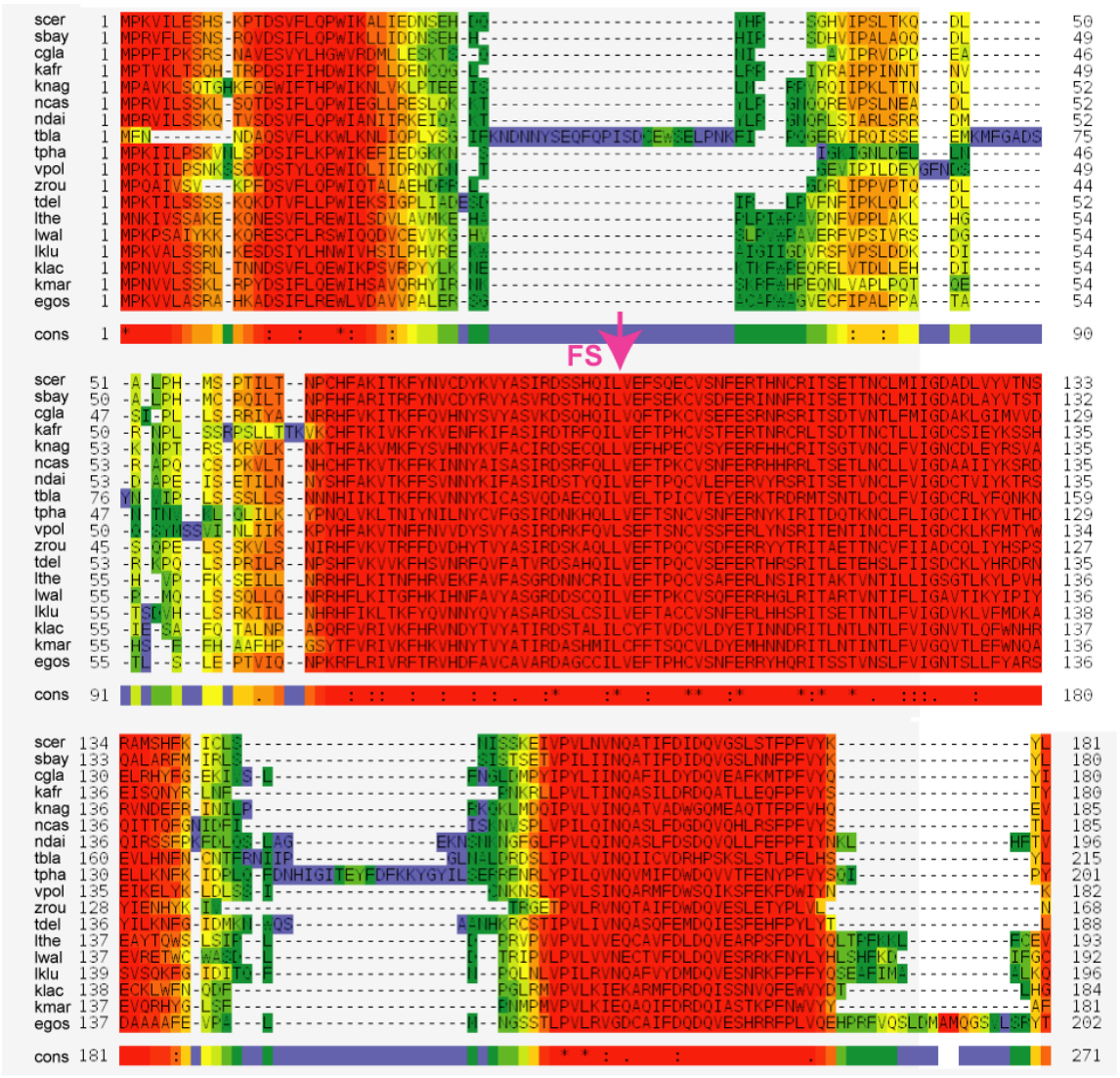

Figure 4.4 Protein alignment (Notredame et al. 2000) of Est3 orthologs in 17 yeast species. The position at which the +1 frameshift occurs in the majority of species is indicated by the pink arrow.

4.3.2 Variation in the frameshift site at the *OAZ1* locus

The most widely studied example of frameshifting in a eukaryotic gene is that of the protein antizyme, which is translated via a programmed frameshift in species ranging from yeasts to human and is encoded by *OAZ1* in *S. cerevisiae*. Antizyme functions to lower cellular levels of polyamines, small organic cations that play a role in many fundamental cellular processes such as protein synthesis, cell division, programmed

cell death and binding and stabilising DNA and RNA (Childs et al. 2003; Wallace et al. 2003). Antizyme is an inhibitor of ornithine decarboxylase, the enzyme that catalyses the rate-limiting step in polyamine biosynthesis (Murakami et al. 1992; Zhang et al. 2003a), and also inhibits the import of polyamines into the cell (Sakata et al. 2000; Belting et al. 2003).

Several feedback mechanisms operate to regulate polyamine levels tightly. Firstly, antizyme inhibits the ubiquitination of, and thus stabilises, an antizyme inhibitor (Bercovich and Kahana 2004). Secondly, synthesis of the full length antizyme protein requires a ribosomal frameshifting event in all species from yeast to mammals. Free polyamine in the cell has been shown to stimulate +1 translational frameshifting in the antizyme gene (Coffino 2001). The effects of the various polyamines regulated by *S. cerevisiae* antizyme (spermidine, spermine and putrescine) on ribosomal function is analysed in detail in Rato et al. (2011), although a recent study reports that it is the nascent antizyme polypeptide itself that operates *in cis* to negatively regulate frameshifting, serving as a sensor for polyamine levels in the cell (Kurian et al, 2011).

Three antizyme paralogs have been identified in mammals, all requiring a frameshift (Ivanov et al. 2000a). In yeasts, a single antizyme gene has been identified in species including the fission yeasts *Schizosaccharomyces pombe*, *S. octosporus*, *S, japonicus*, a number of the Saccharomycetaceae family of species, and *Yarrowia lipolytica* (Ivanov et al. 2006). In all cases so far identified, translation begins at the start of a shorter ORF (ORF1) in the 0 frame and switches to a longer, slightly overlapping ORF (ORF2) in the +1 frame that lacks the ability to initiate independently (Matsufuji et al. 1995). The +1 frameshift "jump" occurs after the last sense codon in ORF1 (UCC in all vertebrates as well as *Yarrowia lipolytica* (Ivanov et al. 2000b), GCG in *S. cerevisiae* (Palanimurugan et al. 2004)), resulting in continuation of translation in the +1 frame. The UGA stop codon in ORF1 is universally conserved (Ivanov et al. 2000a); frameshifting results in the U being skipped and the translation

of GAX, where X represents the next base in the nucleotide sequence after UGA (Figure 4.5).

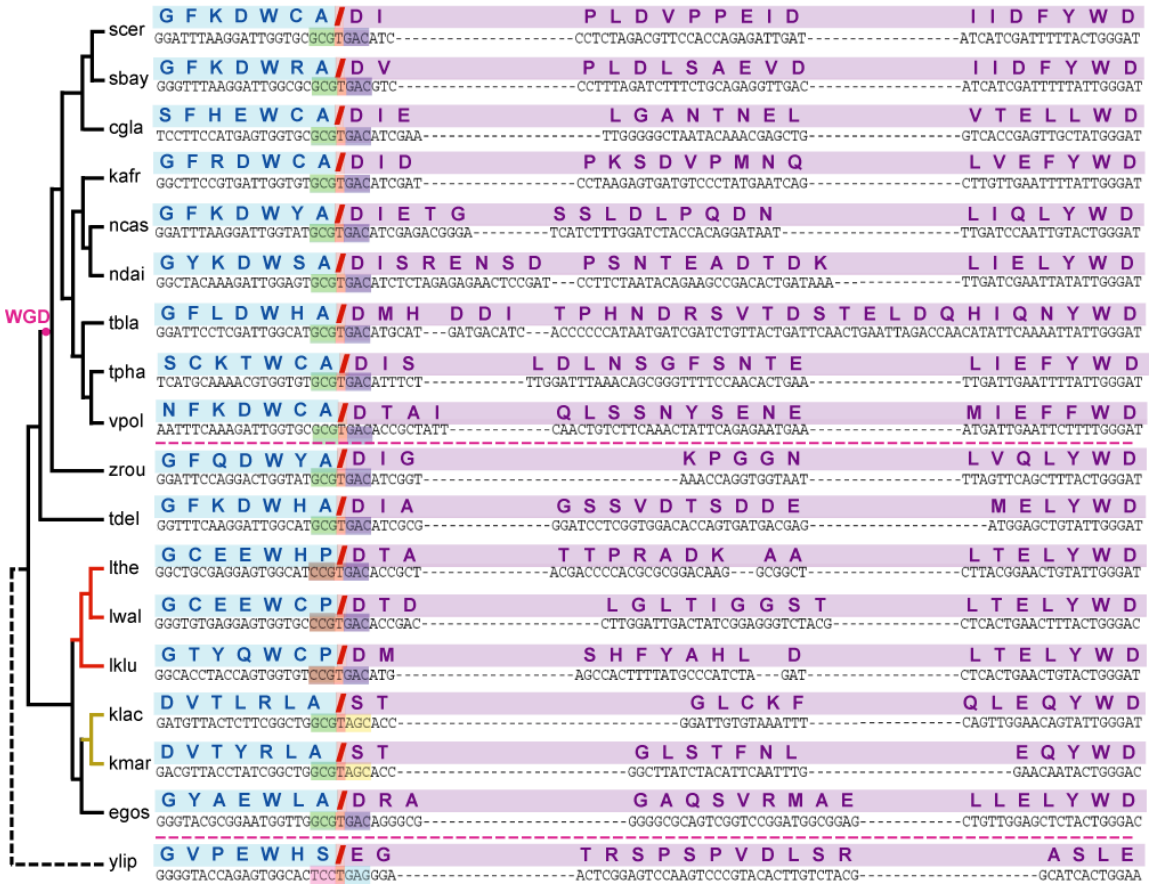

Figure 4.5 ClustalW nucleotide alignment (Larkin et al. 2007) of the region surrounding the +1 ribosomal frameshift in *OAZ1* in 18 yeast species. Post-WGD sequences are separated from non-WGD sequences by the dashed pink line. The branches leading to the orthologs with alternate frameshift heptamers are highlighted. The frameshift sequence of the *Yarrowia lipolytica* antizyme gene is included as an outgroup (Ivanov et al. 2006), separated by the others by the dashed red line. The protein sequences for each species are included above the nucleotide sequences, with the frameshift indicated by the red dash.

In addition to the frameshift heptamer, several cis-acting RNA sequences located both upstream and downstream of the frameshift are necessary for efficient frameshifting in vertebrates (Matsufuji et al. 1995; Ivanov et al. 1998a; Ivanov et al. 2000b; Howard et al. 2001; Petros et al. 2005). The 50 nt region directly upstream of the frameshift in vertebrate antizyme 1 contains a stimulatory region consisting of

three distinct modules that evolved separately (Ivanov et al. 2000a), and downstream of the frameshift there is a stimulatory pseudoknot. The rate of frameshifting for antizyme 1 has been measured at 6-30%. In the absence of these elements, frameshift efficiency is reduced by over an order of magnitude (Ivanov et al. 2006). In invertebrates, cis-acting sequences are to date less well-characterised. *Schizosaccharomyces pombe* contains a poorly characterised 3' signal extending for up to 150 nucleotides from the point of frameshift (Ivanov et al. 2000b), and phylogenetic comparisons suggest that other 3' signals exist in certain branches of fungi and nematodes (Ivanov et al. 2004). One unusual feature of the *Saccharomyces* group species is that both computational RNA-folding techniques and nucleotide sequence conservation show little evidence to suggest the existence of frameshift-stimulating cis-acting sequences (Ivanov et al. 2006). Ivanov and colleagues identified only 8 nucleotides perfectly conserved in 11 *Saccharomyces* group species and *Y. lipolytica*, four of which are contained within the frameshift heptamer, and an additional 12 nucleotides conserved in 10 of 12 species. This suggests that distinct features of the *Saccharomyces* group frameshift sequence may make it efficient enough to remove the requirement for extensive additional frameshift-stimulating sequences.

Some properties of the GCG-U-GAC heptamer illustrate some of the frameshift-stimulating principles described for *EST3* and in the Introduction, but with some distinctions. At the ribosomal P site, the cognate tRNA for GCG, $tRNA^{Ala}_{CGC}$ is absent in *S. cerevisiae* (Ivanov et al. 2006). GCG/GCA triplets are decoded by $tRNA^{Ala}_{UGC}$, which has a 5-carboamoylmethyluridine (ncm^5U) wobble base that recognises both A- and G- ending codons (Johansson et al. 2008). Farabaugh notes that yeast tRNA families usually include a tRNA with a C in the wobble position to recognise the G-ending codon, and suggests this means the ncm^5U to G interaction may be inefficient (Farabaugh 2010).

This may lead to competition between a weak interaction with a near-cognate tRNA in the 0 frame and a stronger interaction between the CGU triplet and its more

abundant cognate tRNA in the +1 frame. While this potentially weak, near-cognate interaction is occurring at the ribosomal P site, a pause is induced at the A site by the UGA stop codon. Competition to decode in the +1 frame is fierce at this position; 16 genes coding for the cognate $tRNA^{Ser}_{GUC}$ exist in *S. cerevisiae*. The bases immediately downstream of the heptamer may also contribute to disrupting normal ribosomal operation. The stop codon UGA followed by C is regarded to be a very weak termination signal (Brown et al. 1990; Tate et al. 1999), and experiments have shown that a significant level of translational readthrough can be induced in *S. cerevisiae* by the sequence CA(A/G) 3' of a stop codon (Namy et al. 2001); in this case the UGA stop codon is followed by the bases CAU. The mechanism by which the frameshift occurs during the UGA-induced pause is unclear. Unlike *EST3*, +1 slippage of the tRNA-mRNA interaction at the P-site appears unlikely. It may be that dissociation and repairing occurs at the P site in the +1 frame (Ivanov et al. 2006), or that occlusion of the mRNA base 3' of the zero frame P-site codon leads to +1 frameshifting (Stahl et al. 2001; Hansen et al. 2003; Baranov et al. 2004).

We found that the requirement for a +1 frameshift was maintained in all 18 species studied (Figure 4.5). *The K. naganishii* ortholog was excluded due to insufficient sequence data. However, while the GCG-U-GAC sequence is conserved in all the post-WGD species studied as well as *Z. rouxii, T. delbrueckii* and *E. gossypii*, we confirmed that the branch leading to the genus *Lachancea* features a slightly different CCG-U-GAC frameshift signal as reported by Ivanov et al. (2006). We identified a third heptamer, GCG-U-AGC, that acts as a frameshift signal in both *K. lactis* and *K. marxianus* (Figure 4.5) and involves a different stop codon (UAG) instead of UGA. The distribution of the GCG-U-GAC heptamer-containing orthologs indicates that this represents the ancestral state within this species, with a G->C transversion occurring in the *Lachancea* branch and a separate change of sequence occurring in the *Kluyveromyces* species studied. However, the frameshift sequence in the ancestor of all known *OAZ1*-containing species is likely to have been UCC-U-GAX (where X is variable), the sequence found in *Yarrowia lipolytica* and vertebrates (Ivanov et al. 2000a). Frameshifting sequences could not be compared in the CTG group yeasts

studied, as surprisingly they were found to lack identifiable *OAZ1* homologs. Like *EST3*, *OAZ1* shows quite high divergence over its entire length (Figure 4.6) and it is interesting to note that among these species the most divergent species are the two *Kluyveromyces* species containing the GCG-U-AGC heptamer.

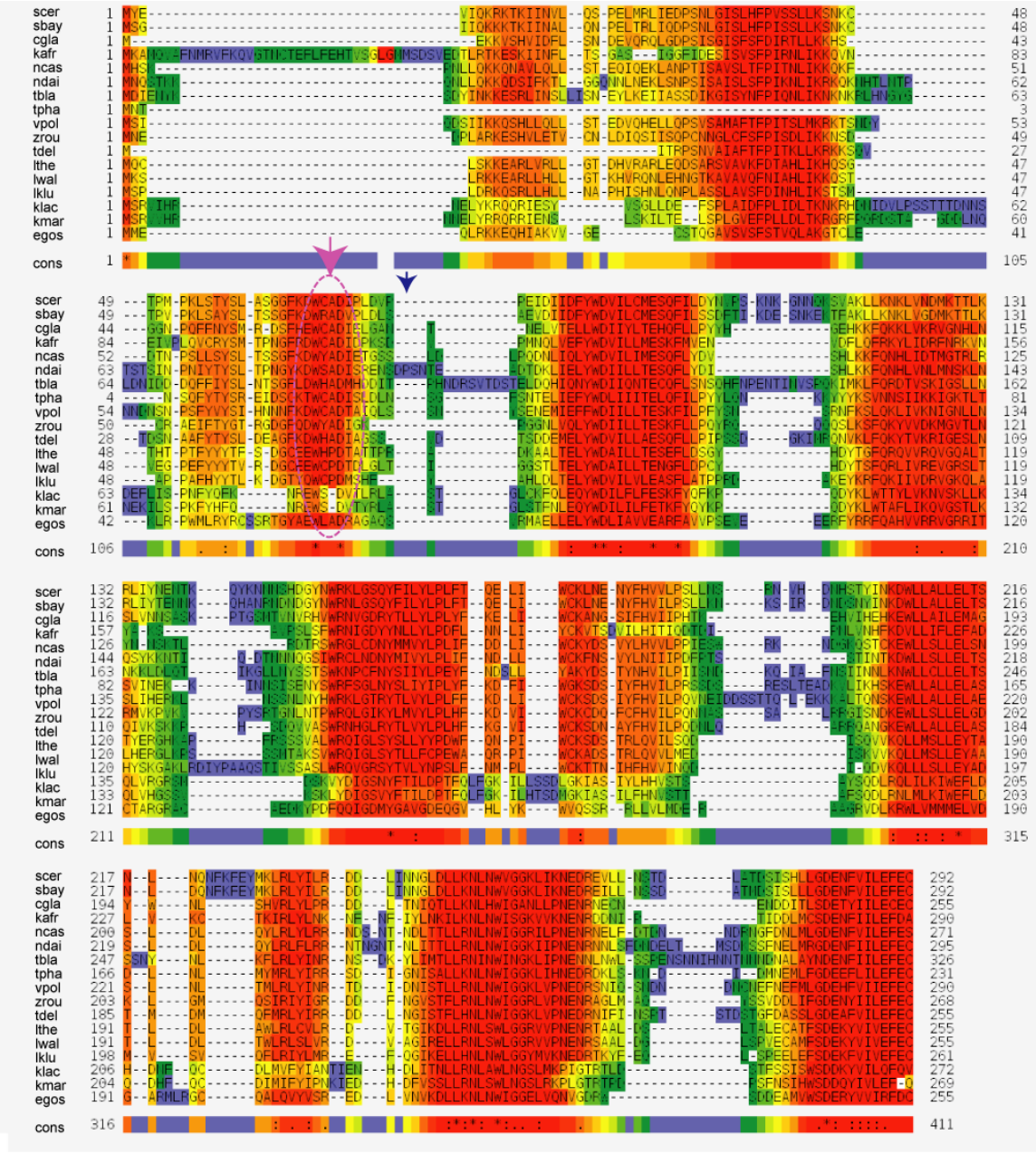

Figure 4.6 Protein alignment (Notredame et al. 2000) of Oaz1 orthologs in 16 yeast species. The position at which the +1 frameshift occurs in the majority of species is indicated by the pink arrow and

circled; the adjacent navy arrow denotes the frameshift in the *Kluyveromyces* species. The *K. naganishii* ortholog is excluded due to insufficient sequence data.

A cognate tRNA for the GCG codon at the 5' end of the heptamer is entirely absent in the post-WGD species; this is also the case for *Z. rouxii, T. delbrueckii* and *K. lactis*, all of which use a heptamer beginning with GCG. Ivanov and colleagues note that *L. kluyveri* and *L. waltii* are missing the cognate tRNAs for the 5' CCG triplet in the heptamers in these species (Ivanov et al. 2006); we found that this is also the case in the third "CCG"-containing species, *L. thermotolerans*. This fits in with a model where frameshifting is encouraged by the absence of a readily available cognate tRNA to interact with this codon in the 0 frame (Sundararajan et al. 1999; Ivanov et al. 2006). The only species to contain the cognate tRNA for the 5' triplet in the heptamer is *E. gossypii*, which contains two $tRNA^{Ala}_{CGC}$ genes. The downstream ORF in *E. gossypii* has an unusually high GC content of 66% (genome average: 52% (Dietrich et al. 2004)). The *S. cerevisiae* ORF2 has 32% GC content (Ivanov et al. 2006) (genome average: 38% (Goffeau et al. 1996)), and the average GC content of ORF2 in all the Saccharomycetaceae genera we studied with the exclusion of *E. gossypii* is 36%. Ivanov and colleagues propose that the increased GC content in *E. gossypii* may result in frameshift-stimulating RNA secondary structures downstream of the heptamer that could compensate for an absence of frameshift induction by the GCG triplet in this species (Ivanov et al. 2006). It is interesting to note that in the closely related *K. lactis* ortholog, GC content in ORF2 is 38% which is right on the genome average of 38% (Dujon et al. 2004), and is missing the cognate tRNA for the 5' GCG codon.

Thus, the higher-than-average GC content is unique to *E. gossypii*, and supports a model where compensation provided by RNA structures negates the need to switch to a more strongly frameshift-inducing triplet such as CCG in *E. gossypii*. The GCG-U-AGC heptamer in the *Kluyveromyces* species is likely to promote frameshifting similarly to the more common GCG-U-GAC. While the cognate tRNA for the A-site +1 frame AGC triplet in *K. lactis* is less abundant than the cognate for GAC in *S.*

cerevisiae (two copies versus 16), the sequence still contains a pause-inducing UAG stop codon in the 0 frame at the A-site.

It is worth noting that while antizyme frameshifting is exclusively +1 in all species studied to date, a sequence stimulating +1 frameshifting in one context may stimulate minus direction frameshifting in another. Matsufuji et al. (1996) expressed a cassette containing the mammalian antizyme 1 frameshift sequence as well as cis-acting sequences in *S. cerevisiae* and found that a high level (16%) of -2 frameshifting only occurred at the frameshift site. This was confirmed by amino acid sequencing that identified an extra proline corresponding to the additional CCU codon at the frameshift site that was the result of the backwards 'slip'. The cis-acting sequences also had a radically different influence in *S. cerevisiae*. The 5' region which stimulates frameshifting threefold in its native environment (Ivanov et al. 1998b) had no stimulating effect, and the stimulating effect of the 3' pseudoknot was increased from 2.5-fold to 30 fold. In contrast, the insertion of a similar cassette into *S. pombe* resulted in frameshifting that was 80% +1 and 20% -2 (Ivanov et al. 1998b), indicating that the *S. pombe* cellular machinery interacts with the cassette in a manner much more similar to the way mammalian machinery interacts with these frameshifting sequences than the way budding yeast does. These results highlight that the genomic (local and global) and cellular contexts of frameshifting elements need to be studied when trying to ascertain the nature and quantify the extent of the effects they may have in an organism.

4.3.3 Frameshifting and ohnolog retention at the *ABP140* locus

A particularly interesting example of +1 ribosomal frameshifting in yeast is that of *ABP140*. The Abp140 protein was first purified by Asakura et al. (1998) in an experiment to identify a gene for an actin filament (F-actin)-binding protein that appeared to be missing from the *S. cerevisiae* genome. The sequence of this protein was found to correspond to a fusion of two adjacent ORFs, *YOR239W* and

YOR240W, that required a +1 frameshift. Asakura et al. identified a CUU-A-GGC heptamer towards the 3' end of *YOR239W* that perfectly matches the Ty1 frameshift site, and results in a protein translation perfectly matching the Abp140 protein sequence.

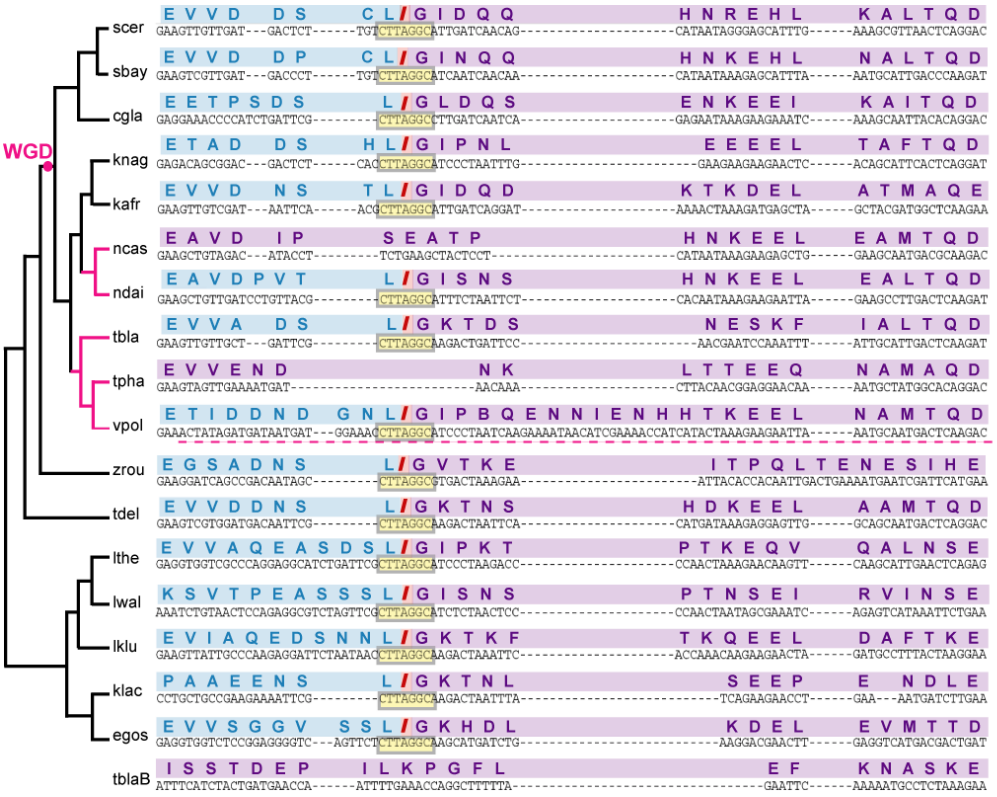

Figure 4.7 ClustalW nucleotide alignment (Larkin et al. 2007) of the region surrounding the location of the +1 ribosomal frameshift in 18 full-length *ABP140* homologs, including two sequences from *T. blattae*. Frameshifting occurs at the CUUAGGC heptamer highlighted (blue, red, yellow) resulting in the ribosome "skipping" the middle A to produce a full-length protein sequence. The heptamer is perfectly conserved in all species containing the frameshift; however in *N. castellii* and *T. phaffii* (tree branches highlighted in red) the frameshift has been lost. The tree branches highlighted in pink indicate post-WGD species in which a potential second ohnolog lacking the frameshift has been identified (Figure 4.10). The full-length *T. blattae* second ohnolog (tblaB) is also included. The location of the WGD event is highlighted in the species tree. This tree is not drawn to scale.

We have identified a frameshift-containing *ABP140* homolog in 15 of the 17 yeast species studied. The requirement for a frameshift has been lost in two separate events in *N. castellii* and *T. phaffii* (Figure 4.7), but a full-length gene showing homology to the frameshift-containing homologs exists in these species (Figure 4.8). The

frameshift CUU-A-GGC heptamer is conserved perfectly in the 15 species with a frameshift-containing *ABP140* gene. The *ABP140* region extending from the location of the frameshift to the 3' end of the gene (from here on referred to as ORF2) shows very high protein sequence conservation both in the homologs containing and lacking the frameshift (Figure 4.9). However, the region upstream of the frameshift (from here on referred to as ORF1) shows far more divergence both in length and in sequence, and contains runs of repetitive sequence. The exception to this is a highly-conserved 15 amino-acid region at the 5' end of the gene in all species (Farabaugh et al. 2006). Interestingly, we identified a second potential ohnolog in five post-whole genome duplication yeast species, only one of which is approximately equal in length to full-length *ABP140*. These paralogs were not identified by Farabaugh et al. An open reading frame encoding a full-length homolog lacking a frameshift exists at the location corresponding to *ABP140* on the *T. blattae* sister chromosome. This ORF, *Tbla_B*, shows sequence homology with its sister ohnolog, *Tbla_A*, and the other *ABP140* homologs but is considerably more diverged over its entire length than the other full-length homologs (Figure 4.9).

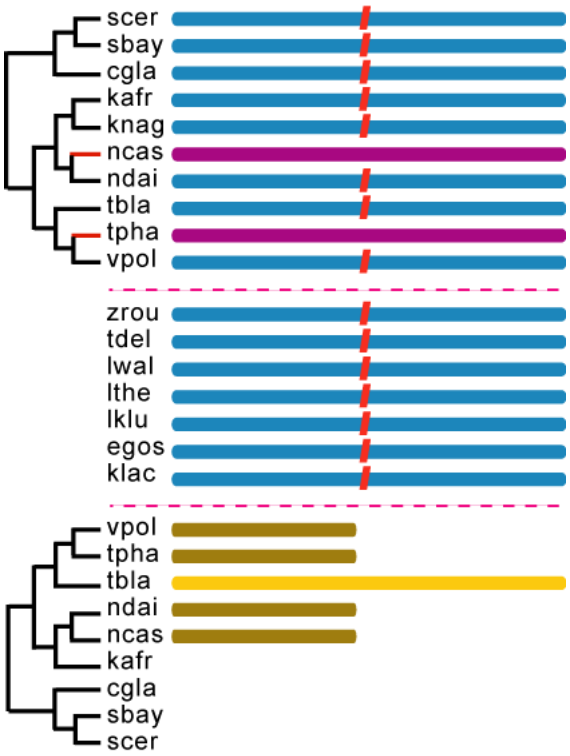

Figure 4.8 Cartoon illustrating the distribution of frameshift-containing homologs (frameshift indicated by a red dash) at the *ABP140* locus. Post-WGD species are represented with double tracks above and below the non-WGD species, represented by single tracks in the centre. In four of the post-WGD species a second potential ohnolog aligning to the 5' region of *ABP140* and extending as far as the location of the frameshift appears to have been retained following WGD. *T. blattae* alone contains both a full-length *ABP140* ohnolog containing a frameshift and a second full-length potential ohnolog where the frameshift has been lost. The frameshift appears to have been lost in the full length *N. castellii* and *T. phaffii ABP140* ohnologs in two separate events.

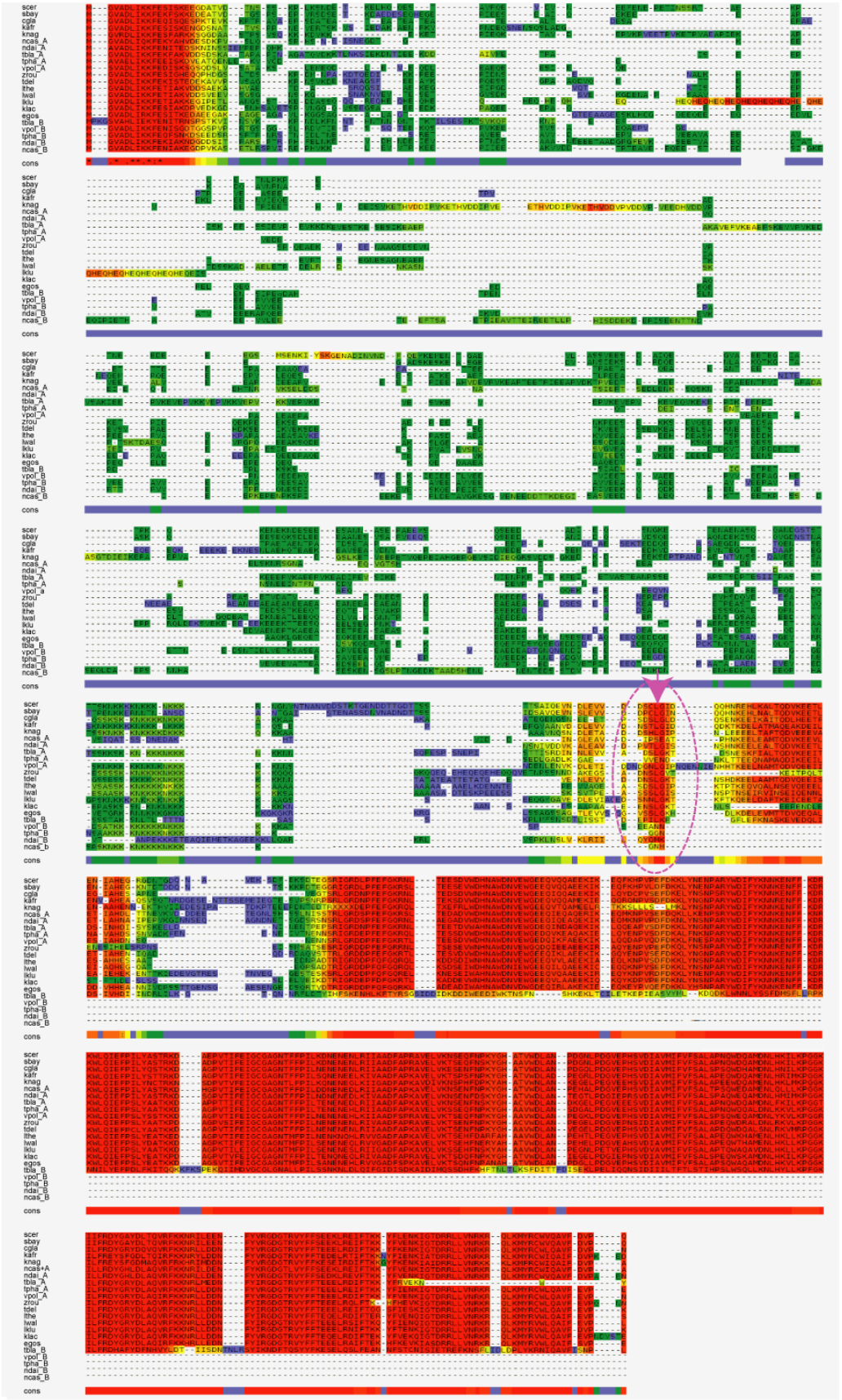

Figure 4.9 TCOFFEE alignment (Notredame et al. 2000) of the protein sequences of *ABP140* homologs in 17 yeast species. Species acronyms are as for Figure 4.3. *T. blattae* contains a second full-length *ABP140* ohnolog (Tbla_B) lacking a frameshift requirement, and *V. polyspora*, *T. phaffii*, *N. dairenensis* and *N. castellii* each contain a truncated second ohnolog aligning to the 5' ("ORF1") region of full-length *ABP140*. The position at which the frameshift occurs is indicated by the pink arrow and circle.

Four other species (*N. castellii*, *N. dairenensis*, *T. phaffii* and *V. polyspora*) contain open reading frames that aligns to ORF1, but do not appear to contain any sequence aligning to ORF2 (Figure 4.8; Figure 4.10). All five of these "B" copies show evidence of protein conservation, producing pairwise omega (*Ka/Ks*) values that almost all range between 0.05 and 0.4 (with none above 0.7) when compared to each other and their full-length counterparts. No evidence was found for the existence of B copies in the remaining post-duplicated species, or for the existence of the 3" end of the *ABP40* as a separate ORF or in a different frame in the species containing B copies.

Figure 4.10 Amino acid alignment of the truncated "B copy" ohnologs of Abp140 that appear to have been retained after whole genome duplication in *N. castellii*, *N. dairenensis*, *T. phaffii* and *V. polyspora* (Notredame et al. 2000). A full length ohnolog that has lost the frameshift has been retained in *T. blattae* in addition to the frameshift-containing ohnolog. Only the ORF1 region up to a position corresponding to the location of the frameshift in *S. cerevisiae* is shown in *S. cerevisiae* and *T. blattae*; the red arrows indicate that these genes extend in the 3' direction. The short retained orthologs align to the more divergent 5' half of Abp140 upstream of the frameshift (position indicated by the blue arrow). The highly conserved 15 amino acid N-terminal region is highlighted in blue.

Thus it appears that in some species, following WGD the ORF2 region of one of the copies of *ABP140* was lost while retaining the ORF1 region. Given that the remaining ORF1 segments align to the full length gene up to the point of frameshift, it appears

that the frameshift is likely to have played a crucial role in this gene degradation. One possible explanation is that the frameshift signal became lost in these species without the frame being corrected, resulting in abolition of translation of the ORF2 region, and eventual loss of transcription and sequence degradation of the ORF2 region. A mutation leading to the correction of the frame of the B copy in *T. blattae* would have resulted in the retention of a full-length ortholog.

The pattern of loss of the ORF2 region and retention of the ORF1 region appears to have been reversed in the CTG group of yeast species, where *ABP140* orthologs exist but lack the frameshift signal. Interestingly, these orthologs are on average only 55% of the length of the full-length orthologs present in the family Saccharomycetaceae, and align only with the ORF2 region of *S. cerevisiae ABP140* downstream of the frameshift signal (Figure 4.11). Evidence suggests that the ORF1 region of the gene has been lost in these species; in every CTG clade species studied except *C. lusitaniae* and *C. tropicalis* the next upstream ORF is a highly conserved homolog of *S. cerevisiae GAL7*.

Several features of *ABP140* frameshifting remain in need of characterisation. Firstly, the frequency with which frameshifting occurs in this gene has not been measured, and it should be determined whether ribosomes that fail to successfully frameshift produce proteins with a biologically relevant function. Secondly, characterisation of any nearby downstream (or upstream) regions that enhance the frequency of frameshifting must be a priority. Without experimental verification, the possibility that the B copies are pseudogene remnants of WGD ohnologs retaining some of the hallmarks of genuine protein-coding genes also cannot be ruled out. Provided the truncated ORFs that constitute potential ohnologs are transcribed and translated genes, a question is raised regarding the biological significance of retaining only the ORF1 region of *ABP140*, particularly given that it represents what is on the whole the far less well-conserved region of the full-length gene; the ORF2 region contains an S-adenosylmethionine (SAM) domain that is highly conserved across eukaryotes (D'Silva et al. 2011). Is this ORF capable of encoding a functional protein by itself? It

may be that dosage issues come into play. The percent frequency of frameshifting varies from gene to gene and is always less than 100% efficient. The Ty1 retrotransposon takes advantage of the frequency of frameshifting to maintain a 50:1 ratio of 5'-only Gag to full-length Gag-Pol in the virus-like particle (Dinman and Wickner 1992). Reducing or increasing frameshifting at this locus and thus changing the Gag:Gag-Pol ratio results in severely reduced frequency of transposition (Xu and Boeke 1990; Kawakami et al. 1993). It may be significant that *N. castellii* and *T. phaffii*, the two species with no frameshift-containing *ABP140* orthologs whatsoever, are two of the four species to have retained a second, ORF1 region-only, "B" copy.


C. albicans SC5314
C. albicans WO1
C. dubliniensis
C. tropicalis
C. parapsilosis
C. orthopsilosis
L.elongisporous
D. hansenii
Sch. stipitis
M.guilliermondii
C.lusitaniae
S.cerevisiae
E. gossypii
Z. rouxii
Cons

Figure 4.11 Amino acid alignment of Abp140 homologs from the CTG group yeasts (*C. albicans* strains SC5314 and WO1, *C. dubliniensis, C. tropicalis, C. parapsilosis, C. orthopsilosis, L. elongisporus, D. hansenii, S. stipitis, M. guilliermondii, C. lusitaniae*) and full length *S. cerevisiae, E. gossypii* and *Z. rouxii* Abp140 homologs (Notredame et al. 2000). The position of the frameshift in *S. cerevisiae, E. gossypii* and *Z. rouxii* is indicated with arrows coloured blue, green and yellow respectively.

Despite its conservation in all the Saccharomycetacea and CTG group yeasts studied, *ABP140* has been found to be inessential; knocking out the gene has no effect on cell growth or F-actin organisation (Asakura et al. 1998; Niewmierzycka and Clarke 1999). A recent paper by Noma and colleagues showed that Abp140 functions in tRNA modification, introducing a 3-methylcytidine (m(3)C) modification at position 32 of the tRNAs for threonine and serine (Noma et al. 2011). While these are not tRNAs that correspond to codons within the frameshift site, it may be significant that (like *OAZ1*) *ABP140* is a frameshift-containing gene that produces a protein that plays a role in translation.

4.3.4 Unusual gene evolution at the *URA6* locus

Our analysis of possible ribosomal frameshifting sites in yeast genome sequences also uncovered one locus, *URA6,* at which an unusual evolutionary process appears to have occurred involving either the introduction of a frameshift, a gene split, or both. Pyrimidines, forming the backbone of the cytosine (C), thymine (T) DNA nucleotides and uracil (U) RNA nucleotide, are essential building blocks for DNA and RNA synthesis. *URA6* codes for a uridylate kinase which catalyses the conversion of uridine monophosphate (UMP) into uridine-5'-diphosphate (UDP), the seventh step in the *de novo* biosynthesis of pyrimidines (Jong et al. 1993). It is a homolog of the mammalian (uridine monophosphate kinase-coding gene) *CMPK1*. Other than at the N-terminal end, homologs show very high sequence conservation across eukarya (Figure 4.12).

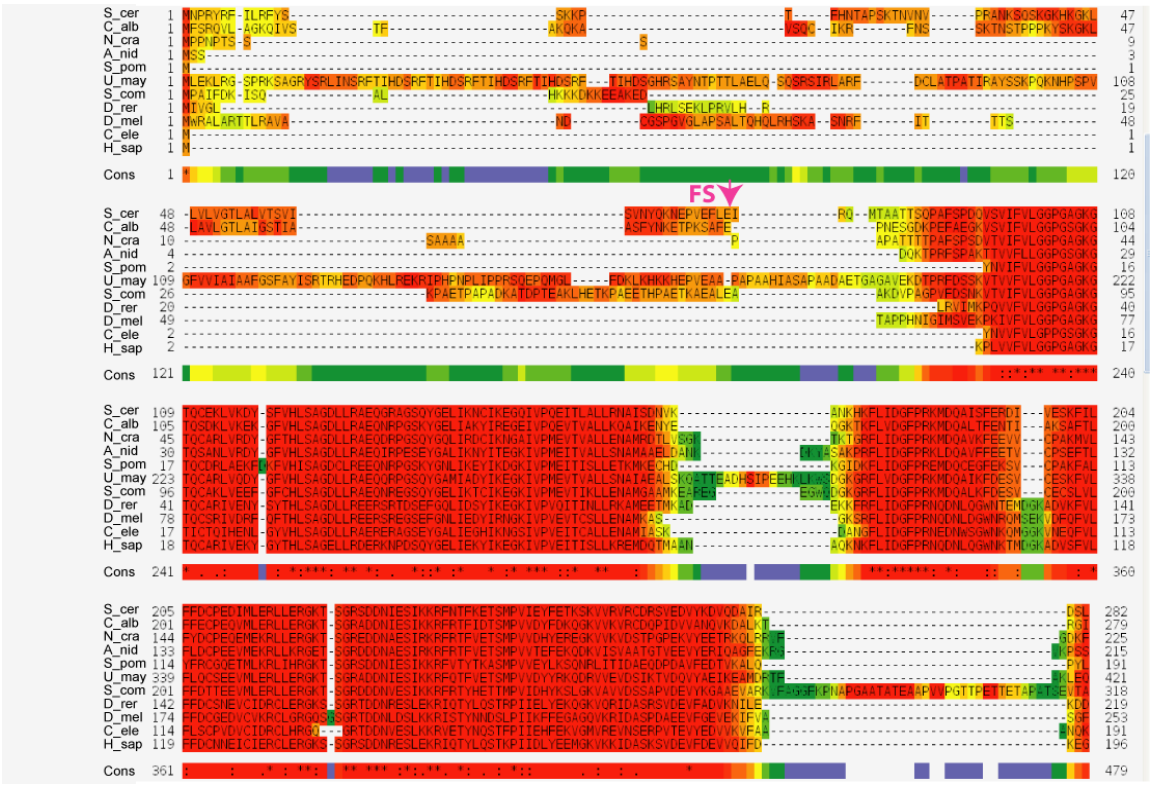

Figure 4.12 Amino acid alignment (Notredame et al. 2000) of Ura6 homologs across a selection of eukaryotes. Species acronyms are as follows: *S. cerevisiae* (S_cer), *C. albicans* (C_alb) *Neurospora crassa* (N_cra), *Aspergillus nidulans* (A_nid), *Schizosaccharomyces pombe* (S_pom), *Ustilago maydis* (U_may), *Schizophyllum commune* (S_com), *Danio reri* (D_rer), *Drosophila melanogaster* (D_mel), *Caenorhabditis elegans* (C_ele). The position of the frameshift/gene split in *S. cerevisiae* is indicated by the pink arrow.

URA6 was initially annotated as a gene coding for a 204 amino acid protein in *S. cerevisiae.* Later an uncharacterised 75 codon open reading frame situated just upstream of *URA6*, called *YKL023C-A*, was identified by comparative sequencing of six *Saccharomyces* species (Cliften et al. 2003; Kellis et al. 2003). As part of a large-scale cDNA analysis, Miura et al. found this ORF to be cotranscribed with *URA6* in a single polycistronic transcript, and found no evidence to indicate that *URA6* was transcribed monocistronically (Miura et al. 2006). We noticed that the annotated *URA6* homologs in yeast species in the non-WGD clade code for proteins that are approximately 100 amino acids longer than *S. cerevisiae URA6*, and this extended N-terminal region shows protein sequence alignment with Ykl023c-a (Figure 4.13). This indicates that the *YKL023C-A/URA6* ORF pair (referred to from here on as ORF1 and

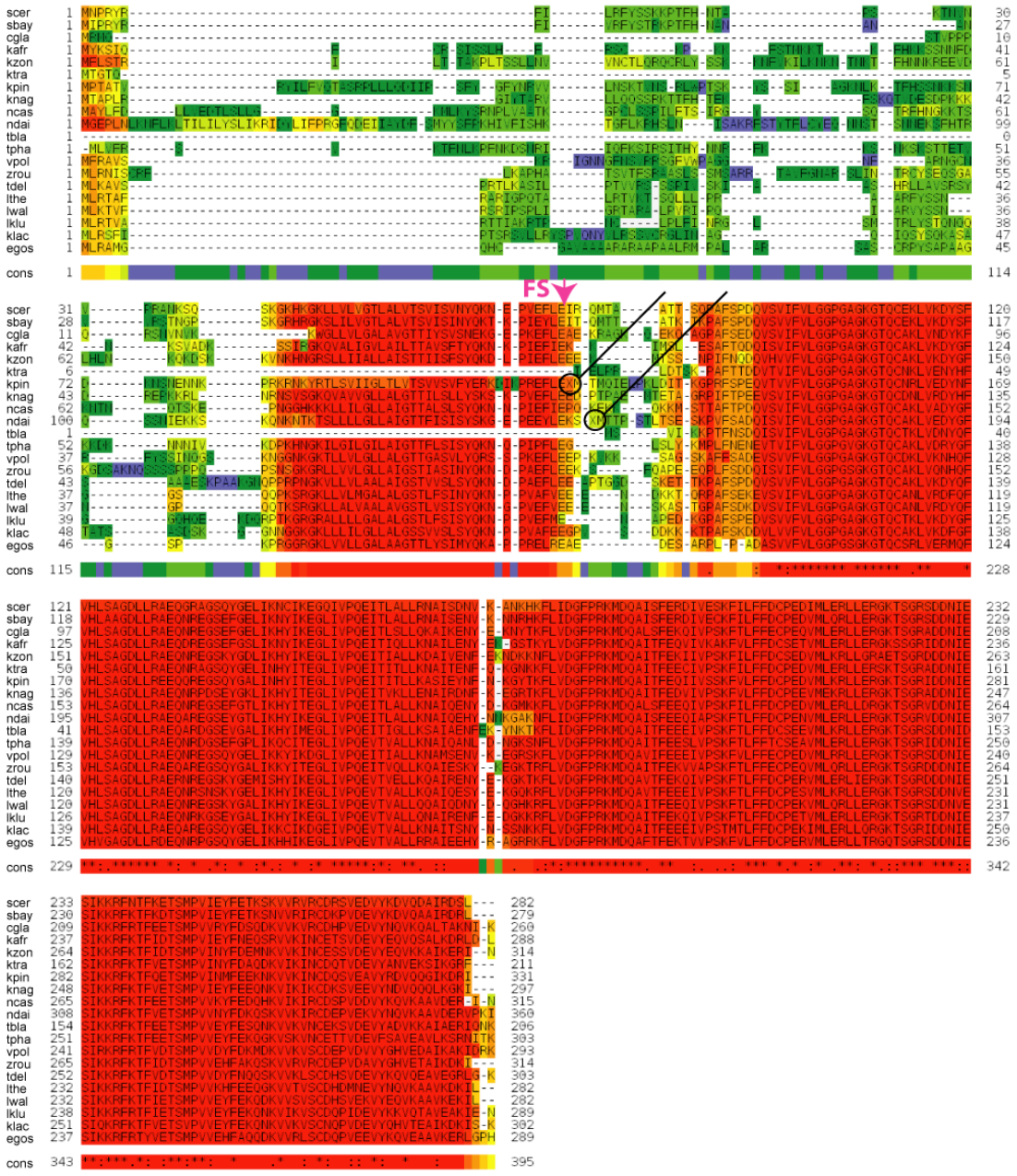

Figure 4.13 MCOFFEE alignment (Wallace et al. 2006) of the protein sequences of 20 yeast Ura6 orthologs. Species acronyms are as for Figure 4.3, with the addition of *Kazachstania zonata* (kzon), *K. transvaalensis* (ktra) and *K. pintolopesii* (kpin) Under the assumption that this is a frameshifting locus, the *S. cerevisiae* and *S. bayanus* orthologs require a -1 frameshift at the position indicated by the black arrow to produce a full-length protein, and a +1 frameshift at this location is required in *K. africana* and *T. phaffii*. This protein sequence conservation could also be explained by a gene split at this location. A gene split has occurred in *K. pintolopesii* and *N. dairenensis* and the "X" circled in black indicates the end of the upstream ORF and the beginning of the downstream ORF in these species. The *T. blattae*, *K. transvaalensis* and *K. zonata* orthologs have lost the 5' end of the gene.

ORF2 respectively) represent either a gene that has become split in the genus *Saccharomyces* or two separate parts of a gene in which a frameshift occurs.

URA6 exists as a single uninterrupted ORF (from here on denoted *URA6*^{full}) in all of the non-WGD species as well as in the CTG group yeasts studied and *Yarrowia lipolytica* (data not shown), indicating that this is clearly the ancestral state. Of the post-WGD species studied, *V. polyspora, N. castellii, K. naganishii* and *C. glabrata* have a *URA6*^{full} homolog. In order to learn more about this unusual locus, we studied the sequence of *URA6* in three additional *Kazachstania* species for which we had obtained lower-quality whole genome sequence data: *K. transvaalensis* and *K. pintolopesii*, which are more closely related to the *URA6*^{full}-containing *K. naganishii*; and *K. zonata*,which is more closely related to the frameshifting/split *URA6*-containing *K. africana*.

In *K. transvaalensis* and *T. blattae*, only an ORF corresponding to ORF2 in *S. cerevisiae* could be identified (Figures 4.13, 4.14). In *K. pintolopesii* and *N. dairenensis*, a gene split appears to have occurred at the approximate point at which the split/frameshift has occurred in *S. cerevisiae*; in both species the two resulting ORFs do not overlap and cannot be bridged by a frameshift. The ORFs are separated by a distance of 181 bases in *K. pintolopesii* and 37 bases in *N. dairenensis*, and the protein sequences of both ORFs in each of these species align to the regions corresponding to ORF1 and ORF2 in the *URA6*^{full}-containing species. Lastly, *K. africana* and *T. phaffii* both have ORF1 and ORF2 orthologs. In these species a full-length readthrough could be obtained by a +1 frameshift, although it also possible that a gene split has occurred in these species resulting in two separate ORFs. No evidence was found to suggest the existence of an intron in any of the homologs at this locus.

Therefore we propose two alternative scenarios to explain evolution at the *URA6* locus in the post-WGD Saccharomycetaceae species; Scenario (a) proposes a single or multiple gene split events occurring at this locus. Scenario (b) proposes a

complex pattern of +1 frameshifting, -1 frameshifting and gene split events occurring at this single location in the *URA6* gene at different points in the phylogenetic tree. Both these scenarios are problematic, as they appear to require multiple independent events and are thus unparsimonious, but it is unclear how the data can be explained by a single event or more simple chain of events. Ribosomal frameshifting has not been demonstrated experimentally in the *URA6* gene of any species.

Cartoon structures of the *URA6/YKL023C-A* locus under the two alternative scenarios of frameshift introduction and gene splitting are shown in Figure 4.14, and the nucleotide sequences at the junction site are shown in Figure 4.15.

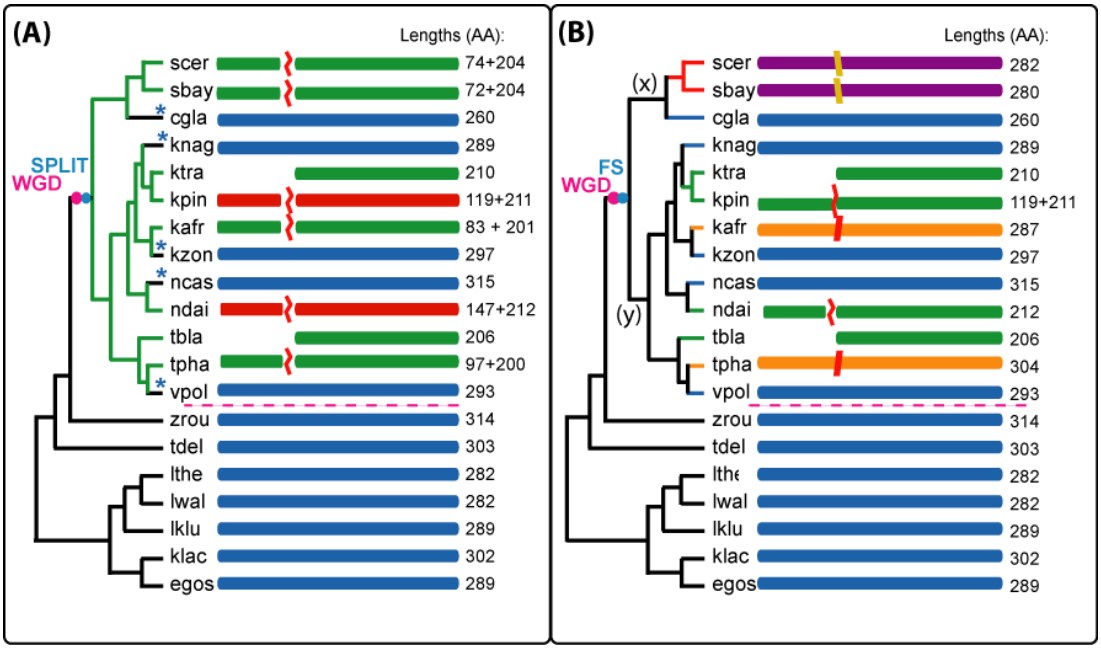

Figure 4.14 Cartoon illustrating the two different models proposed to explain evolution at the *URA6* locus in the 20 yeast species studied. Panel (A) illustrates the gene split model, with the split points indicated by the red "lightning bolt". Here split genes are coloured green, with the exception of *K. pintolopesii* and *N. dairenensis*; these are coloured red to indicate that the two ORFs resulting from gene split cannot be accounted for by a frameshift model. Branches leading to split genes are coloured green. Asterisks denote lineages which have reverted to $URA6^{full}$ at this locus. Panel (B) illustrates the frameshifting model; here the yellow back slashes indicate a -1 frameshift, and the red forward slashes indicate a +1 frameshift. Orthologs that have undergone a gene split or have lost the ORF1 region of full-length *URA6* are coloured green. The branches along which the frameshift signal has been lost resulting in a gene split or truncation are coloured green. The branches in which the frame has become corrected to produce a full-length, single-frame ortholog are coloured blue.

Scenario (A): Multiple gene split events at the URA6 locus

Under Scenario (A), I propose that one or more gene split events may have occurred in the lineages that emerged post-whole genome duplication. The most parsimonious model would appear to be one in which a gene split occurred in one of the two *URA6* ohnologs formed by whole genome duplication in the ancestor of these species. Provided both ohnologs were retained in these lineages until relatively recently in evolutionary history, a model involving seemingly arbitrary loss of either the full-length or split gene would satisfactorily account for the distribution of split and full-length genes illustrated in Figure 4.14A. The presence of ORF2 only in *K. transvaalensis* and *T. blattae* can be explained by two separate recent losses of ORF1 from the split ORF pair retained in these species.

However, there is evidence to rule out this hypothesis. According to this model, the split homologs should all be orthologous to each other, and paralogous to the full-length homologs. Yet analysis of synteny at this locus using YGOB indicates that these genes are all orthologous regardless of whether they are split or intact. Furthermore, this model requires the loss of one of two ohnologs in all lineages in recent evolutionary history. Both the timing and number of separate loss events necessary to support this model seem unlikely.

An alternative gene split model assumes the loss of one ohnolog following whole gene duplication in the ancestor of these species and the introduction of a gene split in the remaining ohnolog. This would need to be followed by five separate "gene re-fusion" events in different lineages to produce the distribution seen in these species (indicated by asterisks in Figure 4.14A). If we instead assume that $URA6^{full}$ represents the ancestral state in the surviving ohnolog, then no less than six separate gene split events is required, with each gene split occurring in roughly the same location in the gene independently.

Both of these models require an unlikely number of events to occur; however there appears to be no clear way to account for the data with a single event or simple set of events under this "gene split" scenario or that of frameshifting (discussed below).

Miura et al's (2006) observation that in *S. cerevisiae* ORF1 and ORF2 are never transcribed as separate mRNAs does not rule out a gene split model. Reinitiation may lead to the translation of both ORFs from a single bicistronic mRNA. While ORF1 is too large to fit with a model of reinitiation (at ORF2) after translation of a small upstream ORF (ORF1) (Kozak 2001), it is feasible that reinitiation at ORF2 occurs via an internal ribosomal entry site (IRES). However, IRES sites are difficult to identify and it is likely that many IRES sites have yet to be characterised (Baird et al. 2006; Mokrejs et al. 2006), making it difficult to verify this hypothesis.

Scenario (B): +1 frameshifting, -1 frameshifting and gene splitting at the URA6 locus

Under Scenario (B), I propose that in *S. cerevisiae, S. bayanus* and other *Saccharomyces*species, a -1 frameshift occurs at the sequence GAAUAA (corresponding to the last 6 nucleotides in the ORF1 sequence), resulting in the translation of the sequence GAAUAA as GAA-AUA-A (Figure 4.15). This would allow the UAA stop codon to be read through and the frame to be switched to that of the annotated *URA6* ORF (ORF2), producing a 282 amino acid protein that aligns over its entire length with the $URA6^{full}$ orthologs in the non-WGD species. Both the GAAUAA sequence and the necessity for a -1 frameshift are conserved in all the other *Saccharomyces*species studied (*S. bayanus, S. paradoxus, S. mikatae, S. kudriavzevii*) (data not shown). This sequence has little similarity to the "slippery heptamer" X-XX.Y-YY.Z sequence to which most -1 frameshift sites identified to date conform (Jacks et al. 1988a; Farabaugh 2010). It is also difficult to see how this sequence would allow -1 frameshifting by a mechanism involving a "backwards slip" by the ribosome. However, stop codons in the -1 and 0 frame pinpoint this sequence as the only possible frameshift position in *S. cerevisiae* (Figure 4.15) and this is

supported by high conservation of a protein sequence motif in the 0 frame directly upstream of the proposed site (Figure 4.15). Thus it appears that frameshifting at this site either takes place by a novel mechanism, or occurs at very low frequency.

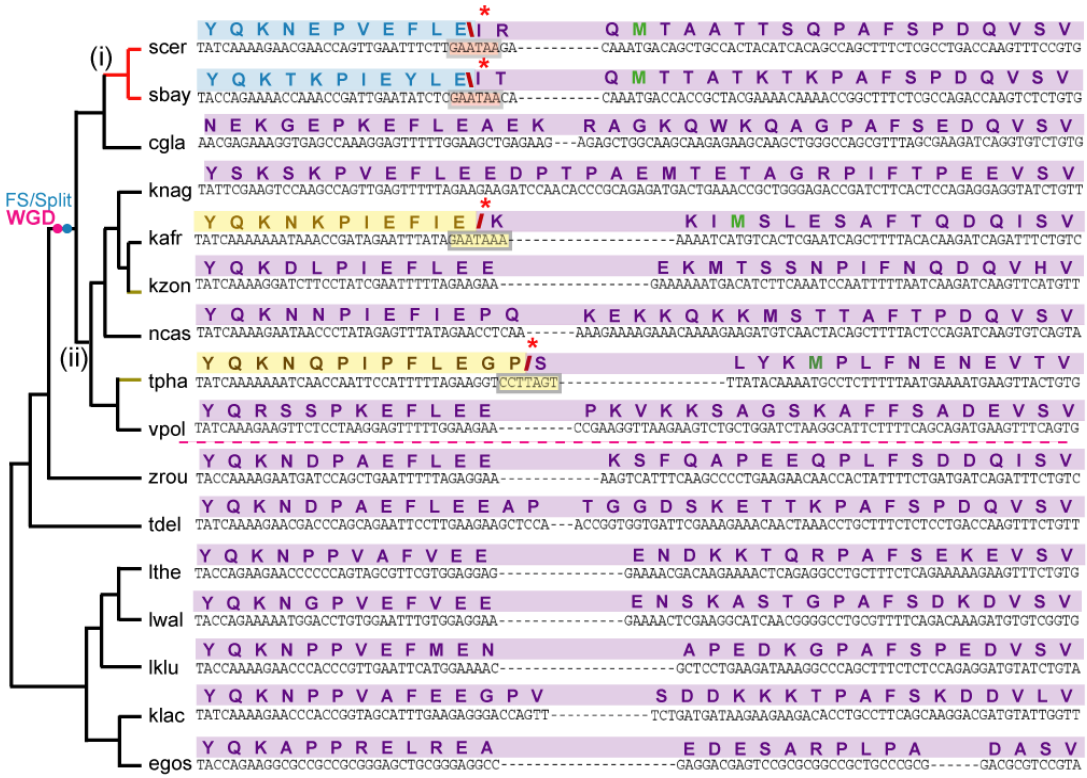

Figure 4.15 ClustalW nucleotide alignment (Larkin et al. 2007) of the region including and surrounding the *URA6* frameshift/gene split in 16 yeast species. *N. dairenensis*, *K. pintolopesii*, *K. transvaalensis* and *T. blattae* have been excluded for clarity; the truncated 3' only nature of the *K. transvaalensis* and *T. blattae* orthologs and the distance of noncoding region between the ORFs making up the split homolog in *N. dairenensis* and *K.pintolopesii* make it difficult to align these sequences meaningfully at this location. Branches containing potential frameshifts are highlighted in red and yellow on the species tree. The position of the most recent common ancestor of all frameshift/split-containing species is indicated on the tree by the blue circle which coincides with the position of the WGD (indicated by the pink circle). The dashed pink line separates post-WGD and non-WGD species. (i) A -1 frameshift at the six base pair sequence highlighted in red would be necessary in order to produce full-length Ura6 in *S. cerevisiae* and *S. bayanus* via frameshifting. The protein sequence of the *S. cerevisiae* Ura6 frameshifting product is shown above the nucleotide sequence, with the position of the frameshift indicated by the red backslash. (ii) A +1 frameshift would be required to produce a single full length Ura6 product in *K. africana* and *T. phaffii*. Protein conservation data indicates that the location of the frameshift in these species would have to be at the heptamer highlighted in yellow in *K. africana*, and at a heptamer located within the 13bp sequence highlighted in *T.* phaffii. The tree is not drawn to scale.

It may also be worth noting that while on average the genic regions corresponding to ORF 1 and ORF2 in the species studied have the same GC content (42.5% and 43%

respectively), the species containing an apparent -1 frameshift show a notably asymmetrical %GC (38.2% and 45.6% respectively for *S. cerevisiae* compared to a genome average 37.7%; 40.6% and 48.9% respectively for *S. bayanus* with a genome average of 39.6%). This may indicate the presence of frameshift-stimulating RNA secondary structures such as pseudoknots downstream of the GAAUAA sequence, as proposed by Ivanov et al. in the case of *E. gossypii OAZ1* (Ivanov et al. 2006). Downstream pseudoknots are widely considered to play a role in -1 frameshifting, although the mechanism by which they do so is still unclear (Kontos et al. 2001; Plant et al. 2003; Namy et al. 2006).

In two species, *K. africana* and *T. phaffii*, two adjacent ORFs corresponding to *S. cerevisiae* ORF1 and ORF2 exist, and a +1 frameshift at a position that aligns closely with that of the proposed -1 frameshift in *S. cerevisiae* could allow these ORFs to be bridged in order to produce a full-length protein. Protein conservation data and the position of the stop codon constrain the location of a potential +1 frameshift in *K. africana* to the heptamer GAA-U-AAA at positions 247-253 in the nucleotide sequence (Figure 4.15), although this does not match any previously reported +1 frameshifting heptamers. In *T. phaffii*, the location of stop codons indicates that a potential frameshift must occur at a heptamer within the 13 base sequence GAAGGUCCUUAGU (nucleotide positions 289-301). We propose that it is most likely to occur at CCU-U-AGU. In both these cases, it seems likely that the stop codon in the 0 frame would generate a pause that allows a shift into the +1 frame, as happens in *OAZ1* and *ABP140* frameshifting. However, in other respects these heptamers are not a perfect fit for the tRNA-based models proposed by Ivanov et al., Farabaugh et al. and others for *OAZ1, EST3* and *ABP140* and the yeast Ty elements. Firstly, a cognate tRNA exists in each case to decode the 0 frame triplet (GAA in *K. africana*, and CCU in *T. phaffii*) at the P site, removing the requirement for a near-cognate, non-legal wobble base-pairing interaction.
The second issue relates to tRNA abundance at the A site. Based on the number of genes coding for the relevant tRNAs (which has been found to correlate well with tRNA abundance (Percudani et al. 1997)), while the cognate tRNA for the AAA

codon in the +1 frame at the A site is abundant in *K. africana* (8 copies of the $tRNA^{Lys}_{UUU}$ gene), the tRNA corresponding to AGU is not especially common in *T. phaffii* , with 3 copies of the $tRNA^{Ser}_{GCU}$ gene in the *T. phaffii* genome (http://wolfe.gen.tcd.ie/ygob; all tRNA annotations for YGOB species carried out using tRNA-scan (Lowe and Eddy 1997)).

Several alternative explanations can be offered in support of the frameshifting hypothesis. Frameshifting at this heptamer may be promoted by cis-acting sequences such as those identified for other +1 frameshifting genes. Alternatively, these heptamers may stimulate a high degree of frameshifting by themselves, but by a different mechanism than those proposed to date. Without experimental measurements of full-length to truncated product, we also cannot rule out the possibility that frameshifting may simply be inefficient, and thus occur at a low frequency in these species. It is interesting to note that the proposed *T. phaffii URA6* +1 frameshift heptamer CCU-U-AGU overlaps by 6 bases with the CUU-A-GUU sequence that promotes +1 frameshifting in *EST3*. Whereas this heptamer in *URA6* is in the wrong frame to function by the mechanism proposed for *EST3*, it is conceivable that other properties of this combination of nucleotides could have a frameshift-stimulating effect regardless of frame.

The available data paints a complex picture of frameshifting in *URA6*. As previously described,$URA6^{full}$ clearly represents the ancestral state for the species studied. The ancestral point at which a frameshift in *URA6* appears in the tree is indicated in Figure 4.14B, and coincides with the whole genome duplication. Of the post-WGD species studied, only four (*C. glabrata, K. zonata, K. naganishii* and *N. castellii*) have a $URA6^{full}$ ortholog; the others contain a +1 frameshift, a -1 frameshift, a truncated 3'-only *URA6* ($URA6^{3'}$) or a gene split. Under the assumption that it is more likely for a frameshift to be lost in several lineages tham for frameshifts to be gained at roughly the same location in different species, it is parsimonious to suggest that an event occurred on the branch leading to the post-WGD species to disrupt the *URA6* reading frame which was initially resolved by a frameshift in this ancestor.

The patchy distribution of frameshift and truncation events among the post-WGD species indicates that several subsequent events must have occurred in different lineages. Under the assumption that a -1 frameshift-containing gene (denoted *URA6*$^{-}$) was the ancestral state at position (x) and a +1 frameshift-containing gene (*URA6*$^{+}$) existed at position (y) in Figure 4.14B, an event must have occurred to convert the *URA6*$^{+}$ to *URA6*$^{-}$ or vice versa along one of these branches, depending on which of them is the original state. From there, it appears that the frameshift-containing gene must have corrected itself back to *URA6*^{full} in five separate events: from *URA6*$^{-}$ to *URA6*^{full} in *C. glabrata*, and three separate instances of *URA6*$^{+}$ to *URA6*^{full} in the branches leading to *K. zonata*, *K. naganishii*, *N. castellii* and *V. polyspora*. Evolutionary conversion of +1 frameshifts to standard 0 reading frames have previously been identified in *EST3* and *ABP140*. Furthermore, interconversion between -1 frameshifting and +1 frameshifting may not be that difficult even if a different mechanism is employed; the -2 frameshift that occurs when the mammalian antizyme frameshift-stimulating sequence (which stimulates a +1 frameshift in mammals) is expressed in yeast indicates that even the same sequence may stimulate plus direction and minus direction frameshifting in different contexts (Matsufuji et al. 1996).

Given that the protein alignments indicate a truncation point that coincides quite closely with the position of the putative frameshift, it is logical to infer that the loss of the putative frameshift signal may have led to the loss of the *URA6* region upstream of the frameshift in *N. dairenensis*, *T. blattae*and *K. traansvalensis*, and to the gene split seen in *K. pintolopesii*. Phylogenetic distribution indicates that three separate frameshift signal loss events would be required, with a single frameshift signal loss leading to the gene split in *K. pintolopesii* and a *URA6*$^{3'}$ ortholog in *K. transvaalensis* (Figure 4.14).

Differentiating between gene split and frameshift

Both the gene split and the frameshift models I present to explain evolution at the *URA6* locus are problematic. Both appear to require an unlikely number of events. The six separate gene split events or five gene fusion events required for the gene split model seems astronomically unlikely, particularly given both that all the gene split/fusion events are constrained to recent evolutionary history, and that the model requires the split to happen at the same location in multiple orthologs in independent events. However, it is not inpossible that genes such as *URA6* have specific points at which a split is far more likely to occur, meaning that independent gene split events in different orthologs could produce similar split ORFs.
Similarly for the frameshift model, it is quite feasible that loss of a frameshift signal could lead to a gene split and/or loss of part of a gene, and the evidence that differences in genomic context can cause a -2 frameshifting sequence to become a +1 would suggest that minor changes in these sequences or their surroundings could lead to a change in the type of frameshift promoted. However, it is less clear how unlikely reversion to a full-length non-frameshifting ORF is, and our model requires several of these steps. Yet it is difficult to envisage a simpler model to account for the data.

One obvious step to test these models is to express and translate the *S. cerevisiae* Ura6 protein *in vitro*. Under the frameshifting model, we would expect to get at least a fraction of full-length Ura6, depending on how efficient frameshifting is. Furthermore, we would not expect any protein products corresponding to ORF2 only. Under the gene split model on the other hand, we expect two separate protein products corresponding to ORF1 and ORF2, and no protein corresponding to $URA6^{full}$. It is clear that an exceptionally complex process of structural evolution has occurred at the *URA6* locus in post-WGD species, and that the tools of molecular biology will need to be brought to bear in order to determine the exact manner in which *URA6* is translated in *S. cerevisiae* and other post-WGD species.

Conservation of the ORF1 region

To verify that ORF1 codes for a protein that is evolutionarily conserved, we carried out *Ka/Ks* comparisons of the region corresponding to ORF1 in both frameshift/aplit-containing and *URA6*^{full} orthologs. With the exception of pairwise comparisons involving the *T. delbrueckii* and *N. dairenensis* orthologs (which surprisingly both show a *Ka/Ks* >1 over this region in most comparisons against their orthologs), all other pairwise comparisons produced *Ka/Ks* values between 0.01 and 0.67 (with the majority of values below 0.4), indicating that despite the very high divergence upstream of the frameshift the protein sequence of these regions shows conservation.

Similarly to *ABP140*, the *URA6* region upstream of the frameshift/split is far more divergent in both sequence and length than the region downstream in both frameshift/split-containing and *URA6*^{full} orthologs, with only a 30 residue sequence just upstream of the frameshift showing significant protein conservation (Figure 4.13).

The region of the gene corresponding to ORF2 downstream of the frameshift shows exceptionally high levels of protein conservation across eukaryotes (Figure 4.12), consistent with the ancient and fundamental cellular role of the Ura6 protein. Outside of the Saccharomycetaceae however, many of the species studied, including *Aspergillus nidulans*, *Drosophila melanogaster* and *Homo sapiens*, have a homolog aligning only to ORF2 and little to no upstream sequence. Others, such as the basidiomycete *Ustilago maydis*, have a homolog with an N-terminal region up to 190 amino acids in length. The extreme variation in length and sequence at the 5' end of *URA6* make it difficult to draw any conclusions about whether *URA6* with the 5' extension was the ancestral state or whether these 5' extensions happened independently in different lineages. However, the fact that this extension is missing in so many species and so variable in the species that have it indicates that it is not required for the function of this essential enzyme. It may have species-specific functions, such as protein localisation signals.

4.4 Discussion

The three frameshift-containing yeast chromosomal genes identified to date and the potential fourth example, *URA6*, share some intriguing commonalities. These are genes that are exceptionally highly conserved across species, coding for proteins that play integral roles in cellular function: telomere maintenance, polyamine regulation, tRNA modification and pyrimidine synthesis. All four are conserved in all species studied within the Saccharomycetaceae, and for three of the four (*EST3, ABP140*, and *URA6*) a homolog existed in each CTG group species studied. *OAZ1, ABP140* and *URA6* contain domains highly conserved across all eukaryotes.

URA6 and the full-length *ABP140* homologs both contain ORF1 regions upstream of the position of the frameshift (in frameshift-containing genes) that shows high divergence in length and sequence among the Saccharomycetaceae. *ABP140* homologs in the CTG species are entirely lacking the ORF1 region. *URA6* homologs outside of the Saccharomycetaceae show extreme variation length and sequence in the region corresponding to ORF1, whereas the region corresponding to ORF2 is highly conserved across eukaryotes.

In the cases of *EST3* and *ABP140*, the available data points to a point of origination of the frameshift signal on the branch leading to the Saccharomycetaceae (Figure 4.1). The *OAZ1* frameshift is much older based on its presence across eukaryotes, and the *URA6* frameshift/gene split appears to have originated on the branch on which the whole genome duplication occurs.

It is tempting to speculate that the WGD may have given evolution the opportunity to tinker with the translation of *ABP140* and *URA6*. Only species having undergone WGD have the *URA6* gene split/frameshift, although within the post-WGD species, several gene-splitting/frameshift introduction events appear to have occurred (Figure 4.14). The gene redundancy produced by the WGD means that in the ancestor of

post-WGD species there existed an extra copy of *URA6* that may have been under relaxed selective pressure, and thus more amenable to changes in translational regulation. However, a curious observation is that the multitude of different states that exist in present day species point to no clear advantage between the full-length ORF and the gene in a split or frameshift-requiring state.It may of course be that the gene split/introduction of the frameshift and potential loss of the intact ohnolog was a pure evolutionary accident. In the case of *ABP140*, the WGD may have allowed for tweaking of the ratio of full-length product to ORF1-only (i.e. translation events in which the ribosome fails to shift to the +1 frame) through the retention of "ORF1-only" ohnologs (Figure 4.8). In two cases the species has lost the frameshift in the full-length ohnolog but gained an "ORF1-only" ohnolog, potentially maintaining a situation in which ORF1-only proteins and full-length proteins are produced in a specific ratio.

In the Saccharomycetaceae species studied, the frameshifting requirement is conserved in all species for *OAZ1*, in all except the *Kluyveromyces* species for *EST3*, and shows patchier retention both for *ABP140* and among the post-WGD species for *URA6*. Introduction of a frameshift must have an effect on the levels of full-length protein produced from a locus, given that the percentage frequency at which successful frameshifting occurs varies both between genes and with cellular conditions. The translation product of a failure to shift frames may or may not be functional and relevant to the cell. As previously discussed, the frameshift may result in a required stochiometric ratio between full-length and non-frameshifted products. The distributions of frameshift retentions among closely related species may allow us to see how the cell has adapted to the introduction or loss of the frameshift. For example, in a pair of closely related species where one species has a frameshift, have promoter sequences and/or levels of transcription increased in the frameshift-containing species to compensate for the lower percentage of full-length translation events? If, as suggested by Plant et al. (2004) for -1 frameshift sites, the frameshift is used as a method to target proteins for nonsense-mediated decay, how is this level of regulation achieved in the non-frameshift-containing species? If the truncated product

is functional, what role does it play and how do species lacking this truncated protein compensate?

The genes discussed in this chapter represent the only known chromosomal genes to date to employ frameshifting in yeasts, however genes with undiscovered frameshift signals may well exist. A study by Shah and colleagues (2002) identified a heptamer supporting significant frameshifting, GCU-C-AGA, that is unrelated to all currently known +1 frameshift sites, indicating that more +1 frameshift sites may lie undiscovered. Studies by Jacobs et al. (2007) and Theis et al. (2008) identified a multitude of sites that could potentially support -1 frameshifting. These four genes also represent examples of genes in which a programmed frameshift leads to the translation of a longer protein than would be produced by remaining in the 0 frame. However, it is possible that programmed ribosomal frameshifts exist that stimulate the production of a shorter protein than is produced when frameshifting does not occur. In *E. coli* translation of the *dnaX* gene involves programmed -1 frameshifting, leading to synthesis of equal ratios of a C-terminally truncated form of the protein (the γ subunit of DNA polymerase III) as well as the full-length, non-frameshifted product (the τ subunit) (Blinkowa and Walker 1990; Tsuchihashi 1991). Thus, it is likely that further examples of loci employing programmed ribosomal frameshifting to subvert the standard genetic code and produce alternate gene products await discovery.

4.5 Methods

Sequencing of *K. naganishii, K. africana, K. zonata, K. pintolopesii, K. transvaalensis, N. dairenensis, T. blattae, T. phaffii, T. delbrueckii* and *K. marxianus* was carried out using the Roche FLX platform and assembled as described in Gordon et al. (in preparation; Appendix II). Genome annotations were carried out using YGAP (the *Y*east *G*enome *A*nnotation *P*ipeline) with manual modification (Proux-Wéra et al., in preparation). Genome sequence and annotations for *K. naganishii, K. Africana, N. dairenensis, T. blattae, T. phaffii, T. delbrueckii* are available at

http://wolfe.gen.tcd.ie/ygob. Genome sequences and annotations corresponding to the Saccharomycetaceae are listed in Table 4.1 and Appendix I: Table S2.1.

Protein and nucleotide sequences for the CTG group yeasts were retrieved from the Candida Gene Order Browser (Fitzpatrick et al. 2010); the genome sequences and annotations used in CGOB are listed in Appendix I: Table S2.1. Protein and nucleotide sequences for homologs outside these yeast groups were retrieved from GenBank (Benson et al. 2011).

Annotation modification to include the frameshift was carried out using the annotation tool Artemis (Rutherford et al. 2000). Gene coordinate modifications are listed in Table 4.2. Nucleotide alignments were carried out using ClustalW (Larkin et al. 2007) and MUSCLE (Edgar 2004) and included manual modification. Protein sequence alignments were carried out using T-Coffee (Notredame et al. 2000). *Ka/Ks* calculations for the *ABP140* and *URA6* ORF1 regions were carried out using yn00, part of the PAML suite of software (Yang 2007).

The phylogenetic relationship between the *Kazachstania* species was determined by analysing the protein sequences of the genes *ECM38* (*YLR299W), RPO41 (YFL036W)* and *URA6 (YKL024C)* using Phyml (Guindon et al. 2009) as part of the Seaview progam (Gouy et al. 2010) (Figure 4.16). 100 bootstraps were used.

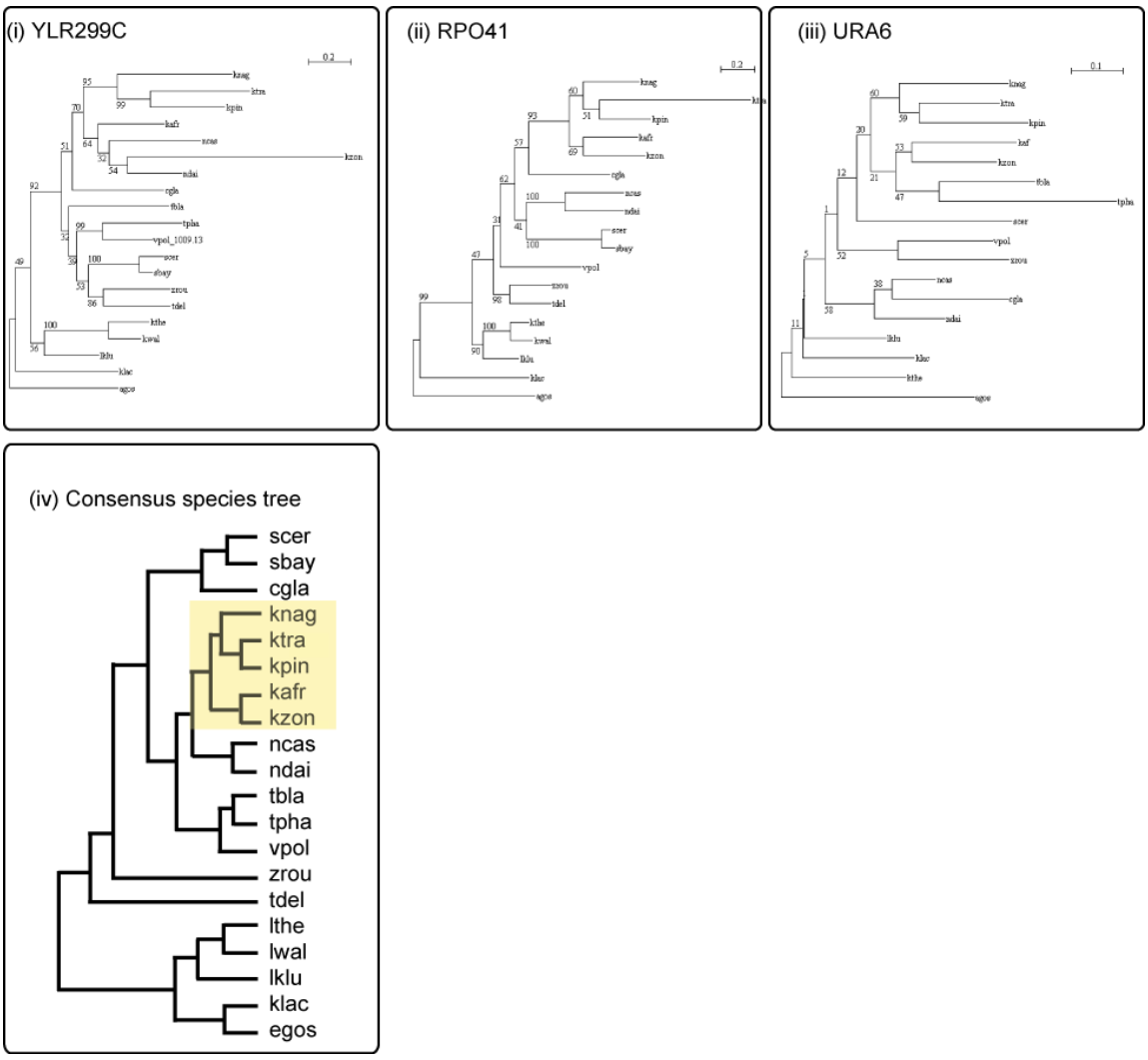

Figure 4.16 Bootstrap consensus tree of *Kazachstania* species from analysis of the protein sequences of three genes: (i) *ECM38* (*YLR299W)*; (ii) *RPO41 (YFL036W)*; (iii) *URA6 (YKL024C)* (Guindon et al. 2009) All three analyses separate the *Kazachstania* species into the same grouping, highlighted in yellow in the species tree in (iv).

Table 4.1 Genome sequences and annotations for the Saccharomycetaceae yeast species used in this study, in addition to those already listed in Table 2.2.

Species	Clade	Coverage	Sequence	Gene annotation
Kazachstania Africana	Kazachstania (2)	>20x	Gordon et al (in preparation; Appendix II)	Proux-Wéra et al (in preparation)
Kazachstania naganishii	Kazachstania (2)	>20x	Gordon et al (in preparation)	Proux-Wéra et al (in preparation)
Naumovozyma	Naumovozyma (3)	26x	Gordon et al (in	Proux-Wéra et

castellii			preparation)	al (in preparation)
Naumovozyma dairenensis	Naumovozyma (3)	>20x	Gordon et al (in preparation)	Proux-Wéra et al (in preparation)
Tetrapisispora blattae	Tetrapisispora (5)	>20x	Gordon et al (in preparation)	Proux-Wéra et al (in preparation)
Tetrapisispora phaffii	Tetrapisispora (5)	>20x	Gordon et al (in preparation)	Proux-Wéra et al (in preparation)
Torulaspora delbrueckii	Torulaspora (9)	>20x	Gordon et al (in preparation)	Proux-Wéra et al (in preparation)
Additional *Kazachstania* species used in the *URA6* comparative analysis:				
Kazachstania transvaalensis	Kazachstania (Clade 2)	>20x	Unpublished data	Incomplete
Kazachstania zonata	Kazachstania (2)	>20x	Unpublished data	Incomplete
Kazachstania pintolopesii	Kazachstania (2)	>20x	Unpublished data	Incomplete
Additional *Kluyveromyces* species used in the *EST3* comparative analysis:				
Kluyveromyces marxianus	Kluyveromyces (Clade 11)	>20x	Unpublished data	Incomplete

Table 4.2 Coordinates of the genes Saccharomycetaceae homologs of the frameshifting genes studied in this chapter. Coordinates marked with an asterisk represent coordinates modified to correctly include the location of the frameshift. For*S. cerevisiae*, *S. bayanus*, *K. africana* and *T. phaffii* two sets of coordinates are given; these represent the two alternative scenarios of frameshifting and gene split for these orthologs. Abbreviations used: Chr. (chromosome) c (contig) s/scf (scaffold). Note that coordinates in contig or scaffold assemblies are subject to change.

Species	*EST3*	*OAZ1*
S. cerevisiae	Chr. 9 complement(335663..335932,335934..336209)	Chr. 16 (458796..459002,459004..459675)
S. bayanus	c598 (23200..23472,23474..23743)	c521 (10212..10418,10420..11091)*
C. glabrata	Chr. 3 (377164..377427,377429..377707)*	Chr. 13 (747630..747818,747820..748398)
K. naganishii	Chr. 12 complement(177940..178215,178217..178498)*	-
K. Africana	Chr. 12 complement (87343..87603, 87605..87886)*	Chr. 8 (257299..257613, 257615..258171)*
N. castellii	Chr. 5 complement(411717..411992,411994..412275)	Chr. 3 (435379..435594,435596..436195)
N. dairenensis	Chr. 5 complement(804960..805268,805270..805551)	Chr. 5 (641224..641481,641483..642112)
T. blattae	Chr. 1 complement(213576..213869,213871..214224)	Chr. 1 (84243..84962, 84964..85224)*
T. phaffii	Chr. 3 (931833..932096..932098..932439)*	Chr. 3 (382922..382987,382989..383618)*
V. polyspora	s1062 (125085..125363,125365..125634)*	s333 (742..972,974..1615)
Z. rouxii	Chr. 3 complement(410153..410401,410403..410660)	Chr. 6 complement(688910..689509,689511..689717)
T. delbrueckii	Chr. 7 complement(483201..483485, 483487..483768)	Chr. 1 (1326049..1326195,1326197..1326817)*
L. thermotolerans	Chr. 6 complement(1055444..1055740,1055742..1056026)	Chr. 5 (1039996..1040196,1040198..1040764)*
L. waltii	s55 complement(581681..581974,581976..582260)	s27 (986693..986893,986895..987461)*
L. kluyveri	Chr. 6 (413034..413324,413326..413625)	Chr. 8 (951723..951920, 951922..952509)*
K. lactis	Chr. 4 (1215729..1216283)	Chr. 2 (879228..879488,879490..880047)
K. marxianus	scf7180000047547 complement(497605..498147)	s7180000047543 (6535..6789,6791..7344)
E. gossypii	Chr. 4 complement(680590..680913,680915..681199)*	Chr. 7 (210150..210338,210340..210918)

Table 4.2 (cont.)

Species	*ABP140*	*URA6*
S. cerevisiae	Chr. 15 (784858..785688,785690..786745)	**Frameshift:** Chr. 11 complement(392169..392792,392792..393016)* **Split**: Chr. 11 complement(392169..392783) Chr11 complement (392792..393016)*
S. bayanus	c635 (35115..35828,35830..36870)	**Frameshift:** c562 (19488..19703,19703..20326)* **Split:** c562 (19488..19694)* c562 (19703..20326)*
C. glabrata	Chr. 10 complement(361909..362919,362921..363517)*	Chr. 12 (1056341..1057123)
K. naganishii	Chr.1 complement(570688..571758,571760..572797)	Chr. 8 complement(19327..20220)
K. transvaalensis	-	c7180000054305 (4009..4641)*
K. pintolopesii	-	c7180000059644 GENE SPLIT 1: (8460..8819) 2: (9001..9636)

K. africana	Chr. 1 (775451..776122, 776124..777218)	**Frameshift**: Chr. 3 complement(772907..773518,773520..773774)* **Split:** Chr. 3 complement(772907..773509) Chr. 3 complement(773520..773774)*
K. zonata	-	c7180000059644 (3517..4464)*
N. castellii	Chr. 2 (257668..259176) Chr. 3 (302917..303816)	Chr. 8 (676587..677534)
N. dairenensis	Chr. 5 complement(242410..243483,243485..243655) Chr. 5 (500822..501589)	Gene split Chr. 3 complement(19018..19656) Chr. 3 complement(19694..20136)
T. blattae	Chr2 complement(895002..896672) Chr. 3 (433181..434113, 434115..435068)	Chr. 3 (211081..211701)
T. phaffii	Chr. 7 complement(341818..342369) Chr. 4 (251042..252319)	**Frameshift:** Chr. 14 (413443..413739,413741..414355)* **Split**: Chr. 14 (413443..413739) Chr. 14 (413753..414355)*
V. polyspora	s1012 (25467..25886,25888..26913) s1072 complement(49934..50560)	s1057 (29809..30690)
Z. rouxii	Chr. 6 (576854..577543,577545..578543)	Chr. 4 complement(40520..41464)
T. delbrueckii	Chr. 2 (1077285..1077965, 1077967..1078892)*	Chr. 1 (2423160..2424071)
L. thermotolerans	Chr. 4 (896983..897654,897656..898669)	Chr. 3 (55308..56156)
L. waltii	s26 (924376..925101,925103..926116)	s14 (88176..89024)
L. kluyveri	Chr. 8 complement(600377..601444,601446..602225)	Chr. 7 complement(1702194..1703063)
K. lactis	Chr. 6 complement(1047731..1048744,1048746..1049333)	Chr. 5 complement(724562..725470)
E. gossypii	Chr. 3 (577065..577709,577711..578778)	Chr. 3 complement(653511..654380)

Appendix I

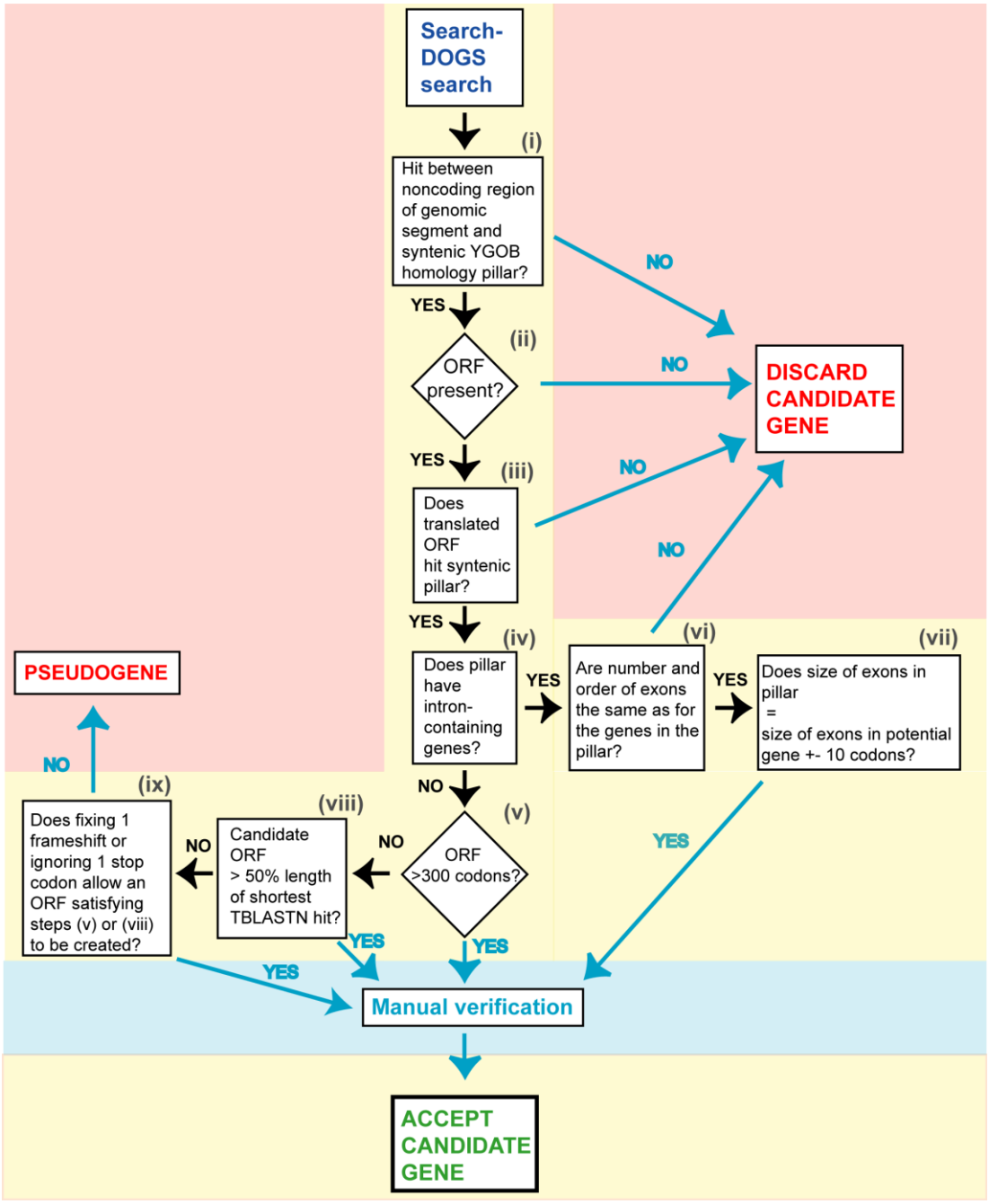

Figure S2.1 Flowchart illustrating the criteria that must be achieved for a SearchDOGS hit to be considered a *bona fide* gene. The automated steps are marked with roman numerals. All candidate genes that pass the automated steps are subjected to a manual examination before they are accepted.

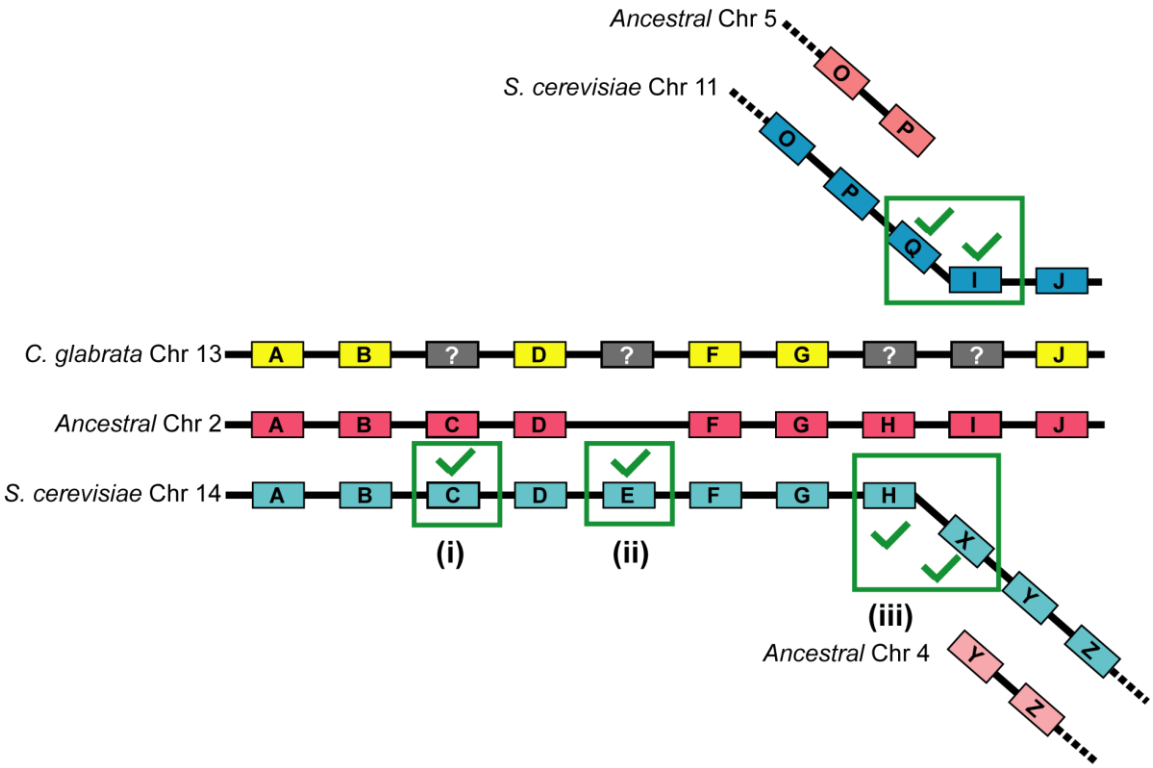

Figure S2

Figure S2.2 Cartoon illustrating the automated SearchDOGS method for establishing orthology between genomic segments. For ease of explanation, only two extant species are shown (*S. cerevisiae* in blue, and *C. glabrata* in yellow), as well as the Ancestral genome (pink; parts of Ancestral chromosomes 2, 4 and 5 are shown). (i) Testing for a possible *C. glabrata* ortholog of gene *C*, which is included in the Ancestral genome. For the *C. glabrata* genomic fragment *B-D*, the two annotated genes have Ancestral orthologs (*Anc_2.B* and *Anc_2.D*) that are less than 10 ancestral genes apart. In this situation, we use the intergenic region between *C. glabrata B* and *D* as a BLASTX query against a database that contains the translations of all genes that map between *Anc_2.B* and *Anc_2.D.* Therefore *S. cerevisiae* gene *C* is included in this database (green tick and box). (ii) Testing for a possible *C. glabrata* ortholog of gene *E*, which is not included in the Ancestral genome. For the *C. glabrata* genomic fragment *D-F*, the two annotated genes have orthologs in another species (*S. cerevisiae*) that are less than 10 genes apart. In this situation, we put each of the genes from that species (*i.e.*, *S. cerevisiae* gene *E*) in the database against which the *C. glabrata* intergenic region will be searched using BLASTX. (iii) Interspecies rearrangements. In this example, an interspecies rearrangement has occurred in *S. cerevisiae* relative to the Ancestor and *C. glabrata*, creating two new gene orders *G-H-X-Y-Z* and *O-P-Q-I-J* in *S. cerevisiae*. To search for possible unannotated genes in the interval between *C. glabrata G* and *J*, we define two orthologous *S. cerevisiae* genomic segments as follows. First, we consider the gene on the left end of the *C. glabrata* segment, *Cgla G*. We identify its *S. cerevisiae* ortholog from the same pillar (*Scer G*), and walk rightwards from this gene until we reach the point where synteny is lost (*Scer Y*; we know that synteny is lost because it is in a pillar with a different part of the ancestral genome, *Anc_4.Y*). We therefore put *S. cerevisiae* genes encountered on this walk (*Scer H* and *X*) into the database against which the *C. glabrata G-J* intergenic interval will be searched by BLASTX. Second, we similarly consider the gene on the right end of the *C. glabrata* segment, *Cgla J*, find its *S. cerevisiae* ortholog (*Scer J*) and walk leftwards in *S. cerevisiae* until synteny is known to be lost (at *Scer P*). We add the *S. cerevisiae* genes from this encountered on this walk (*Scer I* and *Q*) to

the database. Thus the *C. glabrata G-J* intergenic region will be used as a BLASTX query against a database containing *S. cerevisiae H*, *X*, *Q* and I.

Table S2.1 Genome sequences and annotations for the CTG group yeast species used in this study.

Species	Coverage	Sequence	Gene annotation
Candida albicans str. SC5314	Complete	(Braun et al. 2005)	(Braun et al. 2005)
Candida albicans str. WO1	Complete	(Butler et al. 2009)	(Butler et al. 2009)
Candida dubliniensis	Complete	(Jackson et al. 2009)	(Jackson et al. 2009)
Candida tropicalis	Complete	(Butler et al. 2009)	(Butler et al. 2009)
Candida parapsilosis	10x	Sanger Institute (http://www.sanger.ac.uk)	Guide *et al.* (in preparation)
Candida orthopsilosis	10x	Riccombeni *et al.* (in preparation)	Riccombeni *et al.* (in preparation)
Lodderomyces elongisporus	7-10x	(Butler et al. 2009)	(Butler et al. 2009)
Debaryomyces hansenii	Complete	(Dujon et al. 2004)	(Dujon et al. 2004)
Scheffersomyces stipitis	Complete	(Jeffries et al. 2007)	(Jeffries et al. 2007)
Meyerozyma guilliermondii	7-10x	(Butler et al. 2009)	(Butler et al. 2009)
Candida lusitaniae	7-10x	(Butler et al. 2009)	(Butler et al. 2009)

Table S3.1 List of loci at which orthologs of short (<60 codon) *E. coli* K12 MG1655 genes have been identified in the species studied. Species acronyms are as follows: ECK1: *Escherichia coli* K12 *substr.* MG1655, ECO1: *Escherichia coli* O157:H7 *str. Sakai*, ECS8: *Escherichia coli* S88, SBOY: *Shigella boydii* Sb227, SETY: *Salmonella enterica subsp. enterica serovar* Typhi str. Ty2, YPAN: *Yersinia pestis* antiqua, PSYR: *Pseudomonas syringae pv. tomato str.* DC3000, VCHO: *Vibrio cholerae* O395, XCAM: *Xanthomonas campestris pv. campestris str.* ATCC 33913.

Gene name	Length (codons)	Annotated in:	Unannotated ortholog found:	Protein function (Riley et al. 2006)
gnsA	58	ECK1 ECS8 SETY	SBOY	Multicopy suppressor of secG(Cs) and fabA6(Ts); predicted regulator of phosphatidylethanolamine synthesis
yciY	58	ECO1 ECS8	SBOY SETY YPAN	hypothetical protein

yciZ	58	ECK1 ECO1 ECS8	SBOY SETY	hypothetical protein
yjdO	58	ECK1	ECO1 ECS8	predicted protein
ymdF	58	ECK1 SETY	ECO1	conserved protein
ymgI	58	ECK1	ECS8 SBOY	hypothetical protein
rmf	56	ECK1 ECS8 SBOY ECO1 SETY	YPAN VCHO	ribosome modulation factor
yciX	56	ECK1 ECS8	ECO1 SBOY	hypothetical protein
yojO	55	ECK1	SBOY ECO1	hypothetical protein
ytlA	54	ECK1	ECO1 SBOY	predicted protein
hokD	52	ECK1 ECO1 SBOY	ECS8	Qin prophage; small toxic polypeptide
yhrJ	52	ECK1	SBOY	hypothetical protein
Sra	51	ECK1 ECO1 SBOY SETY	ECS8	Stationary-phase-induced ribosome-associated protein
hokB	50	ECK1 ECO1	ECS8 SBOY	toxic polypeptide, small
ecnB	49	ECK1 ECO1 ECS8 SETY	SBOY	entericidin B membrane lipoprotein
ygdT	49	ECK1	ECO1 ECS8 SBOY SETY	hypothetical protein
ygdT	49	ECK1	ECO1 ECS8 SBOY SETY	hypothetical protein
ypdI	48	ECK1	ECO1 ECS8 SBOY	hypothetical protein
yjjy	47	ECK1 ECO1 SBOY SETY	ECS8	predicted protein
ykgO	47	ECK1 ECO1 SETY YPAN VCHO XCAM	ECS8 SBOY	rpmJ (L36) paralog
ylcG	47	ECK1 ECO1	ECS8	expressed protein, DLP12 prophage
yqcG	47	ECK1	SBOY	expressed protein
ybhO	46	ECK1	SBOY ECS8 ECO1	hypothetical protein
sgrT	44	ECK1 ECS8	ECO1 SBOY	Inhibitor of glucose uptake
dinQ	43	ECK1	ECO1 SBOY	Damage inducible, function unknown
ydfB	43	ECK1 ECS8	ECO1	Qin prophage; predicted protein
ymiA	43	ECK1 ECS8	ECO1 SBOY SETY	hypothetical protein
Blr	42	ECK1 ECO1 ECS8	SETY SBOY	beta-lactam resistance membrane protein
ecnA	42	ECK1 ECS8 SETY	ECO1 SBOY	entericidin A membrane lipoprotein, antidote entericidin B
yqfG	42	ECK1	ECO1 ECS8 SBOY	expressed protein
rpmJ	39	ECK1 ECO1 SBOY SETY PSYR VCHO	ECS8 YPAN	50S ribosomal subunit protein L36
ybgT	38	ECK1 ECO1 SETY XCAM	ECS8 SBOY YPAN VCHO	conserved protein
yshB	37	ECK1	ECO1 ECS8 SBOY SETY	expressed protein
ldrA	36	ECK1	ECO1	toxic polypeptide, small
ldrB	36	ECK1 ECS8 SBOY	ECO1	toxic polypeptide, small
ldrB	36	ECK1 ECS8	ECO1 SBOY	toxic polypeptide, small
ldrC	36	ECK1	ECO1 SBOY	toxic polypeptide, small
yniD	36	ECK1 ECS8 SBOY	ECO1	predicted protein
yohO	36	ECK1 ECS8	ECO1 SBOY	predicted protein
ymiB	35	ECK1	ECO1 ECS8 SBOY	expressed protein
yoaI	35	ECK1 SBOY ECO1 ECS8	SETY	predicted protein
ykgR	34	ECK1	ECS8 SBOY	expressed protein
ylcH	34	ECK1	ECO1	hypothetical protein, DLP12 prophage
ilvL	33	ECK1 ECS8 SBOY ECO1 SETY	YPAN	ilvG operon leader peptide
yoaK	33	ECK1	ECO1 ECS8 SETY	expressed protein, membrane-associated
yncL	32	ECK1	ECO1 ECS8 SBOY SETY	hypothetical protein
yneM	32	ECK1	ECO1 ECS8 SETY	expressed protein, membrane-associated
yccB	31	ECK1 SBOY	ECO1 ECS8 SETY	hypothetical protein
yccB	31	ECK1 SBOY	ECO1 ECS8 SETY	hypothetical protein
tisb	30	ECK1	ECO1 ECS8 SBOY SETY	lexA-regulated toxic peptide

TisB	30	ECK1	ECO1 SBOY SETY	lexA-regulated toxic peptide
yncL	30	ECK1	ECO1 ECS8 SBOY SETY	hypothetical protein
ynhF	30	ECK1	ECO1 ECS8 SBOY SETY	hypothetical protein
ynhF	30	ECK1	ECO1 ECS8 SBOY SETY	hypothetical protein
azuC	29	ECK1	ECO1 SBOY	expressed protein
leuL	29	ECK1 ECS8 SBOY ECO1 SETY	YPAN	leu operon leader peptide
uof	29	ECK1	ECO1 SBOY	ryhB-regulated fur leader peptide
ydgU	28	ECK1	ECO1 ECS8 SBOY YPAN	hypothetical protein
yohP	28	ECK1	ECO1 ECS8 SBOY	expressed protein
shoB	27	ECK1	ECO1 SBOY	toxic membrane protein
yqeL	27	ECK1	ECO1	expressed protein
yrbN	27	ECK1	ECO1 ECS8 SBOY SETY	expressed protein
tnaC	25	ECK1 ECS8 ECO1	SBOY	tryptophanase leader peptide
yoaJ	25	ECK1	ECO1 ECS8 SBOY SETY	expressed protein, membrane-associated
ypdK	24	ECK1	ECO1 ECS8 SBOY SETY	expressed protein, membrane-associated
ypdK	24	ECK1	ECO1 ECS8 SBOY SETY	expressed protein, membrane-associated
yobI	22	ECK1	ECO1 ECS8 SBOY	expressed protein
yoeI	21	ECK1	ECO1 ECS8 SBOY SETY	expressed protein
yoeI	21	ECK1	ECO1 ECS8 SBOY SETY	expressed protein
ibsA	20	ECK1	ECO1 ECS8 SBOY SETY	toxic membrane protein
ibsC	20	ECK1	ECO1 ECS8 SBOY	toxic membrane protein
ibsD	20	ECK1	ECO1 ECS8 SBOY SETY	toxic membrane protein
ibsE	20	ECK1	ECO1 ECS8 SBOY SETY	toxic membrane protein
ypfM	20	ECK1	ECO1 ECS8 SBOY SETY	hypothetical protein
ibsB	19	ECK1	ECO1 ECS8 SBOY SETY	toxic membrane protein
yjeV	18	ECK1	ECO1 ECS8 SBOY SETY	expressed protein
hisL	17	ECK1	YPAN	his operon leader peptide
ilvX	17	ECK1	ECO1 ECS8 SBOY YPAN	expressed protein
pheM	15	ECK1 ECS8 ECO1 SETY	SBOY	phenylalanyl-tRNA synthetase operon leader peptide
trpL	15	ECK1 SBOY ECO1 ECS8	SETY	trp operon leader peptide

Table S3.2 Coordinates of potential unannotated *E. coli* K12 genes identified. Species containing annotated orthologs are listed using the species acronyms described for Table S2.1.

Neighbouring genes	Coordinates	Length	Annotated homologs	Nonconsensus start/overlap?
insH essD	complement(574981..576108)	375	ECS8	
yghD yghG	complement(3110076..3110942)	288	ECS8 YPAN VCHO XCAM	ATG start (TTG start in XCAM/VCHO)
cyaY yifL	3991873..3992358	161	ECO1 SBOY	210bp overlap
ybaZ ybaA	475499..475837	112	ECS8	97bp overlap
yahG yahI	338993..339313	106	ECO1	GTG start
ychE oppA	complement(1298626..1298940)	104	ECS8	TTG start
lysP yeiE	complement(2246538..2246846)	102	ECS8	GTG start 88bp overlap
ivbL tisB	3850998..3851288	96	ECO1	14bp overlap
yfgF yfgG	complement(2627177..2627467)	96	ECS8 SETY	156bp overlap
ydfU rem	complement(1642330..1642608)	92	ECO1	
gadE mdtE	complement(3656917..3657195	92	ECO1	
glxR ybbW	536720..536998	92	ECO1	142bp overlap
narI tpr	1285932..1286207	91	ECO1	
ldrD yhjV	complement(3698006..3698278)	90	ECS8	GTG start 105bp overlap
ligB gmk	complement(3819054..3819317)	87	EC81	TTG start 140bp overlap
bolA tig	453947..454210	87	ECS8	GTG start, 67 bp overlap
tolC ygiB	3177618..3177878	86	ECO1 SBOY	113bp overlap
sdhB sucA	757687..757947	86	SBOY ECO1	GTG start
yjiK yjiL	complement(4561691..4561948)	85	SBOY ECO1 ECKS8	4bp overlap
ppdD nadC	117577..117795	85	ECO1	TTG start 12 bp overlap
uof fldA	709914..710168	84	ECS8	GTG start, 35bp overlap
yqgC metK	complement(3084421..3084672)	83	ECO1	GTG start
potA pepT	complement(1184796..1185047)	83	ECS8	GTG start, 22bp overlap
yihG polA	complement(4044745..4044987)	80	ECO1	
ldrD yhjV	complement(3698006..3698245)	80	ECS8	GTG start 105bp overlap
yodB mtfA	2040945..2041187	80	ECO1	GTG start
wrbA ymdF	1067135..1067371	78	ECS8	TTG start, 68bp overlap
yfcJ fabB	2438084..2438305	73	ECO1	59bp overlap
narX narK	1276867..1277085	72	ECO1 ECS8	
ompA sulA	1019434..1019649	72	ECS8	TTG start, 17bp overlap
yhfA crp	3483920..3484135	71	ECS8	
ykgG ykgH	complement(323632..323844)	70	ECO1	46bp overlap
yciN topA	1328737..1328949	70	ECS8	TTG start
ybdR rnk	164537..164743	69	ECO1	TTG start
yoaE manX	complement(1899597..1899806)	69	ECO1	13bp overlap
putA putP	1078160..1078369	69	ECO1	
ygiF ygiM	complement(3199004..3199210)	68	ECS8	
hrpB mrcB	164537..164743	68	ECO1	TTG start 14bp overlap
yjgB insC	complement(4494307..4494513)	68	ECS8	TTG start
bamD raiA	2734935..2735141	68	ECO1 SBOY	TTG start
yfiF trxC	complement(2716540..2716743	67	ECS8	11bp overlap
yfiF trxC	complement(2716540..2716743)	67	ECS8	11bp overlap
gals yeiB	2239680..2239883	67	ECS8	11bp overlap
opgD ydcH	complement(1496456..1496659)	67	ECO1 ECS8 SBOY	GTG start, 80bp overlap
nth dtpA	1710310..1710510	66	ECO1	
kdpF ybfA	complement(727958..728158)	66	ECS8	87bp overlap
ykgG ykgH	complement(323751..323948)	65	ECS8	29bp overlap

ampH sbmA	395649..395843	64	ECS8	
yhjH kdgK	3677164..3677358	64	ECS8	47bp overlap
zupT rib	complement(3181403..3181597)	64	ECS8	TTG start
yjhU yjhF	complement(4518447..4518638)	63	ECS8	GTG start
yedQ yodC	complement(2025962..2026150)	62	ECS8	80bp overlap
ygfT ygfU	3029256..3029444	62	ECS8	GTG start 56bp overlap
mreB csrD	3399217..3399405	62	ECS8	GTG start
betT yahA	complement(331090..331275)	61	ECS8	
dinQ arsR	3645833..3646012	59	ECS8	GTG start 24bp overlap
ydiP ydiQ	complement(1777428..1777604)	58	ECS8	
folE yeiG	2241828..2242004	58	ECS8	73bp overlap
xylH xylR	3732836..3733012	58	ECS8	11bp overlap
ydaN dbpA	1407332..1407505	57	ECS8	
mlaA yfdC	complement(2463055..2463225)	56	ECS8	TTG start
yciK sohB	complement(1327180..1327344)	54	ECS8	TTG start
yhaC garK	3267685..3267849	54	SBOY	GTG start
ydjA sppA	1846754..1846918	54	ECS8	TTG start, 58bp overlap
ldrD yhjV	complement(3698275..3698436)	53	ECS8	TTG start
mhpR mhpA	367675..367833	52	PSYR	GTG start
coaA tufB	complement(4173236..4173391)	51	ECO1 SBOY SETY	
yciN topA	1328685..1328840	51	ECS8	TTG start, 8bp overlap
exbB metC	complement(3149999..3150154)	51	ECO1 SBOY SETY	GTG start 8bp overlap
yfgF yfgG	complement(2627142..2627294)	50	ECO1	
selD ydjA	complement(1845974..1846123)	49	ECS8	GTG start
yjiC iraD	4554907..4555050	47	ECS8	35bp overlap
opgC opgG	complement(1108209..1108352)	47	ECS8	TTG start
ypdI yfdY	complement(2492980..2493117)	45	ECS8	TTG start, 46bp overlap
yhgE pck	complement(3530537..3530668)	44	ECS8	GTG start
mutT yacG	complement(111564..111698)	44	SBOY	GTG start, 50bp overlap
yliL mntR	complement(852092..852220)	42	ECS8	72bp overlap
acs nrfA	4285571..4285690	39	SETY	
aspA fxsA	complement(4366386..4366502)	38	SETY	GTG start
trxA rho	3964254..3964355	33	SBOY ECO1 ECS8	
ybdR rnk	complement(642553..642741)	29	ECS8	

References

Achaz, G., F. Boyer, E.P. Rocha, A. Viari, and E. Coissac. 2007. Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics* **23:** 119-121.

Adkins, J.N., H.M. Mottaz, A.D. Norbeck, J.K. Gustin, J. Rue, T.R. Clauss, S.O. Purvine, K.D. Rodland, F. Heffron, and R.D. Smith. 2006. Analysis of the *Salmonella typhimurium* proteome through environmental response toward infectious conditions. *Mol Cell Proteomics* **5:** 1450-1461.

Aebersold, R. and M. Mann. 2003. Mass spectrometry-based proteomics. *Nature* **422:** 198-207.

Aggarwal, G., E.A. Worthey, P.D. McDonagh, and P.J. Myler. 2003. Importing statistical measures into Artemis enhances gene identification in the *Leishmania* genome project. *BMC Bioinformatics* **4:** 23.

Aivaliotis, M., K. Gevaert, M. Falb, A. Tebbe, K. Konstantinidis, B. Bisle, C. Klein, L. Martens, A. Staes, E. Timmerman et al. 2007. Large-scale identification of N-terminal peptides in the halophilic archaea *Halobacterium salinarum* and *Natronomonas pharaonis*. *J Proteome Res* **6:** 2195-2204.

Allen, J.E., M. Pertea, and S.L. Salzberg. 2004. Computational gene prediction using multiple sources of evidence. *Genome Res* **14:** 142-148.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Andersson, J.O. and S.G. Andersson. 2001. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Mol Biol Evol* **18:** 829-839.

Andersson, S.G. and C.G. Kurland. 1998. Reductive evolution of resident genomes. *Trends Microbiol* **6:** 263-268.

Ansong, C., S.O. Purvine, J.N. Adkins, M.S. Lipton, and R.D. Smith. 2008a. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Brief Funct Genomic Proteomic* **7:** 50-62.

Ansong, C., H. Yoon, A.D. Norbeck, J.K. Gustin, J.E. McDermott, H.M. Mottaz, J. Rue, J.N. Adkins, F. Heffron, and R.D. Smith. 2008b. Proteomics analysis of the causative agent of typhoid fever. *J Proteome Res* **7:** 546-557.

Apweiler, R., A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane et al. 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32:** D115-119.

Armengaud, J. 2009. A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr Opin Microbiol* **12:** 292-300.

Asakura, T., T. Sasaki, F. Nagano, A. Satoh, H. Obaishi, H. Nishioka, H. Imamura, K. Hotta, K. Tanaka, H. Nakanishi et al. 1998. Isolation and characterization of a novel actin filament-binding protein from *Saccharomyces cerevisiae*. *Oncogene* **16:** 121-130.

Atkins, J.F. and P.V. Baranov. 2010. The distinction between recoding and codon reassignment. *Genetics* **185:** 1535-1536.
Atkins, J.F., Gesteland, R.F. 2010. *Recoding: Expansion of Decoding Rules Enriches Gene Expression*. Springer.
Badger, J.H. and G.J. Olsen. 1999. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* **16:** 512-524.
Baertsch, R., M. Diekhans, W.J. Kent, D. Haussler, and J. Brosius. 2008. Retrocopy contributions to the evolution of the human genome. *BMC Genomics* **9:** 466.
Baird, S.D., M. Turcotte, R.G. Korneluk, and M. Holcik. 2006. Searching for IRES. *RNA* **12:** 1755-1785.
Bakke, P., N. Carney, W. Deloache, M. Gearing, K. Ingvorsen, M. Lotz, J. McNair, P. Penumetcha, S. Simpson, L. Voss et al. 2009. Evaluation of three automated genome annotations for *Halorhabdus utahensis*. *PLoS One* **4:** e6291.
Baranov, P.V., O. Fayet, R.W. Hendrix, and J.F. Atkins. 2006. Recoding in bacteriophages and bacterial IS elements. *Trends Genet* **22:** 174-181.
Baranov, P.V., R.F. Gesteland, and J.F. Atkins. 2004. P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA* **10:** 221-230.
Basrai, M.A., P. Hieter, and J.D. Boeke. 1997. Small open reading frames: beautiful needles in the haystack. *Genome Res* **7:** 768-771.
Belcourt, M.F. and P.J. Farabaugh. 1990. Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. *Cell* **62:** 339-352.
Belting, M., K. Mani, M. Jonsson, F. Cheng, S. Sandgren, S. Jonsson, K. Ding, J.G. Delcros, and L.A. Fransson. 2003. Glypican-1 is a vehicle for polyamine uptake in mammalian cells: a pivital role for nitrosothiol-derived nitric oxide. *J Biol Chem* **278:** 47181-47189.
Bennett, M.D., I.J. Leitch, H.J. Price, and J.S. Johnston. 2003. Comparisons with *Caenorhabditis* (approximately 100 Mb) and *Drosophila* (approximately 175 Mb) using flow cytometry show genome size in *Arabidopsis* to be approximately 157 Mb and thus approximately 25% larger than the *Arabidopsis* genome initiative estimate of approximately 125 Mb. *Ann Bot* **91:** 547-557.
Bennett, S. 2004. Solexa Ltd. *Pharmacogenomics* **5:** 433-438.
Bennett, S.T., C. Barnes, A. Cox, L. Davies, and C. Brown. 2005. Toward the 1,000 dollars human genome. *Pharmacogenomics* **6:** 373-382.
Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and E.W. Sayers. 2009. GenBank. *Nucleic Acids Res* **37:** D26-31.
Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and E.W. Sayers. 2011. GenBank. *Nucleic Acids Res* **39:** D32-37.
Bercovich, Z. and C. Kahana. 2004. Degradation of antizyme inhibitor, an ornithine decarboxylase homologous protein, is ubiquitin-dependent and is inhibited by antizyme. *J Biol Chem* **279:** 54097-54102.
Besemer, J. and M. Borodovsky. 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* **33:** W451-454.

Betran, E., W. Wang, L. Jin, and M. Long. 2002. Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. *Mol Biol Evol* **19:** 654-663.

Bevan, M. and S. Walsh. 2005. The *Arabidopsis* genome: a foundation for plant research. *Genome Res* **15:** 1632-1642.

Binns, N. and M. Masters. 2002. Expression of the *Escherichia coli* pcnB gene is translationally limited using an inefficient start codon: a second chromosomal example of translation initiated at AUU. *Mol Microbiol* **44:** 1287-1298.

Binstock, J.F., A. Pramanik, and H. Schulz. 1977. Isolation of a multi-enzyme complex of fatty acid oxidation from *Escherichia coli*. *Proc Natl Acad Sci U S A* **74:** 492-495.

Blattner, F.R., G. Plunkett, 3rd, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453-1462.

Blinkowa, A.L. and J.R. Walker. 1990. Programmed ribosomal frameshifting generates the *Escherichia coli* DNA polymerase III gamma subunit from within the tau subunit reading frame. *Nucleic Acids Res* **18:** 1725-1729.

Bocs, S., S. Cruveiller, D. Vallenet, G. Nuel, and C. Medigue. 2003. AMIGene: Annotation of MIcrobial Genes. *Nucleic Acids Res* **31:** 3723-3726.

Boeke, J.D., D.J. Garfinkel, C.A. Styles, and G.R. Fink. 1985. Ty elements transpose through an RNA intermediate. *Cell* **40:** 491-500.

Bork, P. and A. Bairoch. 1996. Go hunting in sequence databases but watch out for the traps. *Trends Genet* **12:** 425-427.

Bousquet, I., G. Dujardin, and P.P. Slonimski. 1991. ABC1, a novel yeast nuclear gene has a dual function in mitochondria: it suppresses a cytochrome b mRNA translation defect and is essential for the electron transfer in the bc 1 complex. *EMBO J* **10:** 2023-2031.

Brachat, S., F.S. Dietrich, S. Voegeli, Z. Zhang, L. Stuart, A. Lerch, K. Gates, T. Gaffney, and P. Philippsen. 2003. Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biol* **4:** R45.

Braun, B.R., M. van Het Hoog, C. d'Enfert, M. Martchenko, J. Dungan, A. Kuo, D.O. Inglis, M.A. Uhl, H. Hogues, M. Berriman et al. 2005. A human-curated annotation of the *Candida albicans* genome. *PLoS Genet* **1:** 36-57.

Brayman, T.G. and R.P. Hausinger. 1996. Purification, characterization, and functional analysis of a truncated *Klebsiella aerogenes* UreE urease accessory protein lacking the histidine-rich carboxyl terminus. *J Bacteriol* **178:** 5410-5416.

Brett, D., H. Pospisil, J. Valcarcel, J. Reich, and P. Bork. 2002. Alternative splicing and genome complexity. *Nat Genet* **30:** 29-30.

Brierley, I. 1995. Ribosomal frameshifting viral RNAs. *J Gen Virol* **76 (Pt 8):** 1885-1892.

Brown, C.M., P.A. Stockwell, C.N. Trotman, and W.P. Tate. 1990. Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. *Nucleic Acids Res* **18:** 6339-6345.

Brown, D. and K. Sjolander. 2006. Functional classification using phylogenomic inference. *PLoS Comput Biol* **2:** e77.

Brunner, E., C.H. Ahrens, S. Mohanty, H. Baetschmann, S. Loevenich, F. Potthast, E.W. Deutsch, C. Panse, U. de Lichtenberg, O. Rinner et al. 2007. A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat Biotechnol* **25:** 576-583.

Brushaber, K.R., G.A. O'Toole, and J.C. Escalante-Semerena. 1998. CobD, a novel enzyme with L-threonine-O-3-phosphate decarboxylase activity, is responsible for the synthesis of (R)-1-amino-2-propanol O-2-phosphate, a proposed new intermediate in cobalamin biosynthesis in *Salmonella typhimurium* LT2. *J Biol Chem* **273:** 2684-2691.

Bryson, K., V. Loux, R. Bossy, P. Nicolas, S. Chaillou, M. van de Guchte, S. Penaud, E. Maguin, M. Hoebeke, P. Bessieres et al. 2006. AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. *Nucleic Acids Res* **34:** 3533-3545.

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78-94.

Bustamante, C.D., R. Nielsen, and D.L. Hartl. 2002. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Mol Biol Evol* **19:** 110-117.

Butler, G., M.D. Rasmussen, M.F. Lin, M.A. Santos, S. Sakthikumar, C.A. Munro, E. Rheinbay, M. Grabherr, A. Forche, J.L. Reedy et al. 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* **459:** 657-662.

Butler, J.S., M. Springer, J. Dondon, M. Graffe, and M. Grunberg-Manago. 1986. *Escherichia coli* protein synthesis initiation factor IF3 controls its own gene expression at the translational level in vivo. *J Mol Biol* **192:** 767-780.

Byrne, K.P. and K.H. Wolfe. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15:** 1456-1461.

Byrne, K.P. and K.H. Wolfe. 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* **175:** 1341-1350.

Byrnes, J.K., G.P. Morris, and W.H. Li. 2006. Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. *Mol Biol Evol* **23:** 1136-1143.

Castellana, N.E., S.H. Payne, Z. Shen, M. Stanke, V. Bafna, and S.P. Briggs. 2008. Discovery and revision of *Arabidopsis* genes by proteogenomics. *Proc Natl Acad Sci U S A* **105:** 21034-21038.

Castellano, S., N. Morozova, M. Morey, M.J. Berry, F. Serras, M. Corominas, and R. Guigo. 2001. In silico identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep* **2:** 697-702.

Castellano, S., S.V. Novoselov, G.V. Kryukov, A. Lescure, E. Blanco, A. Krol, V.N. Gladyshev, and R. Guigo. 2004. Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution. *EMBO Rep* **5:** 71-77.

Chain, P.S., P. Hu, S.A. Malfatti, L. Radnedge, F. Larimer, L.M. Vergez, P. Worsham, M.C. Chu, and G.L. Andersen. 2006. Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. *J Bacteriol* **188:** 4453-4463.

Chen, G.T., M.J. Axley, J. Hacia, and M. Inouye. 1992. Overproduction of a selenocysteine-containing polypeptide in *Escherichia coli*: the fdhF gene product. *Mol Microbiol* **6:** 781-785.

Chen, P., S.K. Sapperstein, J.D. Choi, and S. Michaelis. 1997. Biogenesis of the *Saccharomyces cerevisiae* mating pheromone a-factor. *J Cell Biol* **136:** 251-269.

Cherry, J.M., C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R.K. Mortimer et al. 1997. Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* **387:** 67-73.

Childs, A.C., D.J. Mehta, and E.W. Gerner. 2003. Polyamine-dependent gene expression. *Cell Mol Life Sci* **60:** 1394-1406.

Cho, B.K., E.M. Knight, and B.O. Palsson. 2006. Transcriptional regulation of the fad regulon genes of *Escherichia coli* by ArcA. *Microbiology* **152:** 2207-2219.

Choudhary, J.S., W.P. Blackstock, D.M. Creasy, and J.S. Cottrell. 2001. Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1:** 651-667.

Chung, W.Y., S. Wadhawan, R. Szklarczyk, S.K. Pond, and A. Nekrutenko. 2007. A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput Biol* **3:** e91.

Clare, J. and P. Farabaugh. 1985. Nucleotide sequence of a yeast Ty element: evidence for an unusual mechanism of gene expression. *Proc Natl Acad Sci U S A* **82:** 2829-2833.

Clare, J.J., M. Belcourt, and P.J. Farabaugh. 1988. Efficient translational frameshifting occurs within a conserved sequence of the overlap between the two genes of a yeast Ty1 transposon. *Proc Natl Acad Sci U S A* **85:** 6816-6820.

Clarke, J., H.C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4:** 265-270.

Claros, M.G. and P. Vincens. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem* **241:** 779-786.

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

Cliften, P.F., L.W. Hillier, L. Fulton, T. Graves, T. Miner, W.R. Gish, R.H. Waterston, and M. Johnston. 2001. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res* **11:** 1175-1186.

Coffino, P. 2001. Regulation of cellular polyamines by antizyme. *Nat Rev Mol Cell Biol* **2:** 188-194.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, 3rd et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537-544.

Cole, S.T., K. Eiglmeier, J. Parkhill, K.D. James, N.R. Thomson, P.R. Wheeler, N. Honore, T. Garnier, C. Churcher, D. Harris et al. 2001. Massive gene decay in the leprosy bacillus. *Nature* **409:** 1007-1011.

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

Conant, G.C. and K.H. Wolfe. 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol Syst Biol* **3:** 129.

Craig, R. and R.C. Beavis. 2004. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20:** 1466-1467.

Curwen, V., E. Eyras, T.D. Andrews, L. Clarke, E. Mongin, S.M. Searle, and M. Clamp. 2004. The Ensembl automatic gene annotation system. *Genome Res* **14:** 942-950.

D'Silva, S., S.J. Haider, and E.M. Phizicky. 2011. A domain of the actin binding protein Abp140 is the yeast methyltransferase responsible for 3-methylcytidine modification in the tRNA anti-codon loop. *RNA* **17:** 1100-1110.

da Silva, A.C., J.A. Ferro, F.C. Reinach, C.S. Farah, L.R. Furlan, R.B. Quaggio, C.B. Monteiro-Vitorello, M.A. Van Sluys, N.F. Almeida, L.M. Alves et al. 2002. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* **417:** 459-463.

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

de Groot, A., R. Dulermo, P. Ortet, L. Blanchard, P. Guerin, B. Fernandez, B. Vacherie, C. Dossat, E. Jolivet, P. Siguier et al. 2009. Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium *Deinococcus deserti*. *PLoS Genet* **5:** e1000434.

Delcher, A.L., D. Harmon, S. Kasif, O. White, and S.L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27:** 4636-4641.

Deng, W., S.R. Liou, G. Plunkett, 3rd, G.F. Mayhew, D.J. Rose, V. Burland, V. Kodoyianni, D.C. Schwartz, and F.R. Blattner. 2003. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J Bacteriol* **185:** 2330-2337.

Dietrich, F.S., S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pohlmann, P. Luedi, S. Choi et al. 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304:** 304-307.

Dignard, D., A.L. El-Naggar, M.E. Logue, G. Butler, and M. Whiteway. 2007. Identification and characterization of MFA1, the gene encoding *Candida albicans* a-factor pheromone. *Eukaryot Cell* **6:** 487-494.

Dimitrov, L.N., R.B. Brem, L. Kruglyak, and D.E. Gottschling. 2009. Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of *Saccharomyces cerevisiae* S288C strains. *Genetics* **183:** 365-383.

Dinman, J.D. 1995. Ribosomal frameshifting in yeast viruses. *Yeast* **11:** 1115-1127.

Dinman, J.D., T. Icho, and R.B. Wickner. 1991. A -1 ribosomal frameshift in a double-stranded RNA virus of yeast forms a gag-pol fusion protein. *Proc Natl Acad Sci U S A* **88:** 174-178.

Dinman, J.D. and R.B. Wickner. 1992. Ribosomal frameshifting efficiency and gag/gag-pol ratio are critical for yeast M1 double-stranded RNA virus propagation. *J Virol* **66:** 3669-3676.

Drillon, G. and G. Fischer. 2011. Comparative study on synteny between yeasts and vertebrates. *C R Biol* **334:** 629-638.

Drinnenberg, I.A., D.E. Weinberg, K.T. Xie, J.P. Mower, K.H. Wolfe, G.R. Fink, and D.P. Bartel. 2009. RNAi in budding yeast. *Science* **326:** 544-550.

Droege, M. and B. Hill. 2008. The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol* **136:** 3-10.

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

Dujon, B. 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet* **22:** 375-387.

Dujon, B. 2010. Yeast evolutionary genomics. *Nat Rev Genet* **11:** 512-524.

Dujon, B., K. Albermann, M. Aldea, D. Alexandraki, W. Ansorge, J. Arino, V. Benes, C. Bohn, M. Bolotin-Fukuhara, R. Bordonné et al. 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XV. *Nature* **387 (Suppl.):** 98-102.

Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuveglise, E. Talla et al. 2004. Genome evolution in yeasts. *Nature* **430:** 35-44.

Eckardt, N.A. 2001. A sense of self: the role of DNA sequence elimination in allopolyploidization. *Plant Cell* **13:** 1699-1704.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792-1797.

Edman, P. 1949. A method for the determination of amino acid sequence in peptides. *Arch Biochem* **22:** 475.

Elgar, G. and T. Vavouri. 2008. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet* **24:** 344-352.

Elias, D.A., M.E. Monroe, M.J. Marshall, M.F. Romine, A.S. Belieav, J.K. Fredrickson, G.A. Anderson, R.D. Smith, and M.S. Lipton. 2005. Global detection and characterization of hypothetical proteins in *Shewanella oneidensis* MR-1 using LC-MS based proteomics. *Proteomics* **5:** 3120-3130.

Emanuelsson, O., H. Nielsen, S. Brunak, and G. von Heijne. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300:** 1005-1016.

Enault, F., K. Suhre, and J.M. Claverie. 2005. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. *BMC Bioinformatics* **6:** 247.

Eng JK, M.A., Yates JR. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5:** 976-989.

Eul, J., M. Graessmann, and A. Graessmann. 1995. Experimental evidence for RNA trans-splicing in mammalian cells. *EMBO J* **14:** 3226-3235.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8:** 186-194.

Farabaugh, P. 2010. Programmed Frameshifting in Budding Yeast. In *Recoding* (ed. J.F. Atkins), pp. 221-248. Springer.

Farabaugh, P.J. 1996. Programmed translational frameshifting. *Microbiol Rev* **60:** 103-134.

Farabaugh, P.J., E. Kramer, H. Vallabhaneni, and A. Raman. 2006. Evolution of +1 programmed frameshifting signals and frameshift-regulating tRNAs in the order Saccharomycetales. *J Mol Evol* **63:** 545-561.

Farabaugh, P.J., H. Zhao, and A. Vimaladithan. 1993. A novel programed frameshift expresses the POL3 gene of retrotransposon Ty3 of yeast: frameshifting without tRNA slippage. *Cell* **74:** 93-103.

Farrer, R.A., E. Kemen, J.D. Jones, and D.J. Studholme. 2009. De novo assembly of the *Pseudomonas syringae* pv. syringae B728a genome using Illumina/Solexa short sequence reads. *FEMS Microbiol Lett* **291:** 103-111.

Fay, J.C. and J.A. Benavides. 2005. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet* **1:** 66-71.

Feng, L., P.R. Reeves, R. Lan, Y. Ren, C. Gao, Z. Zhou, J. Cheng, W. Wang, J. Wang, W. Qian et al. 2008. A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PLoS One* **3:** e4053.

Fikes, J.D., V.A. Bankaitis, J.P. Ryan, and P.J. Bassford, Jr. 1987. Mutational alterations affecting the export competence of a truncated but fully functional maltose-binding protein signal peptide. *J Bacteriol* **169:** 2345-2351.

Finn, R.D., J. Clements, and S.R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39:** W29-37.

Fischer, G., E.P. Rocha, F. Brunet, M. Vergassola, and B. Dujon. 2006. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet* **2:** e32.

Fisk, D.G., C.A. Ball, K. Dolinski, S.R. Engel, E.L. Hong, L. Issel-Tarver, K. Schwartz, A. Sethuraman, D. Botstein, and J.M. Cherry. 2006. *Saccharomyces cerevisiae* S288C genome annotation: a working hypothesis. *Yeast* **23:** 857-865.

Fitzpatrick, D.A., M.E. Logue, J.E. Stajich, and G. Butler. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol* **6:** 99.

Fitzpatrick, D.A., P. O'Gaora, K.P. Byrne, and G. Butler. 2010. Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser. *BMC Genomics* **11:** 290.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496-512.

Foissac, S. and T. Schiex. 2005. Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics* **6:** 25.

Force, A., M. Lynch, F.B. Pickett, A. Amores, Y.L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531-1545.

Fountoulakis, M., M.F. Takacs, P. Berndt, H. Langen, and B. Takacs. 1999. Enrichment of low abundance proteins of *Escherichia coli* by hydroxyapatite chromatography. *Electrophoresis* **20:** 2181-2195.

Fozo, E.M., M.R. Hemm, and G. Storz. 2008. Small toxic proteins and the antisense RNAs that repress them. *Microbiol Mol Biol Rev* **72:** 579-589, Table of Contents.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397-403.

Friedberg, I. 2006. Automated protein function prediction--the genomic challenge. *Brief Bioinform* **7:** 225-242.

Frishman, D., A. Mironov, H.W. Mewes, and M. Gelfand. 1998. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* **26:** 2941-2947.

Fujita, Y., H. Matsuoka, and K. Hirooka. 2007. Regulation of fatty acid metabolism in bacteria. *Mol Microbiol* **66:** 829-839.

Galagan, J.E., S.E. Calvo, K.A. Borkovich, E.U. Selker, N.D. Read, D. Jaffe, W. FitzHugh, L.J. Ma, S. Smirnov, S. Purcell et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422:** 859-868.

Gallien, S., E. Perrodou, C. Carapito, C. Deshayes, J.M. Reyrat, A. Van Dorsselaer, O. Poch, C. Schaeffer, and O. Lecompte. 2009. Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Res* **19:** 128-135.

Gardner, P.P., J. Daub, J.G. Tate, E.P. Nawrocki, D.L. Kolbe, S. Lindgreen, A.C. Wilkinson, R.D. Finn, S. Griffiths-Jones, S.R. Eddy et al. 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res* **37:** D136-140.

Gerstein, A.C., H.J. Chun, A. Grant, and S.P. Otto. 2006. Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. *PLoS Genet* **2:** e145.

Gevaert, K., M. Goethals, L. Martens, J. Van Damme, A. Staes, G.R. Thomas, and J. Vandekerckhove. 2003. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol* **21:** 566-569.

Goffeau, A., B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston et al. 1996. Life with 6000 genes. *Science* **274:** 546, 563-547.

Goldstein, A.L. and J.H. McCusker. 2001. Development of *Saccharomyces cerevisiae* as a model pathogen. A system for the genetic identification of gene products required for survival in the mammalian host environment. *Genetics* **159:** 499-513.

Gordon, J.L., K.P. Byrne, and K.H. Wolfe. 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet* **5:** e1000485.

Gouy, M., S. Guindon, and O. Gascuel. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27:** 221-224.

Gouy, M., S. Guindon, and O. Gascuel. 2010 SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27:** 221-224.

Graur, D., Y. Shuali, and W.H. Li. 1989. Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J Mol Evol* **28:** 279-285.

Green, J.B., R.P. Lower, and J.P. Young. 2009. The NfeD protein family and its conserved gene neighbours throughout prokaryotes: functional implications for stomatin-like proteins. *J Mol Evol* **69:** 657-667.

Green, R.E., A.S. Malaspinas, J. Krause, A.W. Briggs, P.L. Johnson, C. Uhler, M. Meyer, J.M. Good, T. Maricic, U. Stenzel et al. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134:** 416-426.

Greig, D., R.H. Borts, E.J. Louis, and M. Travisano. 2002a. Epistasis and hybrid sterility in *Saccharomyces*. *Proc Biol Sci* **269:** 1167-1171.

Greig, D., E.J. Louis, R.H. Borts, and M. Travisano. 2002b. Hybrid speciation in experimental populations of yeast. *Science* **298:** 1773-1775.

Gromadski, K.B. and M.V. Rodnina. 2004. Kinetic determinants of high-fidelity tRNA discrimination on the ribosome. *Mol Cell* **13:** 191-200.

Gu, X., Z. Zhang, and W. Huang. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci U S A* **102:** 707-712.

Guda, C., E. Fahy, and S. Subramaniam. 2004. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* **20:** 1785-1794.

Guindon, S., F. Delsuc, J.F. Dufayard, and O. Gascuel. 2009. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol* **537:** 113-137.

Gupta, N., S. Tanner, N. Jaitly, J.N. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R.D. Smith et al. 2007. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res* **17:** 1362-1377.

Hammell, A.B., R.C. Taylor, S.W. Peltz, and J.D. Dinman. 1999. Identification of putative programmed -1 ribosomal frameshift signals in large DNA databases. *Genome Res* **9:** 417-427.

Han, M.J. and S.Y. Lee. 2006. The *Escherichia coli* proteome: past, present, and future prospects. *Microbiol Mol Biol Rev* **70:** 362-439.

Hansen, T.M., P.V. Baranov, I.P. Ivanov, R.F. Gesteland, and J.F. Atkins. 2003. Maintenance of the correct open reading frame by the ribosome. *EMBO Rep* **4:** 499-504.

Harris, T.D., P.R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. Dimeo, J.W. Efcavitch et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* **320:** 106-109.

Harrison, P., A. Kumar, N. Lan, N. Echols, M. Snyder, and M. Gerstein. 2002. A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J Mol Biol* **316:** 409-419.

Hayashi, T., K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C.G. Han, E. Ohtsubo, K. Nakayama, T. Murata et al. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* **8:** 11-22.

Heckman, D.S., D.M. Geiser, B.R. Eidell, R.L. Stauffer, N.L. Kardos, and S.B. Hedges. 2001. Molecular evidence for the early colonization of land by fungi and plants. *Science* **293:** 1129-1133.

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

Hemm, M.R., B.J. Paul, J. Miranda-Rios, A. Zhang, N. Soltanzad, and G. Storz. 2010 Small stress response proteins in *Escherichia coli*: proteins missed by classical proteomic studies. *J Bacteriol* **192:** 46-58.

Hemm, M.R., B.J. Paul, T.D. Schneider, G. Storz, and K.E. Rudd. 2008. Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol* **70:** 1487-1501.

Herr, A.J., R.F. Gesteland, and J.F. Atkins. 2000. One protein from two open reading frames: mechanism of a 50 nt translational bypass. *EMBO J* **19:** 2671-2680.

Hixson, K.K., J.N. Adkins, S.E. Baker, R.J. Moore, B.A. Chromy, R.D. Smith, S.L. McCutchen-Maloney, and M.S. Lipton. 2006. Biomarker candidate identification in *Yersinia pestis* using organism-wide semiquantitative proteomics. *J Proteome Res* **5:** 3008-3017.

Horner, D.S., G. Pavesi, T. Castrignano, P.D. De Meo, S. Liuni, M. Sammeth, E. Picardi, and G. Pesole. 2010. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinform* **11:** 181-197.

Horsfield, J.A., D.N. Wilson, S.A. Mannering, F.M. Adamski, and W.P. Tate. 1995. Prokaryotic ribosomes recode the HIV-1 gag-pol-1 frameshift sequence by an E/P site post-translocation simultaneous slippage mechanism. *Nucleic Acids Res* **23:** 1487-1494.

Howard, M.T., B.H. Shirts, J. Zhou, C.L. Carlson, S. Matsufuji, R.F. Gesteland, R.S. Weeks, and J.F. Atkins. 2001. Cell culture analysis of the regulatory frameshift event required for the expression of mammalian antizymes. *Genes Cells* **6:** 931-941.

Hughes, T.R., M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102:** 109-126.

Hurles, M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS Biol* **2:** E206.

IBM. 2009. *IBM research aims to build nanoscale DNA sequencer to help drive down cost of personalised genetic analysis*. IBM, New York.

Icho, T. and R.B. Wickner. 1989. The double-stranded RNA genome of yeast virus L-A encodes its own putative RNA polymerase by fusing two open reading frames. *J Biol Chem* **264:** 6716-6723.

Ingolia, N.T., L.F. Lareau, and J.S. Weissman. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147:** 789-802.

Ivanov, I.P., C.B. Anderson, R.F. Gesteland, and J.F. Atkins. 2004. Identification of a new antizyme mRNA +1 frameshifting stimulatory pseudoknot in a subset of diverse invertebrates and its apparent absence in intermediate species. *J Mol Biol* **339:** 495-504.

Ivanov, I.P., R.F. Gesteland, and J.F. Atkins. 1998a. A second mammalian antizyme: conservation of programmed ribosomal frameshifting. *Genomics* **52:** 119-129.

Ivanov, I.P., R.F. Gesteland, and J.F. Atkins. 2000a. Antizyme expression: a subversion of triplet decoding, which is remarkably conserved by evolution, is a sensor for an autoregulatory circuit. *Nucleic Acids Res* **28:** 3185-3196.

Ivanov, I.P., R.F. Gesteland, and J.F. Atkins. 2006. Evolutionary specialization of recoding: frameshifting in the expression of *S. cerevisiae* antizyme mRNA is via an atypical antizyme shift site but is still +1. *RNA* **12:** 332-337.

Ivanov, I.P., R.F. Gesteland, S. Matsufuji, and J.F. Atkins. 1998b. Programmed frameshifting in the synthesis of mammalian antizyme is +1 in mammals, predominantly +1 in fission yeast, but -2 in budding yeast. *RNA* **4:** 1230-1238.

Ivanov, I.P., S. Matsufuji, Y. Murakami, R.F. Gesteland, and J.F. Atkins. 2000b. Conservation of polyamine regulation by translational frameshifting from yeast to mammals. *EMBO J* **19:** 1907-1917.

Jacks, T., H.D. Madhani, F.R. Masiarz, and H.E. Varmus. 1988a. Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. *Cell* **55:** 447-458.

Jacks, T., M.D. Power, F.R. Masiarz, P.A. Luciw, P.J. Barr, and H.E. Varmus. 1988b. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* **331:** 280-283.

Jacks, T. and H.E. Varmus. 1985. Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. *Science* **230:** 1237-1242.

Jackson, A.P., J.A. Gamble, T. Yeomans, G.P. Moran, D. Saunders, D. Harris, M. Aslett, J.F. Barrell, G. Butler, F. Citiulo et al. 2009. Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. *Genome Res* **19:** 2231-2244.

Jacobs, J.L., A.T. Belew, R. Rakauskaite, and J.D. Dinman. 2007. Identification of functional, endogenous programmed -1 ribosomal frameshift signals in the genome of *Saccharomyces cerevisiae*. *Nucleic Acids Res* **35:** 165-174.

Jacq, C., J.R. Miller, and G.G. Brownlee. 1977. A pseudogene structure in 5S DNA of *Xenopus laevis*. *Cell* **12:** 109-120.

Jaffe, J.D., H.C. Berg, and G.M. Church. 2004a. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4:** 59-77.
Jaffe, J.D., N. Stange-Thomann, C. Smith, D. DeCaprio, S. Fisher, J. Butler, S. Calvo, T. Elkins, M.G. FitzGerald, N. Hafez et al. 2004b. The complete genome and proteome of *Mycoplasma mobile*. *Genome Res* **14:** 1447-1461.
Janetzky, B. and L. Lehle. 1992. Ty4, a new retrotransposon from *Saccharomyces cerevisiae*, flanked by tau-elements. *J Biol Chem* **267:** 19798-19805.
Jeffries, T.W., I.V. Grigoriev, J. Grimwood, J.M. Laplaza, A. Aerts, A. Salamov, J. Schmutz, E. Lindquist, P. Dehal, H. Shapiro et al. 2007. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nat Biotechnol* **25:** 319-326.
Jin, Q., Z. Yuan, J. Xu, Y. Wang, Y. Shen, W. Lu, J. Wang, H. Liu, J. Yang, F. Yang et al. 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res* **30:** 4432-4441.
Johansson, M.J., A. Esberg, B. Huang, G.R. Bjork, and A.S. Bystrom. 2008. Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. *Mol Cell Biol* **28:** 3301-3312.
Johnson, A., P. Gin, B.N. Marbois, E.J. Hsieh, M. Wu, M.H. Barros, C.F. Clarke, and A. Tzagoloff. 2005. COQ9, a new gene required for the biosynthesis of coenzyme Q in *Saccharomyces cerevisiae*. *J Biol Chem* **280:** 31397-31404.
Jong, A., Y. Yeh, and J.J. Ma. 1993. Characteristics, substrate analysis, and intracellular location of *Saccharomyces cerevisiae* UMP kinase. *Arch Biochem Biophys* **304:** 197-204.
Jukes, T.H. and S. Osawa. 1990. The genetic code in mitochondria and chloroplasts. *Experientia* **46:** 1117-1126.
Jungblut, P.R., E.C. Muller, J. Mattow, and S.H. Kaufmann. 2001. Proteomics reveals open reading frames in *Mycobacterium tuberculosis* H37Rv not predicted by genomics. *Infect Immun* **69:** 5905-5907.
Jungreis, I., M.F. Lin, R. Spokony, C.S. Chan, N. Negre, A. Victorsen, K.P. White, and M. Kellis. 2011. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Res* **21:** 2096-2113.
Kaessmann, H., N. Vinckenbosch, and M. Long. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10:** 19-31.
Kalume, D.E., S. Peri, R. Reddy, J. Zhong, M. Okulate, N. Kumar, and A. Pandey. 2005. Genome annotation of *Anopheles gambiae* using mass spectrometry-derived data. *BMC Genomics* **6:** 128.
Karolchik, D., R.M. Kuhn, R. Baertsch, G.P. Barber, H. Clawson, M. Diekhans, B. Giardine, R.A. Harte, A.S. Hinrichs, F. Hsu et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36:** D773-779.
Kastenmayer, J.P., L. Ni, A. Chu, L.E. Kitchen, W.C. Au, H. Yang, C.D. Carter, D. Wheeler, R.W. Davis, J.D. Boeke et al. 2006. Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res* **16:** 365-373.
Kawakami, K., S. Pande, B. Faiola, D.P. Moore, J.D. Boeke, P.J. Farabaugh, J.N. Strathern, Y. Nakamura, and D.J. Garfinkel. 1993. A rare tRNA-Arg(CCU)

that regulates Ty1 element ribosomal frameshifting is essential for Ty1 retrotransposition in *Saccharomyces cerevisiae*. *Genetics* **135:** 309-320.
Keasling, J.D. 2010. Manufacturing molecules through metabolic engineering. *Science* **330:** 1355-1358.
Kellis, M., B.W. Birren, and E.S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428:** 617-624.
Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241-254.
Kircher, M. and J. Kelso. 2010. High-throughput DNA sequencing--concepts and limitations. *Bioessays* **32:** 524-536.
Knop, M. 2011. Yeast cell morphology and sexual reproduction--a short overview and some considerations. *C R Biol* **334:** 599-606.
Kobayashi, N., T.K. McClanahan, J.R. Simon, J.M. Treger, and K. McEntee. 1996. Structure and functional analysis of the multistress response gene DDR2 from *Saccharomyces cerevisiae*. *Biochem Biophys Res Commun* **229:** 540-547.
Kolker, E., K.S. Makarova, S. Shabalina, A.F. Picone, S. Purvine, T. Holzman, T. Cherny, D. Armbruster, R.S. Munson, Jr., G. Kolesov et al. 2004. Identification and functional analysis of 'hypothetical' genes expressed in *Haemophilus influenzae*. *Nucleic Acids Res* **32:** 2353-2361.
Kolker, E., A.F. Picone, M.Y. Galperin, M.F. Romine, R. Higdon, K.S. Makarova, N. Kolker, G.A. Anderson, X. Qiu, K.J. Auberry et al. 2005. Global profiling of *Shewanella oneidensis* MR-1: expression of hypothetical genes and improved functional annotations. *Proc Natl Acad Sci U S A* **102:** 2099-2104.
Kontos, H., S. Napthine, and I. Brierley. 2001. Ribosomal pausing at a frameshifter RNA pseudoknot is sensitive to reading phase but shows little correlation with frameshift efficiency. *Mol Cell Biol* **21:** 8657-8670.
Korf, I., P. Flicek, D. Duan, and M.R. Brent. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1:** S140-148.
Korlach, J., K.P. Bjornson, B.P. Chaudhuri, R.L. Cicero, B.A. Flusberg, J.J. Gray, D. Holden, R. Saxena, J. Wegener, and S.W. Turner. Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol* **472:** 431-455.
Kozak, M. 2001. Constraints on reinitiation of translation in mammals. *Nucleic Acids Res* **29:** 5226-5232.
Kriventseva, E.V., I. Koch, R. Apweiler, M. Vingron, P. Bork, M.S. Gelfand, and S. Sunyaev. 2003. Increase of functional diversity by alternative splicing. *Trends Genet* **19:** 124-128.
Kryukov, G.V., S. Castellano, S.V. Novoselov, A.V. Lobanov, O. Zehtab, R. Guigo, and V.N. Gladyshev. 2003. Characterization of mammalian selenoproteomes. *Science* **300:** 1439-1443.
Kucerova, E., S.W. Clifton, X.Q. Xia, F. Long, S. Porwollik, L. Fulton, C. Fronick, P. Minx, K. Kyung, W. Warren et al. 2010. Genome sequence of *Cronobacter sakazakii* BAA-894 and comparative genomic hybridization analysis with other *Cronobacter* species. *PLoS One* **5:** e9556.

Kumar, K., V. Desai, L. Cheng, M. Khitrov, D. Grover, R.V. Satya, C. Yu, N. Zavaljevski, and J. Reifman. 2011. AGeS: a software system for microbial genome sequence annotation. *PLoS One* **6:** e17469.

Kurian, L., R. Palanimurugan, D. Godderz, and R.J. Dohmen. 2011. Polyamine sensing by nascent ornithine decarboxylase antizyme stimulates decoding of its mRNA. *Nature* **477:** 490-494.

Kuo, C.H., N.A. Moran, and H. Ochman. 2009. The consequences of genetic drift for bacterial genome complexity. *Genome Res* **19:** 1450-1454.

Kuo, C.H. and H. Ochman. 2009. Deletional bias across the three domains of life. *Genome Biol Evol* **1:** 145-152.

Kuo, C.H. and H. Ochman. 2010. The extinction dynamics of bacterial pseudogenes. *PLoS Genet* **6**.

Kurland, C.G. 1992. Translational accuracy and the fitness of bacteria. *Annu Rev Genet* **26:** 29-50.

Kurtzman, C.P. 2003. Phylogenetic circumscription of *Saccharomyces, Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Res* **4:** 233-245.

Kurtzman, C.P. 2011. Discussion of Teleomorphic and Anamorphic Ascomycetous Yeasts and Yeast-like Taxa. In *The Yeasts, a Taxonomic studiy* (eds. C.P. Kurtzman J.W. Fell, and T. Boekhout), pp. 293-307. Elsevier, Amsterdam.

Kurtzman, C.P. and C.J. Robnett. 2003. Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Res* **3:** 417-432.

Kuster, B., P. Mortensen, J.S. Andersen, and M. Mann. 2001. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1:** 641-650.

Lafontaine, I. and B. Dujon. 2010. Origin and fate of pseudogenes in Hemiascomycetes: a comparative analysis. *BMC Genomics* **11:** 260.

Lafontaine, I., G. Fischer, E. Talla, and B. Dujon. 2004. Gene relics in the genome of the yeast *Saccharomyces cerevisiae*. *Gene* **335:** 1-17.

Lagesen, K., P. Hallin, E.A. Rodland, H.H. Staerfeldt, T. Rognes, and D.W. Ussery. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35:** 3100-3108.

Lan, R. and P.R. Reeves. 2002. *Escherichia coli* in disguise: molecular origins of *Shigella*. *Microbes Infect* **4:** 1125-1132.

Langille, M.G., W.W. Hsiao, and F.S. Brinkman. 2008. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* **9:** 329.

Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23:** 2947-2948.

Larsen, B., R.F. Gesteland, and J.F. Atkins. 1997. Structural probing and mutagenic analysis of the stem-loop required for *Escherichia coli* dnaX ribosomal frameshifting: programmed efficiency of 50%. *J Mol Biol* **271:** 47-60.

Larsen, T.S. and A. Krogh. 2003. EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* **4:** 21.

Lawrence, J.G., R.W. Hendrix, and S. Casjens. 2001. Where are the pseudogenes in bacterial genomes? *Trends Microbiol* **9:** 535-540.

Lawrence, J.G. and J.R. Roth. 1995. The cobalamin (coenzyme B12) biosynthetic genes of *Escherichia coli*. *J Bacteriol* **177:** 6371-6380.

Lee, J., E.K. Mandell, T. Rao, D.S. Wuttke, and V. Lundblad. 2010. Investigating the role of the Est3 protein in yeast telomere replication. *Nucleic Acids Res* **38:** 2279-2290.

Lendvay, T.S., D.K. Morris, J. Sah, B. Balasubramanian, and V. Lundblad. 1996. Senescence mutants of *Saccharomyces cerevisiae* with a defect in telomere replication identify three additional EST genes. *Genetics* **144:** 1399-1412.

Lerat, E. and H. Ochman. 2004. Psi-Phi: exploring the outer limits of bacterial pseudogenes. *Genome Res* **14:** 2273-2278.

Lerat, E. and H. Ochman. 2005. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res* **33:** 3125-3132.

Li, W.H., T. Gojobori, and M. Nei. 1981. Pseudogenes as a paradigm of neutral evolution. *Nature* **292:** 237-239.

Li, W.H., C.I. Wu, and C.C. Luo. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* **2:** 150-174.

Liao, P.Y., P. Gupta, A.N. Petrov, J.D. Dinman, and K.H. Lee. 2008. A new kinetic model reveals the synergistic effect of E-, P- and A-sites on +1 ribosomal frameshifting. *Nucleic Acids Res* **36:** 2619-2629.

Link, A.J., L.G. Hays, E.B. Carmack, and J.R. Yates, 3rd. 1997a. Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis* **18:** 1314-1334.

Link, A.J., K. Robison, and G.M. Church. 1997b. Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* **18:** 1259-1313.

Lipton, M.S., L. Pasa-Tolic, G.A. Anderson, D.J. Anderson, D.L. Auberry, J.R. Battista, M.J. Daly, J. Fredrickson, K.K. Hixson, H. Kostandarithes et al. 2002. Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc Natl Acad Sci U S A* **99:** 11049-11054.

Lisman, Q., D. Urli-Stam, and J.C. Holthuis. 2004. HOR7, a multicopy suppressor of the Ca2+-induced growth defect in sphingolipid mannosyltransferase-deficient yeast. *J Biol Chem* **279:** 36390-36396.

Liti, G., D.M. Carter, A.M. Moses, J. Warringer, L. Parts, S.A. James, R.P. Davey, I.N. Roberts, A. Burt, V. Koufopanou et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458:** 337-341.

Liti, G. and J. Schacherer. 2011. The rise of yeast population genomics. *C R Biol* **334:** 612-619.

Liu, Y., P.M. Harrison, V. Kunin, and M. Gerstein. 2004. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol* **5:** R64.

Lonnerberg, P. and C.F. Ibanez. 1999. Novel, testis-specific mRNA transcripts encoding N-terminally truncated choline acetyltransferase. *Mol Reprod Dev* **53:** 274-281.

Lowe, T.M. and S.R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25:** 955-964.

Lukjancenko, O., T.M. Wassenaar, and D.W. Ussery. 2010. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol* **60:** 708-720.

Lynch, M. and A. Force. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459-473.

Maar, D., D. Liveris, J.K. Sussman, S. Ringquist, I. Moll, N. Heredia, A. Kil, U. Blasi, I. Schwartz, and R.W. Simons. 2008. A single mutation in the IF3 N-terminal domain perturbs the fidelity of translation initiation at three levels. *J Mol Biol* **383:** 937-944.

Maldonado, R. and A.J. Herr. 1998. Efficiency of T4 gene 60 translational bypassing. *J Bacteriol* **180:** 1822-1830.

Maltsev, N., E. Glass, D. Sulakhe, A. Rodriguez, M.H. Syed, T. Bompada, Y. Zhang, and M. D'Souza. 2006. PUMA2--grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res* **34:** D369-372.

Maniloff, J. 1996. The minimal cell genome: "on being the right size". *Proc Natl Acad Sci U S A* **93:** 10004-10006.

Mardis, E.R. 2006. Anticipating the 1,000 dollar genome. *Genome Biol* **7:** 112.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Markowitz, V.M., I.M. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, I. Anderson, A. Lykidis, K. Mavromatis et al. 2010. The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res* **38:** D382-390.

Marquez, V., D.N. Wilson, W.P. Tate, F. Triana-Alonso, and K.H. Nierhaus. 2004. Maintaining the ribosomal reading frame: the influence of the E site during translational regulation of release factor 2. *Cell* **118:** 45-55.

Martens, J.H., H. Barg, M.J. Warren, and D. Jahn. 2002. Microbial production of vitamin B12. *Appl Microbiol Biotechnol* **58:** 275-285.

Matsufuji, S., T. Matsufuji, Y. Miyazaki, Y. Murakami, J.F. Atkins, R.F. Gesteland, and S. Hayashi. 1995. Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. *Cell* **80:** 51-60.

Matsufuji, S., T. Matsufuji, N.M. Wills, R.F. Gesteland, and J.F. Atkins. 1996. Reading two bases twice: mammalian antizyme frameshifting in yeast. *EMBO J* **15:** 1360-1370.

Mayer, V.W. and A. Aguilera. 1990. High levels of chromosome instability in polyploids of *Saccharomyces cerevisiae*. *Mutat Res* **231:** 177-186.

McCutcheon, J.P. and S.R. Eddy. 2003. Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res* **31:** 4119-4128.

Medigue, C. and I. Moszer. 2007. Annotation, comparison and databases for hundreds of bacterial genomes. *Res Microbiol* **158:** 724-736.

Meinnel, T., C. Sacerdot, M. Graffe, S. Blanquet, and M. Springer. 1999. Discrimination by *Escherichia coli* initiation factor IF3 against initiation on non-canonical codons relies on complementarity rules. *J Mol Biol* **290:** 825-837.

Mejlhede, N., P. Licznar, M.F. Prere, N.M. Wills, R.F. Gesteland, J.F. Atkins, and O. Fayet. 2004. -1 frameshifting at a CGA AAG hexanucleotide site is required for transposition of insertion sequence IS1222. *J Bacteriol* **186:** 3274-3277.

Merico, A., P. Sulo, J. Piskur, and C. Compagno. 2007. Fermentative lifestyle in yeasts belonging to the Saccharomyces complex. *FEBS J* **274:** 976-989.

Merrill, C., L. Bayraktaroglu, A. Kusano, and B. Ganetzky. 1999. Truncated RanGAP encoded by the Segregation Distorter locus of Drosophila. *Science* **283:** 1742-1745.

Metzker, M.L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* **11:** 31-46.

Mighell, A.J., N.R. Smith, P.A. Robinson, and A.F. Markham. 2000. Vertebrate pseudogenes. *FEBS Lett* **468:** 109-114.

Mira, A., H. Ochman, and N.A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17:** 589-596.

Miranda, J.J., P. De Wulf, P.K. Sorger, and S.C. Harrison. 2005. The yeast DASH complex forms closed rings on microtubules. *Nat Struct Mol Biol* **12:** 138-143.

Miura, F., N. Kawaguchi, J. Sese, A. Toyoda, M. Hattori, S. Morishita, and T. Ito. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci U S A* **103:** 17846-17851.

Miura, K., Y. Tomioka, H. Suzuki, M. Yonezawa, T. Hishinuma, and M. Mizugaki. 1997. Molecular cloning of the nemA gene encoding N-ethylmaleimide reductase from *Escherichia coli*. *Biol Pharm Bull* **20:** 110-112.

Mokrejs, M., V. Vopalensky, O. Kolenaty, T. Masek, Z. Feketova, P. Sekyrova, B. Skaloudova, V. Kriz, and M. Pospisek. 2006. IRESite: the database of experimentally verified IRES structures (www.iresite.org). *Nucleic Acids Res* **34:** D125-130.

Morris, D.K. and V. Lundblad. 1997. Programmed translational frameshifting in a gene required for yeast telomere replication. *Curr Biol* **7:** 969-976.

Morrissy, A.S., R.D. Morin, A. Delaney, T. Zeng, H. McDonald, S. Jones, Y. Zhao, M. Hirst, and M.A. Marra. 2009. Next-generation tag sequencing for cancer gene expression profiling. *Genome Res* **19:** 1825-1835.

Murakami, Y., S. Matsufuji, T. Kameji, S. Hayashi, K. Igarashi, T. Tamura, K. Tanaka, and A. Ichihara. 1992. Ornithine decarboxylase is degraded by the 26S proteasome without ubiquitination. *Nature* **360:** 597-599.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344-1349.

Nakao, Y., T. Kanamori, T. Itoh, Y. Kodama, S. Rainieri, N. Nakamura, T. Shimonaga, M. Hattori, and T. Ashikari. 2009. Genome sequence of the lager brewing yeast, an interspecies hybrid. *DNA Res* **16:** 115-129.

Nakazawa, T., M. Yamaguchi, T.A. Okamura, E. Ando, O. Nishimura, and S. Tsunasawa. 2008. Terminal proteomics: N- and C-terminal analyses for high-fidelity identification of proteins using MS. *Proteomics* **8:** 673-685.

Namy, O., I. Hatin, and J.P. Rousset. 2001. Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep* **2:** 787-793.

Namy, O., S.J. Moran, D.I. Stuart, R.J. Gilbert, and I. Brierley. 2006. A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. *Nature* **441:** 244-247.

Namy, O., J.P. Rousset, S. Napthine, and I. Brierley. 2004. Reprogrammed genetic decoding in cellular gene expression. *Mol Cell* **13:** 157-168.

Nash, R., S. Weng, B. Hitz, R. Balakrishnan, K.R. Christie, M.C. Costanzo, S.S. Dwight, S.R. Engel, D.G. Fisk, J.E. Hirschman et al. 2007. Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res* **35:** D468-471.

Nie, H., F. Yang, X. Zhang, J. Yang, L. Chen, J. Wang, Z. Xiong, J. Peng, L. Sun, J. Dong et al. 2006. Complete genome sequence of *Shigella flexneri* 5b and comparison with *Shigella flexneri* 2a. *BMC Genomics* **7:** 173.

Nielsen, P. and A. Krogh. 2005. Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* **21:** 4322-4329.

Niewmierzycka, A. and S. Clarke. 1999. S-Adenosylmethionine-dependent methylation in *Saccharomyces cerevisiae*. Identification of a novel protein arginine methyltransferase. *J Biol Chem* **274:** 814-824.

Nilsson, A.I., S. Koskiniemi, S. Eriksson, E. Kugelberg, J.C. Hinton, and D.I. Andersson. 2005. Bacterial genome size reduction by experimental evolution. *Proc Natl Acad Sci U S A* **102:** 12112-12116.

Nirenberg, M., T. Caskey, R. Marshall, R. Brimacombe, D. Kellogg, B. Doctor, D. Hatfield, J. Levin, F. Rottman, S. Pestka et al. 1966. The RNA code and protein synthesis. *Cold Spring Harb Symp Quant Biol* **31:** 11-24.

Noma, A., S. Yi, T. Katoh, Y. Takai, and T. Suzuki. 2011 Actin-binding protein ABP140 is a methyltransferase for 3-methylcytidine at position 32 of tRNAs in *Saccharomyces cerevisiae*. *RNA* **17:** 1111-1119.

Notredame, C., D.G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302:** 205-217.

Ochman, H. 2002. Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. *Trends Genet* **18:** 335-337.

Ochman, H. and L.M. Davalos. 2006. The nature and dynamics of bacterial genomes. *Science* **311:** 1730-1733.

Ohama, T., T. Suzuki, M. Mori, S. Osawa, T. Ueda, K. Watanabe, and T. Nakase. 1993. Non-universal decoding of the leucine codon CUG in several Candida species. *Nucleic Acids Res* **21:** 4039-4045.

Oheigeartaigh, S.S., D. Armisen, K.P. Byrne, and K.H. Wolfe. 2011. Systematic discovery of unannotated genes in 11 yeast species using a database of orthologous genomic segments. *BMC Genomics* **12:** 377.

Ohno, S. 1970. *Evolution by gene duplication*, London.

Ohshima, K., M. Hattori, T. Yada, T. Gojobori, Y. Sakaki, and N. Okada. 2003. Whole-genome screening indicates a possible burst of formation of processed

pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol* **4:** R74.

Oliver, S.G., Q.J. van der Aart, M.L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J.P. Ballesta, P. Benit et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357:** 38-46.

Ongay-Larios, L., R. Navarro-Olmos, L. Kawasaki, N. Velazquez-Zavala, E. Sanchez-Paredes, F. Torres-Quiroz, G. Coello, and R. Coria. 2007. *Kluyveromyces lactis* sexual pheromones. Gene structures and cellular responses to alpha-factor. *FEMS Yeast Res* **7:** 740-747.

Osawa, S., T. Ohama, T.H. Jukes, K. Watanabe, and S. Yokoyama. 1989. Evolution of the mitochondrial genetic code. II. Reassignment of codon AUA from isoleucine to methionine. *J Mol Evol* **29:** 373-380.

Oshiro, G., L.M. Wodicka, M.P. Washburn, J.R. Yates, 3rd, D.J. Lockhart, and E.A. Winzeler. 2002. Parallel identification of new genes in *Saccharomyces cerevisiae*. *Genome Res* **12:** 1210-1220.

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

Pal, S., D.H. Park, and B.V. Plapp. 2009. Activity of yeast alcohol dehydrogenases on benzyl alcohols and benzaldehydes: characterization of ADH1 from *Saccharomyces carlsbergensi*s and transition state analysis. *Chem Biol Interact* **178:** 16-23.

Palanimurugan, R., H. Scheel, K. Hofmann, and R.J. Dohmen. 2004. Polyamines regulate their synthesis by inducing expression and blocking degradation of ODC antizyme. *EMBO J* **23:** 4857-4867.

Palleja, A., E.D. Harrington, and P. Bork. 2008. Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics* **9:** 335.

Pande, S., A. Vimaladithan, H. Zhao, and P.J. Farabaugh. 1995. Pulling the ribosome out of frame by +1 at a programmed frameshift site by cognate binding of aminoacyl-tRNA. *Mol Cell Biol* **15:** 298-304.

Park, J., S.A. Teichmann, T. Hubbard, and C. Chothia. 1997. Intermediate sequences increase the detection of homology between sequences. *J Mol Biol* **273:** 349-354.

Parra, G., P. Agarwal, J.F. Abril, T. Wiehe, J.W. Fickett, and R. Guigo. 2003. Comparative gene prediction in human and mouse. *Genome Res* **13:** 108-117.

Parra, G., E. Blanco, and R. Guigo. 2000. GeneID in *Drosophila*. *Genome Res* **10:** 511-515.

Parra, G., K. Bradnam, and I. Korf. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23:** 1061-1067.

Payne, S.H., S.T. Huang, and R. Pieper. 2010. A proteogenomic update to *Yersinia*: enhancing genome annotation. *BMC Genomics* **11:** 460.

Percudani, R., A. Pavesi, and S. Ottonello. 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J Mol Biol* **268:** 322-330.

Perkins, D.N., D.J. Pappin, D.M. Creasy, and J.S. Cottrell. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20:** 3551-3567.

Peterson, J.D., L.A. Umayam, T. Dickinson, E.K. Hickey, and O. White. 2001. The Comprehensive Microbial Resource. *Nucleic Acids Res* **29:** 123-125.

Petros, L.M., M.T. Howard, R.F. Gesteland, and J.F. Atkins. 2005. Polyamine sensing during antizyme mRNA programmed frameshifting. *Biochem Biophys Res Commun* **338:** 1478-1489.

Petrov, D.A., T.A. Sangster, J.S. Johnston, D.L. Hartl, and K.L. Shaw. 2000. Evidence for DNA loss as a determinant of genome size. *Science* **287:** 1060-1062.

Piskur, J., E. Rozpedowska, S. Polakova, A. Merico, and C. Compagno. 2006. How did *Saccharomyces* evolve to become a good brewer? *Trends Genet* **22:** 183-186.

Plant, E.P. and J.D. Dinman. 2008. The role of programmed-1 ribosomal frameshifting in coronavirus propagation. *Front Biosci* **13:** 4873-4881.

Plant, E.P., K.L. Jacobs, J.W. Harger, A. Meskauskas, J.L. Jacobs, J.L. Baxter, A.N. Petrov, and J.D. Dinman. 2003. The 9-A solution: how mRNA pseudoknots promote efficient programmed -1 ribosomal frameshifting. *RNA* **9:** 168-174.

Plant, E.P., P. Wang, J.L. Jacobs, and J.D. Dinman. 2004. A programmed -1 ribosomal frameshift signal can function as a cis-acting mRNA destabilizing element. *Nucleic Acids Res* **32:** 784-790.

Podlaha, O. and J. Zhang. 2009. Processed pseudogenes: the 'fossilized footprints' of past gene expression. *Trends Genet* **25:** 429-434.

Poptsova, M.S. and J.P. Gogarten. 2010. Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology* **156:** 1909-1917.

Porreca, G.H., J. Shendure, and G.M. Church. 2006. Polony DNA sequencing. In *Curr Protoc Mol Biol*.

Pramanik, A., S. Pawar, E. Antonian, and H. Schulz. 1979. Five different enzymatic activities are associated with the multienzyme complex of fatty acid oxidation from *Escherichia coli*. *J Bacteriol* **137:** 469-473.

Quinlan, A.R., D.A. Stewart, M.P. Stromberg, and G.T. Marth. 2008. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* **5:** 179-181.

Randerath, E., R.C. Gupta, L.L. Chia, S.H. Chang, and K. Randerath. 1979. Yeast tRNA Leu UAG. Purification, properties and determination of the nucleotide sequence by radioactive derivative methods. *Eur J Biochem* **93:** 79-94.

Rato, C., S.R. Amirova, D.G. Bates, I. Stansfield, and H.M. Wallace. 2011. Translational recoding as a feedback controller: systems approaches reveal polyamine-specific effects on the antizyme ribosomal frameshift. *Nucleic Acids Res* **39:** 4587-4597.

Raux, E., A. Lanois, F. Levillayer, M.J. Warren, E. Brody, A. Rambach, and C. Thermes. 1996. *Salmonella typhimurium* cobalamin (vitamin B12) biosynthetic genes: functional studies in *S. typhimurium* and *Escherichia coli*. *J Bacteriol* **178:** 753-767.

Reed, J.L., I. Famili, I. Thiele, and B.O. Palsson. 2006. Towards multidimensional genome annotation. *Nat Rev Genet* **7:** 130-141.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16:** 276-277.

Riley, M., T. Abe, M.B. Arnaud, M.K. Berlyn, F.R. Blattner, R.R. Chaudhuri, J.D. Glasner, T. Horiuchi, I.M. Keseler, T. Kosuge et al. 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res* **34:** 1-9.

Rison, S.C., J. Mattow, P.R. Jungblut, and N.G. Stoker. 2007. Experimental determination of translational starts using peptide mass mapping and tandem mass spectrometry within the proteome of *Mycobacterium tuberculosis*. *Microbiology* **153:** 521-528.

Ro, D.K., E.M. Paradise, M. Ouellet, K.J. Fisher, K.L. Newman, J.M. Ndungu, K.A. Ho, R.A. Eachus, T.S. Ham, J. Kirby et al. 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440:** 940-943.

Rodnina, M.V., K.B. Gromadski, U. Kothe, and H.J. Wieden. 2005. Recognition and selection of tRNA in translation. *FEBS Lett* **579:** 938-942.

Rogozin, I.B., K.S. Makarova, D.A. Natale, A.N. Spiridonov, R.L. Tatusov, Y.I. Wolf, J. Yin, and E.V. Koonin. 2002. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res* **30:** 4264-4271.

Rowland, S.L., W.F. Burkholder, K.A. Cunningham, M.W. Maciejewski, A.D. Grossman, and G.F. King. 2004. Structure and mechanism of action of Sda, an inhibitor of the histidine kinases that regulate initiation of sporulation in *Bacillus subtilis*. *Mol Cell* **13:** 689-701.

Royce, T.E., J.S. Rozowsky, P. Bertone, M. Samanta, V. Stolc, S. Weissman, M. Snyder, and M. Gerstein. 2005. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet* **21:** 466-475.

Rudd, K.E. 2000. EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res* **28:** 60-64.

Rudd, K.E., I. Humphery-Smith, V.C. Wasinger, and A. Bairoch. 1998. Low molecular weight proteins: a challenge for post-genomic research. *Electrophoresis* **19:** 536-544.

Ruiz i Altaba, A. 1999. Gli proteins encode context-dependent positive and negative functions: implications for development and disease. *Development* **126:** 3205-3216.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.A. Rajandream, and B. Barrell. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16:** 944-945.

Sakai, H., K.O. Koyanagi, T. Imanishi, T. Itoh, and T. Gojobori. 2007. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. *Gene* **389:** 196-203.

Sakata, K., K. Kashiwagi, and K. Igarashi. 2000. Properties of a polyamine transporter regulated by antizyme. *Biochem J* **347 Pt 1:** 297-303.

Salzberg, S.L. 2007. Genome re-annotation: a wiki solution? *Genome Biol* **8:** 102.

Samayoa, J., F. Yildiz, and K. Karplus. 2011. Identification of Prokaryotic Small Proteins using a Comparative Genomic Approach. *Bioinformatics*.

Sanger, F., G.M. Air, B.G. Barrell, N.L. Brown, A.R. Coulson, C.A. Fiddes, C.A. Hutchison, P.M. Slocombe, and M. Smith. 1977a. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265:** 687-695.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977b. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74:** 5463-5467.

Sankoff, D., C. Zheng, and Q. Zhu. 2011. The collapse of gene complement following whole genome duplication. *BMC Genomics* **11:** 313.

Scannell, D.R., K.P. Byrne, J.L. Gordon, S. Wong, and K.H. Wolfe. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440:** 341-345.

Scannell, D.R., A.C. Frank, G.C. Conant, K.P. Byrne, M. Woolfit, and K.H. Wolfe. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci U S A* **104:** 8397-8402.

Schacherer, J., D.M. Ruderfer, D. Gresham, K. Dolinski, D. Botstein, and L. Kruglyak. 2007. Genome-wide analysis of nucleotide-level variation in commonly used *Saccharomyces cerevisiae* strains. *PLoS One* **2:** e322.

Schacherer, J., J.A. Shapiro, D.M. Ruderfer, and L. Kruglyak. 2009. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458:** 342-345.

Schneider, J., J. Blom, S. Jaenicke, B. Linke, K. Brinkrolf, H. Neuweger, A. Tauch, and A. Goesmann. 2010. RAPYD--rapid annotation platform for yeast data. *J Biotechnol* **155:** 118-126.

Sevinsky, J.R., B.J. Cargile, M.K. Bunger, F. Meng, N.A. Yates, R.C. Hendrickson, and J.L. Stephenson, Jr. 2008. Whole genome searching with shotgun proteomic data: applications for genome annotation. *J Proteome Res* **7:** 80-88.

Shah, A.A., M.C. Giddings, J.B. Parvaz, R.F. Gesteland, J.F. Atkins, and I.P. Ivanov. 2002. Computational identification of putative programmed translational frameshift sites. *Bioinformatics* **18:** 1046-1053.

Shao, M., P. Liu, J. Zhang, and R. Adzic. 2007. Origin of enhanced activity in palladium alloy electrocatalysts for oxygen reduction reaction. *J Phys Chem B* **111:** 6772-6775.

Skovgaard, M., L.J. Jensen, S. Brunak, D. Ussery, and A. Krogh. 2001. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet* **17:** 425-428.

Small, I., N. Peeters, F. Legeai, and C. Lurin. 2004. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* **4:** 1581-1590.

Smogorzewska, A. and T. de Lange. 2004. Regulation of telomerase by telomeric proteins. *Annu Rev Biochem* **73:** 177-208.

Souciet, J.L., B. Dujon, C. Gaillardin, M. Johnston, P.V. Baret, P. Cliften, D.J. Sherman, J. Weissenbach, E. Westhof, P. Wincker et al. 2009. Comparative genomics of protoploid Saccharomycetaceae. *Genome Res* **19:** 1696-1709.

Spingola, M., L. Grate, D. Haussler, and M. Ares, Jr. 1999. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* **5:** 221-234.

Stahl, G., S. Ben Salem, Z. Li, G. McCarty, A. Raman, M. Shah, and P.J. Farabaugh. 2001. Programmed +1 translational frameshifting in the yeast *Saccharomyces cerevisiae* results from disruption of translational error correction. *Cold Spring Harb Symp Quant Biol* **66:** 249-258.

Stahl, G., G.P. McCarty, and P.J. Farabaugh. 2002. Ribosome structure: revisiting the connection between translational accuracy and unconventional decoding. *Trends Biochem Sci* **27:** 178-183.

Stamm, S., S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T.A. Thanaraj, and H. Soreq. 2005. Function of alternative splicing. *Gene* **344:** 1-20.

Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24:** 637-644.

Stoesser, G., P. Sterk, M.A. Tuli, P.J. Stoehr, and G.N. Cameron. 1997. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **25:** 7-14.

Stothard, P. and D.S. Wishart. 2006. Automated bacterial genome analysis and annotation. *Curr Opin Microbiol* **9:** 505-510.

Struhl, G., K. Fitzgerald, and I. Greenwald. 1993. Intrinsic activity of the Lin-12 and Notch intracellular domains in vivo. *Cell* **74:** 331-345.

Stucka, R., C. Schwarzlose, H. Lochmuller, U. Hacker, and H. Feldmann. 1992. Molecular analysis of the yeast Ty4 element: homology with Ty1, copia, and plant retrotransposons. *Gene* **122:** 119-128.

Studer, R.A. and M. Robinson-Rechavi. 2009. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet* **25:** 210-216.

Sugino, R.P. and H. Innan. 2006. Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends Genet* **22:** 642-644.

Sugita, T. and T. Nakase. 1999. Nonuniversal usage of the leucine CUG codon in yeasts: Investigation of basidiomycetous yeast. *J Gen Appl Microbiol* **45:** 193-197.

Sundararajan, A., W.A. Michaud, Q. Qian, G. Stahl, and P.J. Farabaugh. 1999. Near-cognate peptidyl-tRNAs promote +1 programmed translational frameshifting in yeast. *Mol Cell* **4:** 1005-1015.

Taliaferro, D. and P.J. Farabaugh. 2007. An mRNA sequence derived from the yeast EST3 gene stimulates programmed +1 translational frameshifting. *RNA* **13:** 606-613.

Tam, O.H., A.A. Aravin, P. Stein, A. Girard, E.P. Murchison, S. Cheloufi, E. Hodges, M. Anger, R. Sachidanandam, R.M. Schultz et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453:** 534-538.

Tanner, S., Z. Shen, J. Ng, L. Florea, R. Guigo, S.P. Briggs, and V. Bafna. 2007. Improving gene annotation using peptide mass spectrometry. *Genome Res* **17:** 231-239.

Tate, W.P., J.B. Mansell, S.A. Mannering, J.H. Irvine, L.L. Major, and D.N. Wilson. 1999. UGA: a dual signal for 'stop' and for recoding in protein synthesis. *Biochemistry (Mosc)* **64:** 1342-1353.

Tateno, Y., T. Imanishi, S. Miyazaki, K. Fukami-Kobayashi, N. Saitou, H. Sugawara, and T. Gojobori. 2002. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res* **30:** 27-30.

Taylor, J.W. and M.L. Berbee. 2006. Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia* **98:** 838-849.

Theis, C., J. Reeder, and R. Giegerich. 2008. KnotInFrame: prediction of -1 ribosomal frameshift events. *Nucleic Acids Res* **36:** 6013-6020.

Thieme, F., R. Koebnik, T. Bekel, C. Berger, J. Boch, D. Buttner, C. Caldana, L. Gaigalat, A. Goesmann, S. Kay et al. 2005. Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium *Xanthomonas campestris* pv. vesicatoria revealed by the complete genome sequence. *J Bacteriol* **187:** 7254-7266.

Thompson, R.C. 1988. EFTu provides an internal kinetic standard for translational accuracy. *Trends Biochem Sci* **13:** 91-93.

Toh, H., B.L. Weiss, S.A. Perkin, A. Yamashita, K. Oshima, M. Hattori, and S. Aksoy. 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res* **16:** 149-156.

Torrents, D., M. Suyama, E. Zdobnov, and P. Bork. 2003. A genome-wide survey of human pseudogenes. *Genome Res* **13:** 2559-2567.

Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* **5:** e1000344.

Tsuchihashi, Z. 1991. Translational frameshifting in the *Escherichia coli* dnaX gene in vitro. *Nucleic Acids Res* **19:** 2457-2462.

Umezawa, Y., T. Shimada, A. Kori, K. Yamada, and A. Ishihama. 2008. The uncharacterized transcription factor YdhM is the regulator of the nemA gene, encoding N-ethylmaleimide reductase. *J Bacteriol* **190:** 5890-5897.

Vallenet, D., S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli et al. 2009. MicroScope: a platform for microbial genome annotation and comparative genomics. *Database (Oxford)* **2009:** bap021.

Vallenet, D., L. Labarre, Z. Rouy, V. Barbe, S. Bocs, S. Cruveiller, A. Lajus, G. Pascal, C. Scarpelli, and C. Medigue. 2006. MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* **34:** 53-65.

Van Domselaar, G.H., P. Stothard, S. Shrivastava, J.A. Cruz, A. Guo, X. Dong, P. Lu, D. Szafron, R. Greiner, and D.S. Wishart. 2005. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res* **33:** W455-459.

Velculescu, V.E., L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett, Jr., P. Hieter, B. Vogelstein, and K.W. Kinzler. 1997. Characterization of the yeast transcriptome. *Cell* **88:** 243-251.

Vimaladithan, A. and P.J. Farabaugh. 1994. Special peptidyl-tRNA molecules can promote translational frameshifting without slippage. *Mol Cell Biol* **14:** 8107-8116.

Vorholter, F.J., S. Schneiker, A. Goesmann, L. Krause, T. Bekel, O. Kaiser, B. Linke, T. Patschkowski, C. Ruckert, J. Schmid et al. 2008. The genome of *Xanthomonas campestris* pv. campestris B100 and its use for the reconstruction of metabolic pathways involved in xanthan biosynthesis. *J Biotechnol* **134:** 33-45.

Wagner, A. 2005. Energy constraints on the evolution of gene expression. *Mol Biol Evol* **22:** 1365-1374.

Wallace, H.M., A.V. Fraser, and A. Hughes. 2003. A perspective of polyamine metabolism. *Biochem J* **376:** 1-14.

Wallace, I.M., O. O'Sullivan, D.G. Higgins, and C. Notredame. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* **34:** 1692-1699.

Wang, H., M.J. Moore, P.S. Soltis, C.D. Bell, S.F. Brockington, R. Alexandre, C.C. Davis, M. Latvis, S.R. Manchester, and D.E. Soltis. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc Natl Acad Sci U S A* **106:** 3853-3858.

Wang, R., J.T. Prince, and E.M. Marcotte. 2005. Mass spectrometry of the M. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Res* **15:** 1118-1126.

Wang, W., J. Zhang, C. Alvarez, A. Llopart, and M. Long. 2000. The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in *Drosophila melanogaster*. *Mol Biol Evol* **17:** 1294-1301.

Wang, Y., M.S. Goligorsky, M. Lin, J.N. Wilcox, and P.A. Marsden. 1997. A novel, testis-specific mRNA transcript encoding an NH2-terminal truncated nitric-oxide synthase. *J Biol Chem* **272:** 11392-11401.

Washburn, M.P., D. Wolters, and J.R. Yates, 3rd. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19:** 242-247.

Watanabe, T., Y. Totoki, A. Toyoda, M. Kaneda, S. Kuramochi-Miyagawa, Y. Obata, H. Chiba, Y. Kohara, T. Kono, T. Nakano et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453:** 539-543.

Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Wei, C. and M.R. Brent. 2006. Using ESTs to improve the accuracy of de novo gene prediction. *BMC Bioinformatics* **7:** 327.

Weil, D., M.A. Power, G.C. Webb, and C.L. Li. 1997. Antisense transcription of a murine FGFR-3 psuedogene during fetal developement. *Gene* **187:** 115-122.

Weissenbach, J., G. Dirheimer, R. Falcoff, J. Sanceau, and E. Falcoff. 1977. Yeast tRNALeu (anticodon U--A--G) translates all six leucine codons in extracts from interferon treated cells. *FEBS Lett* **82:** 71-76.

Wheeler, D.L., T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. Dicuccio, R. Edgar, S. Federhen et al. 2008. Database resources

of the National Center for Biotechnology Information. *Nucleic Acids Res* **36:** D13-21.
Wicker, T., E. Schlagenhauf, A. Graner, T.J. Close, B. Keller, and N. Stein. 2006. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7:** 275.
Williams, R.E. and N.C. Bruce. 2002. 'New uses for an Old Enzyme'--the Old Yellow Enzyme family of flavoenzymes. *Microbiology* **148:** 1607-1614.
Wolfe, K. 2000. Robustness--it's not where you think it is. *Nat Genet* **25:** 3-4.
Wolfe, K. 2004. Evolutionary genomics: Yeasts accelerate beyond BLAST. *Curr Biol* **14:** R392-R394.
Wolfe, K.H. 2006. Comparative genomics and genome evolution in yeasts. *Philos Trans R Soc Lond B Biol Sci* **361:** 403-412.
Wolfe, K.H. and D.C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708-713.
Wolinsky, H. 2007. The thousand-dollar genome. Genetic brinkmanship or personalized medicine? *EMBO Rep* **8:** 900-903.
Won-Ki Huh, J.V.F., Luke C. Gerke, Adam S. Carroll, Russell W. Howson, Jonathan S. Weissman & Erin K. O'Shea. 2003. Global analysis of protein localization in budding yeast. *Nature* **VOL 425:** 686-691.
Wootton, J.C. and S. Federhen. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266:** 554-571.
Xu, H. and J.D. Boeke. 1990. Host genes that influence transposition in yeast: the abundance of a rare tRNA regulates Ty1 transposition frequency. *Proc Natl Acad Sci U S A* **87:** 8360-8364.
Xu, J., R.W. Hendrix, and R.L. Duda. 2004. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol Cell* **16:** 11-21.
Xu, Z., W. Wei, J. Gagneur, F. Perocchi, S. Clauder-Munster, J. Camblong, E. Guffanti, F. Stutz, W. Huber, and L.M. Steinmetz. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457:** 1033-1037.
Yang, F., J. Yang, X. Zhang, L. Chen, Y. Jiang, Y. Yan, X. Tang, J. Wang, Z. Xiong, J. Dong et al. 2005. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res* **33:** 6445-6458.
Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586-1591.
Yang, Z. and J.P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15:** 496-503.
Yassour, M., T. Kaplan, H.B. Fraser, J.Z. Levin, J. Pfiffner, X. Adiconis, G. Schroth, S. Luo, I. Khrebtukova, A. Gnirke et al. 2009. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci U S A* **106:** 3264-3269.
Yates, J.R., 3rd, J.K. Eng, and A.L. McCormack. 1995. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem* **67:** 3202-3210.
Yusupova, G.Z., M.M. Yusupov, J.H. Cate, and H.F. Noller. 2001. The path of messenger RNA through the ribosome. *Cell* **106:** 233-241.

Zhang, M., C.M. Pickart, and P. Coffino. 2003a. Determinants of proteasome recognition of ornithine decarboxylase, a ubiquitin-independent substrate. *EMBO J* **22:** 1488-1496.

Zhang, Y., D.A. Rodionov, M.S. Gelfand, and V.N. Gladyshev. 2009. Comparative genomic analyses of nickel, cobalt and vitamin B12 utilization. *BMC Genomics* **10:** 78.

Zhang, Z., N. Carriero, and M. Gerstein. 2004. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet* **20:** 62-67.

Zhang, Z. and M. Gerstein. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res* **31:** 5338-5348.

Zhang, Z. and M. Gerstein. 2004. Large-scale analysis of pseudogenes in the human genome. *Curr Opin Genet Dev* **14:** 328-335.

Zhang, Z., P.M. Harrison, Y. Liu, and M. Gerstein. 2003b. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **13:** 2541-2558.

Zheng, C., Q. Zhu, Z. Adam, and D. Sankoff. 2008. Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes. *Bioinformatics* **24:** i96-104.

Zhou, B.S., D.R. Beidler, and Y.C. Cheng. 1992. Identification of antisense RNA transcripts from a human DNA topoisomerase I pseudogene. *Cancer Res* **52:** 4280-4285.

Zhu, H.Q., G.Q. Hu, Z.Q. Ouyang, J. Wang, and Z.S. She. 2004. Accuracy improvement for identifying translation initiation sites in microbial genomes. *Bioinformatics* **20:** 3308-3317.

Appendix II

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

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

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

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

Note – This Appendix has been accepted for publication by the Proceedings of the National Academy of Science. My role in this study was in the editing and correction of sequence data for the new Saccharomycetaceae family species that were included in this study (soon to be publicly available on YGOB). Also, SearchDOGS was included as an annotation step in the Yeast Genome Annotation Pipeline (YGAP) that was used to annotate these genomes.

Abstract

We investigate yeast sex chromosome evolution by comparing genome sequences from 16 species in the family Saccharomycetaceae, including new data from the genera *Tetrapisispora*, *Kazachstania*, *Naumovozyma* and *Torulaspora*. We show that although most yeast species contain a mating-type (*MAT*) locus and silent *HML* and *HMR* loci that are structurally analogous to those of *Saccharomyces cerevisiae*, their detailed organization is highly variable and indicates that the *MAT* locus is a deletion hotspot. Over evolutionary time, chromosomal genes located immediately beside *MAT* have continually been deleted, truncated, or transposed to other places in the genome in a process that is gradually shortening the distance between *MAT* and *HML*. Each time a gene beside *MAT* is removed by deletion or transposition, the next gene

on the chromosome is brought into proximity with *MAT* and is in turn put at risk of removal. This process has also continually replaced the triplicated sequence regions, called Z and X, that allow *HML* and *HMR* to be used as templates for DNA repair at *MAT* during mating-type switching. We propose that the deletion and transposition events result from evolutionary accidents during mating-type switching, combined with natural selection to keep *MAT* and *HML* on the same chromosome. The rate of deletion accelerated greatly after the whole-genome duplication, probably because genes were redundant and could be deleted without requiring transposition. Our results show that mating-type switching imposes a significant mutational cost on the genome, one that must be outweighed by the evolutionary advantages of being able to switch.

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