TWO BIOINFORMATIC STUDIES OF GENOME
DYNAMICS IN BUDDING YEAST SPECIES.

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

by
David B. H. Barton

Department of Genetics
University of Leicester

SEPTEMBER 2006
Abstract

Two Bioinformatic Studies Of Genome Dynamics In Budding Yeast Species. David B. H. Barton.

The genome of the common baking and brewing yeast *Saccharomyces cerevisiae* continues to be the most completely sequenced and annotated of any eukaryotic organism. With this invaluable resource as a reference, several groups (Cliften *et al.* 2003; Kellis *et al.* 2003; Dujon *et al.* 2004; and others) have begun to sequence, assemble and analyse the genomes of various relatives of *S. cerevisiae*. Such analyses have led to more accurate annotation of the reference genome, clearer evidence for which genes are genuine and which seem dubious, and a wealth of information on regulatory motifs. The amount of sequence data is considerable, and many more interesting observations are likely to emerge from secondary studies.

This work describes two such discoveries. It has been determined that the regions immediately surrounding budding yeast centromeres are excessively divergent between the species, and highly polymorphic between strains. This occurs despite the obvious importance of centromeres in chromosome dynamics, and despite the absence of lengthy repetitious DNA that is characteristic of centromeres in other eukaryotes. As such it is a most unexpected finding, and although the mechanisms responsible for this 'supervariation' remain unclear it is sufficient to serve as a fingerprint for readily distinguishing many almost identical strains.

It has also been established that a sizeable subtelomeric section of the *S. cerevisiae* genome has at some point in the past been horizontally transmitted (probably by hybrid introgression) to European strains of a wild relative *Saccharomyces paradoxus*. This is part of a growing body of evidence that there is extensive gene flow even between so-called species and that budding yeast genomes have a more complex and multi-layered history than previously anticipated.
Foreword for the Lay Reader

This work concerns the genomes of budding yeasts. These modest single-celled fungal organisms are of intense interest for a number of reasons. They are industrially important because of their use in baking and brewing. The group is also of medical and agricultural importance: for example the yeast Candida albicans is a human pathogen, the cause of thrush, while Eremothecium gossypii causes diseases in plants. Moreover the budding yeasts are of general scientific importance: they have been the subjects of detailed studied since Pasteur's work in the 1860s, and have provided unparalleled insight into the workings of complex (eukaryotic) cells.

A genome is the complete DNA blueprint of an organism, written in the four-letter alphabet of the genetic code (G, C, A and T). The genome of the common baking and brewing yeast Saccharomyces cerevisiae was the first complex genome to be read—or 'sequenced'—thoroughly, and it remains the most complete. Now an increasing number of closely related genomes, those of various 'cousin species' of S. cerevisiae, have also been sequenced. This enables us to study the evolution of a complex genome: the molecular genealogy of an entire family of organisms.

The comparison of these related blueprints—a relatively new scientific field dubbed comparative genomics—creates exciting new opportunities. We find clues about the purpose and importance of regions of DNA that have previously remained mysterious (functionally important patterns in DNA tend to be conserved by evolution, and when they are not it is of equal interest). We can identify other regions that are particularly variable and may therefore be linked to diversity. We can better understand the processes that drive
genomes to diverge through generations, that create a new species from an old one. And by 'practising' on the relatively small genomes of yeasts we can perhaps develop tools, techniques and insights that may subsequently be useful for tackling the bigger challenges of understanding human evolution and human disease.

Bioinformaticians—scientists who use computers to find patterns in biological data—are often talented computer programmers with a secondary interest in biology. They can also be (as is the case with this author) biologists first and foremost, with information technology a complementary but secondary focus. This particular comparative study is characterised throughout by a very human mode of investigation. Computers are invaluable tools for sifting through the vast amounts of information that comprise even a simple genome, and the number of large scale computer-driven analyses in the literature is increasing every year. Nevertheless, such approaches can easily overlook interesting details that it would take a biological, and biologically-trained, mind to recognise as significant. There can be little doubt that over time the field of computation will increasingly penetrate and empower the field of biology.

The bulk of this thesis describes two novel studies, each of which is a synthesis of these two somewhat different disciplines. Of necessity, this work must stray a little from the conventional format of a biological dissertation if it is to aspire towards clarity. To that end, it will begin with a general introduction of relevance to both studies, with each individual study then following a more traditional format, referring back to the general introduction when necessary.
Declaration

The research detailed in this thesis was conducted by the author at the Department of Genetics in the University of Leicester. All the work presented as such is the author’s own and has not been submitted for any other degree in these or any other institutions.


This thesis is copyright material and no quotation from it may be published without proper acknowledgement.
Acknowledgements

My deepest thanks go to my undergraduate and PhD supervisor Professor Ed J. Louis, for encouraging my transition into scientific research, for granting me extraordinary freedom to pursue those areas that proved most interesting to me (even though the work has diverged significantly from what was originally intended!) and for his faith and patience throughout.

Many thanks are also owed to all those within Ed's lab (past and present) and the wider scientific community who have helped, cajoled and engaged in discussion on the work here described, in particular Gianni Liti, and also Duncan Greig, Duccio Cavalieri, Douda Bensasson, Dianna Fisk, Raymond Dalgleish and Julian Ketley.

My thanks also to those who have supported me in other ways over recent years; my dutiful parents Nicholas and Margaret Barton, my beautiful partner Helen Speechley (supportive way beyond the call of duty), my dear friend and former colleague Carolyn Deeming, and all the many others who have kept me more or less on the straight and narrow throughout this strange (and not always easy) educational transformation.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbreviations</td>
<td>9</td>
</tr>
<tr>
<td>Summary of Figures</td>
<td>10</td>
</tr>
<tr>
<td>Summary of Tables</td>
<td>11</td>
</tr>
<tr>
<td><strong>CHAPTER 1</strong> GENERAL INTRODUCTION</td>
<td>12</td>
</tr>
<tr>
<td>1.1 Yeasts and Genome Projects</td>
<td>13</td>
</tr>
<tr>
<td>1.2 The Computation of Biological Similarity</td>
<td>21</td>
</tr>
<tr>
<td>1.3 The Evolution of DNA Sequences</td>
<td>27</td>
</tr>
<tr>
<td>1.4 Centromere Biology</td>
<td>35</td>
</tr>
<tr>
<td>1.5 Telomere Biology</td>
<td>42</td>
</tr>
<tr>
<td><strong>CHAPTER 2</strong> MATERIALS AND METHODS</td>
<td>55</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>56</td>
</tr>
<tr>
<td>2.2 Sequence and Feature Data</td>
<td>56</td>
</tr>
<tr>
<td>2.3 Additional Data</td>
<td>59</td>
</tr>
<tr>
<td>2.4 Main Data Formats</td>
<td>61</td>
</tr>
<tr>
<td>2.5 Third-Party Programs</td>
<td>66</td>
</tr>
<tr>
<td>2.6 Genome-Wide Analysis Scripts</td>
<td>73</td>
</tr>
<tr>
<td>2.7 Miscellaneous Work</td>
<td>88</td>
</tr>
<tr>
<td><strong>CHAPTER 3</strong> AUXILIARY AND PREPARATORY WORK</td>
<td>95</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>96</td>
</tr>
<tr>
<td>3.2 The Origins of Y' Elements</td>
<td>96</td>
</tr>
<tr>
<td>3.3 Analysis of Overlapping Genes</td>
<td>104</td>
</tr>
<tr>
<td>3.4 Annotation of Telomeric Elements in SGD</td>
<td>111</td>
</tr>
</tbody>
</table>
CHAPTER 4  CENTROMERE-ADJACENT SUPERVARIATION  113
WITHIN THE BUDDING YEASTS

4.1  Introduction  114
4.2  Results  114
4.3  Discussion  134

CHAPTER 5  HYBRID INTROGRESSION OF S. CEREVISIAE  140
SEQUENCE INTO EUROPEAN S. PARADOXUS

5.1  Introduction  141
5.2  Results  141
5.3  Discussion  149

CHAPTER 6  CONCLUSIONS  156

Appendix I  CD-ROM Contents  163
Appendix II  Publications and Presentations Arising  165
References  167
Abbreviations

ALT  Alternative Lengthening of Telomeres
ARS  Autonomous Replication Sequence
ASCII American Standard Code for Information Interchange
BER  Base Excision Repair
BIR  Break-Induced Repair
CDE  Centromeric DNA Element
DOS  Disk Operating System
FTP  File Transfer Protocol
GER  General Excision Repair (=NER)
HR   Homologous Recombination
ITS  Internal Transcribed Spacer
LTR  Long Terminal Repeat
MMR  Mismatch Repair
NCBI National Center for Biotechnology Information
NER  Nucleotide Excision Repair (=GER)
NHEJ Non-Homologous End-Joining
ORC  Origin Recognition Complex
ORF  Open Reading Frame
SAGE Serial Analysis of Gene Expression
SGD  Saccharomyces Genome Database
SQL  Structured Query Language
SSA  Single-Strand Annealing
TAS  Telomere-Associated Sequence
TCR  Transcription-Coupled Repair
# Summary of Figures

1.1 Principles of dot matrix comparison. 23

1.2 Centromeric DNA across yeast species. 37

1.3 Specific and general structure of chromosome ends in *S. cerevisiae* S288C. 45

2.1 The ubiquitous minimal "Hello World" program in Java and Python. 75

4.1 Sample dotplot showing divergence of CEN1 across four closely related *Saccharomyces* species. 119

4.2 Sample dotplot showing divergence of CEN7 across four closely related *Saccharomyces* species, including an additional *S. cerevisiae* strain and two additional *S. paradoxus* strains. 120

4.3 Scatterplot showing number of mismatches vs total length for each region compared between *S. cerevisiae* and *S. paradoxus*. 122

4.4 Bar plot showing the distribution of P-distances for each category of region in each of the six pairwise comparisons. 123

4.5 Correlation between composition and divergence between *S. cerevisiae* and *S. paradoxus*. 124

4.6 Variability of centromeric and adjacent DNA compared to other intergenic regions. 126

4.7 Variation of centromere-binding proteins. 129

4.8 Phylogenetic tree of *Saccharomyces* strains based on intergenic sequence spanning CEN7. 131

4.9 Intergenic sequence containing CEN7 in multiple *S. paradoxus* strains and *S. cerevisiae* S288c. 133

5.1 Scatterplot of gene and intergenic P-distances between *S. cerevisiae* S288C and *S. paradoxus* CBS432 along the length of chromosome XIV. 143

5.2 Plot of identity between *S. cerevisiae* S288C and *S. paradoxus* CBS432 in the left subtelomere of chr XIV. 144

5.3 Phylogenies of *sensu stricto* yeasts, one based on sequence from the left subtelomere of chromosome XIV, and one based on sequence covering CEN7. 146

5.4 Possible order of subtelomeric duplication and hybridisation events. 154

6.1 The challenge of true 'whole genome phylogenies'. 159
## Summary of Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Yeast strains used in this study.</td>
<td>90</td>
</tr>
<tr>
<td>3.1</td>
<td>Incidence of overlapping ORFs in the <em>S. cerevisiae</em> genome according to SGD on 1/7/03.</td>
<td>106</td>
</tr>
<tr>
<td>3.2</td>
<td>SAGE expression data for overlapping ORFs.</td>
<td>108</td>
</tr>
<tr>
<td>3.3</td>
<td>Incidence of overlapping ORFs in the <em>S. cerevisiae</em> genome according to SGD on 10/9/06.</td>
<td>110</td>
</tr>
<tr>
<td>3.4</td>
<td>Telomeric annotations provided to SGD in 2003.</td>
<td>111</td>
</tr>
<tr>
<td>4.1</td>
<td>Intergenic regions containing centromeres in <em>S. cerevisiae</em>.</td>
<td>115</td>
</tr>
<tr>
<td>5.1</td>
<td>Genes wholly or partially within the cross-species transfer region.</td>
<td>147</td>
</tr>
</tbody>
</table>
CHAPTER 1: GENERAL INTRODUCTION
The work described in this thesis is unusual in several ways. First, although it concerns the biology of genome dynamics, it is almost entirely computer-based research, in which the raw materials are electronic files of genome sequence data and the methods are either pre-existing or custom-built computer programs.

Secondly, the work has been very exploratory in nature: casual observations leading to more detailed analyses to check their veracity and implications.

Thirdly, the research encompasses many different areas of biology. This introduction is intended as a review of those different fields of study, but because of the broad scope of the overall work, no individual section of this introduction can be as comprehensive as the author might like, lest the thesis be ten times the size and the key emphases of the work be lost in a mire of incidental detail.

1.1 Yeasts and Genome Projects

Yeasts are single-celled fungi, although the cells of some species can form filamentous aggregates. They are complex eukaryotic cells, with much the same kind of internal compartmentalisation and basic biochemistry as human cells, but the cells are much smaller and often significantly easier to culture.

Yeast are a polyphyletic group: the majority belong to the fungal division Ascomycota in that following meiosis they produce ascospores.
surrounded by an ascus. Some unicellular fungi that are commonly (perhaps misleadingly) classified as yeasts do not belong to this group: *Phaffia rhodozyma*, for example, is part of the Basidiomycota (producing ballistospores in a basidium; Miller *et al.* 1976).

Ascomycota is a broad and diverse division that also includes multicellular fungi such as truffles, the *Neurospora* and *Penicillium* bread moulds, and many of the fungal symbiotes found in lichens. The numerous ‘true yeast’ species in this group are also very diverse in form, behaviour, habitat and biochemistry.

1.1.1. *Saccharomyces cerevisiae* and the *sensu stricto* yeasts.

By far the best studied yeast (indeed the most thoroughly studied eukaryotic organism at this time) is *Saccharomyces cerevisiae*, the common baking and brewing yeast. This organism has been domesticated by humanity for thousands of years, but is also commonly found in the wild, as are a number of closely related species. They are typically found on the leaves, fruit or bark of trees, or in the soil below, or on organisms that might pass through those environments such as the *Drosophila* fruit flies.

*Saccharomyces paradoxus, Saccharomyces mikatae, Saccharomyces kudriavzevii* and *Saccharomyces bayanus* are, in that order, the closest relatives of *S. cerevisiae*. All are members of the ‘*sensu stricto*’ group of *Saccharomyces*: they are *Saccharomyces* yeasts ‘in the strictest sense’ (Vaughan-Martini *et al.* 1993). This group includes other species too, such as the hybrid *S. pastorianus* (a hybrid of *S. cerevisiae* and *S. bayanus* used in lager production and also known as *Saccharomyces carlsbergensis*). There
are numerous cases of yeasts originally defined as distinct species of the 
sensu stricto group, such as *Saccharomyces boulardii*, *Saccharomyces cario
canus*, *Saccharomyces douglasii* and *Saccharomyces mangini*, but later 
recognised as merely subspecies of the above species.

1.1.2. *Sensu lato* yeasts.

Beyond the *sensu stricto* group, a number of other yeasts comprise the 
informal *Saccharomyces* ‘*sensu lato*’ (*in the broad sense*) group. *Candida
glabrata* is a natural commensal organism in human mucosal tissues, 
normally non-pathogenic except in cases of immunodeficiency. Though
classified as *Candida* on the basis of asexuality, molecular studies have 
established beyond doubt that it is much more closely related to *S. cerevisiae*
than it is to *Candida albicans* (described below; Fidel *et al.* 1999).

*Saccharomyces servazzii* and *Saccharomyces unisporus* are wild 
yeasts with a limited repertoire of sugar fermentation (Vaughan-Martini *et al.*
1993; Casaregola *et al.* 2000). *Saccharomyces exiguus* is a wild species used
to make San Francisco sourdough bread and native to that area in soil and
sewage. *Saccharomyces dairenensis* and *Saccharomyces castellii* are also 
found in the wild and are phenotypically indistinguishable although their DNA
sequence is only 80% identical (Vaughan-Martini & Barcaccia 1996).

1.1.3. Pre-‘whole genome duplication’ species.

All the above species have genomes of a similar size to *S. cerevisiae*,
and it is now firmly established that their common ancestor underwent a whole

Other more distant relatives of *S. cerevisiae* generally have smaller genomes and are the descendents of an ancestor which predated this event. *Kluyveromyces lactis* can metabolise lactose and is used to produce rennet for cheese production (Bolotin-Fukuhara et al. 2000). *Saccharomyces kluyveri* is a wild strain found in Europe and North America: like many others it is found on trees, in soil and on *Drosophila* (Neuveglise et al. 2000). *Kluyveromyces waltii* was isolated from trees in Japan (Kodama & Kyono 1974) and recently rose to prominence as an excellent example of a pre-duplication genome (Kellis et al. 2004). *Eremothecium gossypii*, also known as *Ashbya gossypii*, is an insect-transmitted plant pathogen of cotton and fruit, thought to have the smallest genome of any known free-living eukaryote (Dietrich et al. 2004).

More distantly still, *Pichia angusta* (also known as *Hansenula polymorpha*) is a thermotolerant methylotrophic yeast used in the pharmaceutical industry to produce recombinant proteins (Blandin et al. 2000). *Debaryomyces hansenii* is a marine yeast often found in fish and resistant to cold and high salt levels, which like *K. lactis* metabolises lactose and is used in cheese production (Lepingle et al. 2000). *Candida albicans*, habitually diploid, is the causative agent of thrush (and more serious infections in immunodeficient patients) and thus is of significant medical interest (Jones et al. 2004). *Yarrowia lipolytica* can metabolise alkanes, is often found in food, and is used industrially to produce citric acid and various proteins (Beckerich et al. 1998).
Whereas most of the hemiascomycetous yeasts described above are under normal conditions true 'budding yeasts' (the daughter cells of which bud and then grow from different sites on the periphery) *Candida albicans* and *Yarrowia lipolytica* are dimorphic yeasts, which can either bud normally, or form germ tubes or pseudohyphae, and sometimes true hyphae (Herrero *et al.* 1999).

A more distant archaeascomycetous relative, *Schizosaccharomyces pombe*, grows by a different pattern again. The rod-shaped cells of this 'fission yeast' grow longitudinally before dividing into two daughter cells of roughly equal size. This yeast was originally extracted from East African millet beer and is now a popular model organism, particularly for cell cycle studies (*e.g.* Hayles *et al.* 1986).

1.1.4. Yeast phylogenies.

The above yeasts, a tiny but representative fraction of the many species known, are mentioned in approximate order of decreasing relatedness to *S. cerevisiae*. The relationships between some of them are shown as part of Figure 1.2 in Section 1.4 (p.37). It is difficult to attach reliable dates to these relationships, but *S. cerevisiae* and *S. pombe* are reckoned to have diverged some 420 to 330 million years ago (Sipiczki 2000) and the whole genome duplication event is estimated to have occurred 150 to 100 million years ago (Sugino & Innan 2005).

Molecular studies have led to enormous revisions in the taxonomy and phylogeny of these species, and molecular characteristics are evidently much more reliable indicators of yeast phylogeny than phenotypic characteristics.
Phylogenies have for example been based on 18S rRNA sequence (James et al. 1997), on the internal transcribed spacers (ITSs) of the 35S rRNA precursor (McCullough et al. 1998) and on karyotypes (Petersen et al. 1999). Since agreement between different molecular methods is not total, particularly amongst the branches of the sensu lato group, there may yet be new revisions to come, particularly in the light of various past, present and future genome projects that provide a wealth of new sequence information for this group.

1.1.5. Yeast genome projects.

*S. cerevisiae* has been the subject of intense scientific scrutiny since Louis Pasteur’s work on fermentation in the 1860s, and Ojvind Winge’s pioneering genetic work in the 1930s and 40s (Mortimer & Johnston 1986). This became even more the case when in 1996 the lab strain S288C became the first eukaryote to be fully sequenced (Goffau et al. 1996; 1997), thanks to the International Collaboration for Yeast Genome Sequencing. It remains the most completely sequenced (including as it does almost all telomeric sequence) and the most thoroughly annotated genome in the world.

The availability of this invaluable reference sequence, the wide range of other yeast species in collections around the world, and the relatively small size of their (eukaryotic) genomes are factors behind a massive and ongoing effort to sequence and compare the genomes of many other species and strains in the group.

The Génolevures consortium of French laboratories led the charge with partial genomic sequencing of 13 hemiascomycetous yeasts: *S. bayanus*

In a similar way, but focussing on a narrower branch of the *Saccharomyces* lineage in order to study functional non-coding sequences through comparative genomics, Washington University undertook low-coverage sequencing of *S. paradoxus* N17, CBS432 and UFRJ50791 (the latter then considered *S. cariocanus*), *S. mikatae* IFO1815, *S. bayanus* 623-6C and CBS380, *S. servazzii* CBS4311, *S. unisporus* CBS398, *S. exigus* CBS379, *S. dairenensis* CBS421, *S. castellii* NRRLY12630 and *S. kluyveri* NRRLY12651 (Cliften et al. 2001). This was later followed by more thorough sequencing of *S. mikatae* IFO1815, *S. kudriavzevii* IFO1802, *S. bayanus* 623-6C, *S. castellii* NRRLY12630, and *S. kluyveri* NRRLY12651 (Cliften et al. 2003).

At the same time the Broad Institute, focussing on a still narrower range of *S. cerevisiae* relatives and also interested primarily in regulatory regions, were undertaking their own genome projects for *S. paradoxus* CBS432, *S. mikatae* IFO1816 (although initially this was mistakenly believed to be the type strain IFO1815; Ed Louis, personal communication) and *S. bayanus* MCYC623 (Kellis et al. 2003), and later for *K. waltii* NCYC2644 (Kellis et al. 2004).
These latter studies resulted in significant revisions to the annotations for the reference genome *S. cerevisiae* (as described in Chapter 3), and produced contigs of a significant size covering a large proportion of the genomes. However the sequence still contained (and still contains) gaps and significant omissions (of telomeric sequence, for example).

Génolevures later completed sequencing, to the chromosome level, of *C. glabrata* and several of its earlier species subjects (*K. lactis, D. hansenii* and *Y. lipolytica; Dujon et al. 2004*). At roughly the same time, the genome of *E. gossypii* ATCC10895 was completed to the same degree (*Dietrich et al. 2004*), and the incomplete diploid genome of *C. albicans* SC5314 was published (*Jones et al. 2004*). The *S. pombe* genome was completed somewhat earlier (*Wood et al. 2002*), but again with incomplete telomere sequence.

Many additional projects are also underway, but there is already an astonishing quantity of sequence available. This presents an unprecedented opportunity, and challenge, for the field of bioinformatics, which is tasked with making sense of all that data, and finding patterns of similarity within it.
1.2: The Computation of Biological Similarity

"Convert the word WOOL into the word COAT, changing one letter at a time and creating a new proper word at each step."

Common ‘word-transform’ puzzle.

The field of bioinformatics is largely dependent on measures of sequence similarity; a more difficult concept to define than many might suppose. In comparing simple words, one could consider the meaning, pronunciation, or even etymology. Similarity in spelling is thankfully a more tractable problem.

In 1950, three years before Watson and Crick described the structure of DNA, information theory pioneer Richard Hamming, who was working on error-detecting and error-correcting codes for telecommunications, defined the simplest measure of similarity between strings (Hamming 1950). The Hamming distance is the minimum number of symbol substitutions required to transform one string into another of the same length. The Hamming distance between WOOL and COAT is 3. Unlike the classic puzzle there is no requirement for each intermediate to be a valid word.

In 1965, Vladimir Levenshtein reconsidered the problem in a slightly broader manner (Levenshtein 1965). The Levenshtein distance, or ‘edit distance’, is another measure of the similarity between two strings, but the strings can be of different lengths, and at each step a symbol can be substituted (as in Hamming’s method) but also inserted or deleted. Insertions and deletions are treated as a single ‘indel’ operation, since the process is
bidirectional and a symbol could be either inserted into one string or deleted from the other. The edit distance between WOOL and COAT is the same as the Hamming distance: 3. The edit distance between WOOL and COATS is 4, because of an extra indel step: the insertion of an S at the end of COAT, or the deletion of an S from WOOL.

The Hamming and edit distances, and variations upon them (for example one might introduce extra operations, such as the exchange of two neighbouring symbols) have a huge number of applications in the world of computing: not only in specialist roles such as error-correcting and cryptography but also in everyday ones such as spell-checkers and search engines.

The similarity between two strings can also be gauged visually using a 'dot matrix diagram' or 'dot plot' (e.g. Maizel & Lenk 1981), and in fact this visual method closely resembles the way in which edit distances, and related measures of biological similarity described below, are calculated (see Figure 1.1).

1.2.1. Biological similarity.

The edit distance principle can also be used to compare nucleic acid or protein sequences, but it is important to bear in mind that similarity between biological sequences is often real and reflects descent from a common ancestral sequence ('homologous sequence') rather than coincidental similarity. However the picture is further complicated by the fact that unrelated sequences can also acquire similarity through convergent evolution ('analogous sequence').
Figure 1.1. Principles of dot matrix comparison. The sequences to be compared (which can be different sequences, or as in this case the same sequence in which regions of self-similarity are sought) are plotted along each axis to create a 'nxm matrix', and points of agreement between the axes are marked (A). Diagonals indicate extended regions of similarity. In DNA, reversed sequences are meaningless so bottom-left-to-top-right diagonals are uninformative, and a secondary layer showing reverse complement matches is required (not shown). Small alphabets like that of DNA produce noisy plots, but these can be filtered to show only those points that lie within meaningful diagonals (B), or to show different levels of noise with a greyramp scheme (C). The calculation of edit distance and related biological measures of similarity depend on the construction of a similar numerical matrix and the identification of the longest contiguous diagonal.

While the latter factor is difficult to account for, algorithms seeking to measure identity due to homology can attempt to model the actual biological processes through which sequences diverge (see Section 1.3).

Such principles were first applied to biological sequences by Saul Needleman and Christian Wunsch (Needleman & Wunsch 1970). The Needleman-Wunsch algorithm can be applied to DNA or proteins, but uses a different 'similarity matrix' for each. While each indel is given the same ‘gap penalty’ score (with no regards to the biological processes that introduce gaps), the similarity matrix assigns different scores to different kinds of substitutions.

In DNA, for example, transition mutations (purine to purine, i.e. A→G, or pyrimidine to pyrimidine, i.e. C→T) are more common than transversion mutations (A→T, A→C, G→C or G→T). Therefore AAAA and AAAG are scored as more similar thanAAAA and AAAC. In protein, a methionine residue can more readily transform into an isoleucine residue (a single third-
base transition: ATG → ATA) than into a serine residue (a minimum of two transversions: ATG → AGT).

With proteins there is selective pressure to maintain amino acids with similar properties at a given point in the sequence: methionine and isoleucine are both large hydrophobic residues, likely to be found on the inside of proteins, whereas serine is a tiny polar residue. To a certain extent the genetic code reflects such differences, but not always: histidine and glutamine codons differ only in the third base, for example, but histidine is hydrophobic while glutamine is polar.

The well-known Smith-Waterman algorithm (Smith & Waterman 1981) is a variation on the Needleman-Wunsch method, but it ignores low-scoring regions of similarity when calculating distances, since these produce a less certain measure of identity.

There are many other similar algorithms which use the same principles but are computationally more efficient (e.g. Myers & Miller 1988; Huang & Miller 1991). Various more sophisticated similarity matrices have also been developed for use in such algorithms, such as the PAM (Dayhoff et al. 1978) and BLOSUM (Henikoff & Henikoff 1992) matrices, derived from comparisons of actual biological sequence.

1.2.2. Problems faced by optimal alignment algorithms.

Obviously any algorithm that models biological sequence divergence can be confounded by random similarity or similarity due to convergent evolution, and equally is inappropriate for searching for such occurrences. Regulatory motifs are generally small subsequences that could look like
homology while actually arising through randomness or convergent evolution. Protein domains are also driven towards analogous forms by evolutionary pressure.

However such processes are highly unlikely to produce similarity across larger sequences, which consequently can more definitively be attributed to homology. But optimal homology-detecting algorithms are extremely time-intensive, and utterly impractical at the level of chromosomes or whole genomes.

Thus the emphasis in bioinformatics has moved away from finding algorithms that more accurately model the biology of sequence divergence, and towards algorithms that can search very large quantities of data for similarity of any kind.

1.2.3. Scaling up.

The BLAST algorithm (standing for Basic Local Alignment Search Tool; Altschul et al. 1990) and its variants have become the de facto standard for searching very large DNA or protein datasets for regions of local similarity (for example the NCBI database at http://www.ncbi.nlm.nih.gov/blast/).

To speed up searching, BLAST creates a table of all possible sequences of a given length. By default this ‘word size’ is set to 11 for DNA and 3 for protein. Each row of that table then contains pointers to all locations of that subsequence in the main sequence.

This simplifies the process of finding a subsequence within a larger sequence, and BLAST uses this table to find short exact matches between query and subject sequences (“maximal segment pairs” or MSPs). It then
searches outwards from those matches to find nearby matches, modifying the match score value as it does so in a manner analogous to the Smith-Waterman algorithm. In this way, the BLAST algorithm avoids calculating every point of similarity between two sequences (as a dot matrix plot or Smith-Waterman algorithm would).

The most time-consuming step in BLAST and related algorithms is searching through the word table. In the recent years various new algorithms have emerged that improve on this rate-limiting step. QUASAR (Burkhardt et al. 1999), REPuter (Kurtz & Schleiermacher 1999; Kurtz et al. 2000; 2001), MGA (Hohl et al. 2002) and MUMmer (Delcher et al. 1999; 2002) all use variations on the suffix tree data structure to index subsequences in place of a table.

Suffix trees are a clever and entirely unintuitive way of storing sequence information in such a form that it can be queried very efficiently, even though the data structure takes up considerably more space than the sequence on which it is based (Weiner 1973). As the name suggests they are a tree-shaped data structure rather than a table. Crucially, it is possible to construct suffix trees in linear time (McCreight 1976; Ukkonen 1995; Giegerich & Kurtz 1995; 1997).

Such developments are crucial given the exponentially increasing quantity of biological sequence data. The number of base pairs in GenBank alone doubled between 2002 and 2005, and now doubles every 10 months. At the time of writing it exceeds 65 billion (see http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html). It is easier than ever to find sequences similar to a sequence of interest, but harder to tell
whether such similarity represents homology or analogy, unless we make inferences based on taxonomic distances, or perform phylogenetic analyses using the older more biologically-grounded algorithms.

1.3: The Evolution of DNA Sequences

Strictly speaking, biological sequence divergence is not the same thing as evolution. Homologous sequences diverge by a variety of molecular processes, but many changes will be lethal to the organism, and so are not propagated. Between generations, biological sequence divergence has a constraint similar to that of the word-transformation puzzle, in that each intermediate stage must 'make sense'. Arguably this should not be called evolution either: it is 'survival' rather than 'survival of the fittest'. Evolution operates at another level again, adjusting the relative frequencies of different sequence versions (alleles) in a population according to the selective pressures of the population's environment. DNA sequences certainly can be said to evolve, over time, but to do so they must first change, in germ-line cells, by mechanisms that have little to do with evolution except in that they too have evolved.

1.3.1. Spontaneous sequence change?

Sometimes DNA sequence can change on its own. A cytosine residue can spontaneously hydrolyse and become deaminated to a uracil residue. Since uracil is not usually present in DNA (as it is in RNA) such unprovoked
sequence change is easy to detect and cells can reverse the effect accurately. If they did not, DNA replication would be impaired, and the error would not be propagated.

Some cytosine residues, however, are deliberately epigenetically modified to 5-methylcytosine. Such modifications confer various advantages to a cell: they can distinguish newly replicated strands from old, a distinction which is exploited by some DNA repair mechanisms, and they can distinguish invading viral DNA from the cell's own (Colot & Rossignol 1999). However when such an altered residue spontaneously deaminates, it becomes a thymine residue that in double-strand DNA will be mispaired with guanine.

Though cells have repair mechanisms capable of detecting such mispairings, they cannot necessarily tell which residue has been altered, and may 'repair' the wrong one. In higher eukaryotes CpG dinucleotides are common targets for epigenetic modification, and because of spontaneous deamination they are generally rarer than would be expected by chance. This process is probably also largely responsible for the composition bias that is common in life: most genomes are AT-rich.

Depurination, the wholesale loss of an adenine or guanine residue leaving only a hydroxyl (-OH) group, is an extremely common spontaneous process which leaves lesions in DNA strands. Chemical mutagens or radiation can also introduce lesions to DNA strands, or even produce double-strand breaks.

The process of DNA replication is also vulnerable to errors. The major DNA polymerases have a proof-reading 3'-5' exonuclease activity which is very efficient at correcting such errors as they occur, but in a tiny percentage
of cases the wrong base is incorporated into a newly forming strand and left in place. Where there are small tandem repeats of sequence the DNA polymerase complex sometimes slips back or forwards to alter the copy number of those repeats.

All such cases produce deformations in the DNA duplex that are likely to interfere with replication and transcription and may kill the cell if left unchecked. Thus the notion of spontaneous sequence change is misleading. Such small scale changes can usually only persist if they are repaired, but repaired incorrectly.

1.3.2. DNA repair mechanisms.

DNA repair is a complex subject, and the different repair processes often overlap in function, and in the proteins that mediate them. DNA repair mechanisms are also involved in 'deliberate' processes such as meiotic recombination. Though summarised here, the field is more thoroughly reviewed in Cunningham 1996, Jackson 1996, Paques & Haber 1999, Haber 2000, and van Gent et al. 2001.

In many organisms (but not mammals) some lesions, such as cyclobutane pyrimidine photodimers created by UV light, and some forms of alkylation, can be directly reversed by specific enzymes, restoring the affected nucleotides to their original state.

There are a variety of base excision repair (BER) mechanisms consisting of specialised enzymes to recognise and excise specific types of lesion. Excised bases are then replaced by DNA polymerases that use the unaffected complementary strand as a template, sealing the replacement
nucleotides into place with DNA ligase. Depurination lesions are recognised and excised with high efficiency by AP-endonucleases. Assorted DNA glycosylase enzymes exist to recognise other specific types of lesion, including those caused by oxidation, alkylation, the previously-mentioned deamination of cytosine to uracil, and certain common base mismatches such as the G-T mispairing left by the deamination of 5-methylcytosine.

Larger single-strand lesions can be treated by the **general excision repair** (GER) system, also called **nucleotide excision repair** (NER). A large complex scans for any abnormal deformations in the DNA helix, and excises a larger section of the afflicted strand, which is repaired with DNA polymerase and ligase. In *S. cerevisiae*, GER is mediated by a number of RAD genes, but other RAD genes are involved in different repair pathways, sometimes more than one. GER can be tied to transcription (**transcription-coupled repair**; TCR) so as to repair damage to genes before that damage interferes with the approaching transcriptional apparatus.

Replication can be coupled to repair in a similar way. A DNA replication fork reaching a single-stranded break will produce a more serious double-stranded break and stall replication, so there are last-minute **translesion synthesis** mechanisms associated with different minor DNA polymerases (Prakash *et al.* 2005) that repair such breaks before they are encountered, but in a highly error-prone fashion. For example in *S. cerevisiae*: DNA polymerase $\xi$, encoded by *REV3* and *REV7*, forms a complex with deoxycytidyl transferase (*REV1*) and inserts a molecule of deoxycytidine monophosphate into the gap (literally a ‘stop-gap measure’) which allows the replication
machinery to continue, but this causes a guanine residue to be incorporated into the new strand, wrongly in ~75% of cases.

General base mismatches in double-stranded DNA, including those caused by uncorrected errors in DNA replication, are handled by the mismatch repair (MMR) systems. Mismatches are detected by the deformation they produce, and these systems are efficient at using epigenetic DNA modification to distinguish which is the newly formed strand, and then excising a small (in ‘small patch’ MMR systems) or large (in ‘large patch’ MMR systems) section of that strand, which is repaired, as in GER, by polymerases and ligases.

Double-strand DNA breaks, such as those caused by high-energy radiation, are extremely problematic but thankfully much rarer than single-strand lesions. Broken ends of DNA are resected by 5'-3' exonucleases to produce single-stranded tails. If the sequence of the two tails is sufficiently homologous, because of the presence of repetitive sequence on either side of the breakpoint, for example, then they can be directly rejoined in the process of single-strand annealing (SSA). This process reduces the copy number of repeats.

Ends with microhomology between matching ‘sticky ends’ can still be rejoined by non-homologous end-joining (NHEJ) mechanisms, as can non-matching ends albeit with much reduced efficiency. In *S. cerevisiae*, such ends are bound by Ku heterodimers (Ku70p and Ku80p, also found at telomeres; see Section 1.5) and joined by DNA ligase IV (*DNL4*) and Ligase Interacting Factor (*LIF1*). The Mre11p / Rad50p / Xrs2p complex, involved in homologous repair as described below, is also generally required.
While NHEJ is not an extremely error-prone repair process, there is a chance of nucleotide loss or gain, particularly in homopolymer tracts. Under extreme conditions, with many double-strand breaks present in a cell, non-homologous end-joining can join the wrong ends together and create a genome translocation, but such conditions are rare and likely to kill the cell. In cells that have lost telomere function, NHEJ can result in chromosome fusion.

In some cases, however, multiple double-strand breaks are deliberately induced at specific sites, precisely to encourage translocation by NHEJ misrepair. This tactic of ‘V(D)J recombination’ is, for example, used to generate extreme receptor and immunoglobulin diversity in the B- and T-cells of the vertebrate immune system, but a similar mechanism is also employed by parasites (*Trypanosoma* and *Plasmodium* for example) to generate rapid diversity of surface antigens as a means of avoiding immune responses. The regions that undergo this recombination are often close to telomeres (see Section 1.5).

In cells that have replicated their chromosomes, or are diploid and thus have pairs of homologous chromosomes, the homologous sequence can be used as a template for the accurate repair of a double-strand DNA break by **homologous recombination** (HR) mechanisms. In *S. cerevisiae*, the break is recognised by Tel1p (also involved in telomere length regulation) and the Mre11p / Rad50p / Xrs2p complex, and ends are resected to produce single-stranded 3’ overhangs which are bound by Rad51p. This protein finds the homologous template, and with Rad52p and Rad54p the single-strand ends invade the homologous duplex and pair with the template strand. The damaged ends are then rebuilt along the template by polymerases, the
complex junction is disassembled by resolvase proteins, and nicks repaired by ligases.

If one allele is repaired using the other as a template, homologous repair can result in gene conversion. If one end of the double-strand break fails to meet the template, then the other end will be replicated all the way to the chromosome end based on the template sequence (break-induced replication; BIR), resulting in large-scale gene conversion. Depending on the topology of the junction and the way in which it is resolved, homologous repair can also result in crossing-over of homologous chromosomes, or crossing-over between homologous regions of different chromosomes (at subtelomeres, for example; see Section 1.5). While accidental mitotic recombination is normally rare, the process is actively induced in a controlled manner during meiosis.

1.3.3. Meiotic recombination and other large-scale sequence changes.

In sexual organisms, large-scale DNA sequence change occurs regularly in the process of genetic recombination between homologous chromosomes during meiosis. Homologous chromosomes typically recombine at multiple points along their length, a process which in S. cerevisiae is triggered by the deliberate creation of double-strand breaks by Spo11p, commonly at a specific set of loci (meiotic recombination hot-spots) that appear to often lie within the promoter regions of genes (Petes 2001). The breaks are then repaired by regulated homologous recombination. Resolution of the Holliday junctions produces a genetic cross-over of information in 50% of cases.
Abnormal meiotic recombination can result in unequal crossing-over: the duplication of genes on one chromosome and the deletion of genes from its partner. In extreme cases it can produce unbalanced chromosomes that produce lethal or severe phenotypes.

The mechanisms described above can produce gene duplications and even duplications of entire subtelomeric regions (see Section 1.5). *Sensu stricto* and *sensu lato* yeasts are, as previously described, the products of a whole genome duplication, and this may not be at all uncommon (Skrabanek & Wolfe 1998). Probably arising from polyploidy, whole genome duplications can create fertile ground for further sequence change, as the doubling of genes creates redundancy and one or other of the copies is freed of selective constraints.

Many genomes contain large quantities of mobile genetic sequences (such as transposons, retroviral sequences and plasmids) that have their own unique methods of reproduction and transmission and can even facilitate the horizontal transmission of DNA sequence between very different organisms. An overview of such elements in *S. cerevisiae* and other yeasts is provided as part of Chapter 3.

Thus, the forces that shape genome sequence are many and varied, operating at a host of different scales, and in different manners in different organisms. But they can also be different in different parts of a cell. The forces that shape a mitochondrial genome are not necessarily the same as those that operate on nuclear sequence, while telomeric and subtelomeric sequences can experience very different forces and selective pressures to centromeric sequences.
1.4: Centromere Biology

In most higher organisms, the cell manipulates the linear chromosomes during meiotic and mitotic divisions by means of a microtubule spindle attached to a unique region of each chromosome called the centromere.

Usually the centromere is a very long sequence made up of highly repetitive (satellite) heterochromatic DNA. Human centromeres, for example, consist of numerous interspersed and overlapping repetitive elements spanning several hundred kilobases, a structure that is particularly problematic for sequencing projects (She et al. 2004). Such repetition doubtless contributes to the high degree of sequence polymorphism observed in centromeres across the Eukaryota (Talbert et al. 2004).

The three centromeres of the fission yeast S. pombe have a similar structure albeit on a smaller scale, ranging from 38 to 97 kb in length, and likewise are highly polymorphic loci (Steiner et al. 1993).

1.4.1. Point centromeres in budding yeast.

The centromeres of S. cerevisiae and most other budding yeasts are strikingly different in nature. Dubbed 'point centromeres' because of their small size (from ~110 bp in S. cerevisiae to ~190 bp in E. gossypii), they are asymmetric and can be divided into three distinct ‘centromere DNA elements’ (CDEs). Two very short and quite well-conserved consensus sequences (CDEI and CDEIII) are separated by an AT-rich, and AA rich / TT rich tract (CDEII). CDEII DNA is clearly restricted in composition and is also, within any given budding yeast species, somewhat limited in length. While distinctly non-
repetitive (perhaps because the length restriction leads to selection against repeats, which are prone to length polymorphism), the sequence of CDEII is also clearly non-random (Baker & Rogers 2005).

The structure and variation of these elements across the budding yeast family is presented in Figure 1.2, and it is clear that the development of point centromeres predates the whole genome duplication event (Lalo et al. 1993; Wolfe & Shields 1997; Dietrich et al. 2004; Kellis et al. 2004).

It is also apparent that the first half of CDEIII (here dubbed CDEIIIa) is much less well conserved throughout evolution than the second half (CDEIIIb). It must be noted that some authorities extend the range of the CDEIII element to include additional distal sequence, based on experimental footprinting methodologies (Espelin et al. 1997; Pietrasanta et al. 1999). While there is a tendency towards AT-richness in this adjacent sequence, there is no discernable conservation of sequence, nor any clear boundary to the element if defined in this manner. Nevertheless these regions are often included within the definition of CDEIII and are here referred to as CDEIIIc.

Figure 1.2 also draws attention to two fascinating anomalies. *Saccharomyces castellii* is quite a close relative of *S. cerevisiae*, and one that would be expected to possess typical point centromeres. However no such structures (and indeed no centromeric DNA of any kind) have yet been identified. Even in intergenic regions where they might be expected by synteny to reside, point centromeres are absent, and these regions are much shorter than average (Cliften et al. 2006). This is extraordinary.
Figure 1.2. Centromeric DNA across yeast species. Phylogeny based upon a tree by Genolevures. For point centromeres, consensus sequences are shown as sequence logos (Schneider & Stephens 1990) which were calculated using WebLogo (Crooks et al. 2004); http://weblogo.berkeley.edu. The length range and composition of CDEII is summarised. The first part of CDEIII (designated CDEIIa) is barely conserved across these budding yeasts. Some authorities extend the range of the CDEIII element to include additional distal sequence (CDEIIIc) based on footprinting methodologies (Espelin et al. 1997; Pietrasanta et al. 1999). While there is a tendency towards AT-richness in this adjacent sequence, there is no discernable conservation of sequence, nor any clear boundary to the element if defined in this manner.
Looking further afield in the yeast family one might hope to find an intermediate stage of centromere structure that bridges the point centromere form and the large repetitive form found in *S. pombe*. However such a link is currently missing. In *Candida albicans*, the only intermediate species with suitable sequence coverage at present, each centromere appears to have a unique sequence with no discernable shared motif (Sanyal *et al.* 2004). Thus it would seem that there are at least three or four very different centromere-defining schemes within the ascomycetous yeasts.

1.4.2. Centromere-binding proteins.

In *S. cerevisiae*, the centromere DNA elements are directly bound by only a small set of proteins. CDEI is bound by Cbf1p, a leucine zipper protein which also binds the same sequence at *MET* gene promoters (Kuras *et al.* 1996; Wieland *et al.* 2001), although *MET* binding seems to be stabilised by other factors (Kent *et al.* 2004) while centromeric binding seems to require the CBF3 complex described below (Wilmen *et al.* 1994). Cbf1p recruits Isw1p and has a generalised role in nucleosome positioning. Null mutants are viable but grow poorly and suffer increased chromosome loss, and there are changes to the structure of centromeric chromatin (Mellor *et al.* 1990).

In many eukaryotes, centromeric chromatin is based on a specialised centromere-specific nucleosome in which histone H3 is replaced by a variant component (Henikoff *et al.* 2000; Takahashi *et al.* 2000; Sanyal & Carbon 2002). In humans this variant is CENP-A (Sullivan *et al.* 1994). In *S. cerevisiae*, histone H3 is replaced by Cse4p, which is homologous to human CENP-A, and of course homologous to histone H3 (Meluh *et al.* 1998).
AT-rich DNA of CDEII presumably wraps around this specialised nucleosome, and it may be this interaction which imposes a restriction on the composition and length of this DNA element. A standard eukaryotic nucleosome is wrapped by ~146bp of DNA, suggesting that there is only room for CDEII to wrap around a single specialised centromeric nucleosome in *S. cerevisiae*. However, as shown in **Figure 1.2**, the CDEII elements of *K. lactis* and *E. gossypii* are approximately double the length of those in *S. cerevisiae*, implying that these CDEIIs might be wrapped twice around a centromeric nucleosome.

*Mif2p* (homologous to human CENP-C) may bind to CDEII since it contains a DNA-binding AT-hook motif (Meluh & Koshland 1995). It certainly binds to CDEIII *in vitro* (Meluh & Koshland 1997). It is an essential protein that interacts with histone components and may be recruited to the centromere by them. It then binds the MIND and COMA kinetochore complexes where spindle microtubules attach during cell division (Westermann *et al.* 2003).

Finally, a protein complex called CBF3 binds to CDEIII, but this complex also associates with other sequence elements as well as centromere-binding proteins. It consists of four subunits, all of which are essential.

The largest subunit **Cbf2p** (also widely referred to as **Ndc10p**) also independently binds CDEII, evidently by contacting that sequence at various points along its length, and can bind to non-centromeric sequences with a similar composition (Espelin *et al.* 2003).

**Cep3p** is the only subunit in the complex to have a standard DNA-binding motif, and it mediates binding to the CDEIII consensus sequence,
though it is unable to bind on its own (Espelin et al. 1997). Cep3p also physically interacts with Cbf1p (Hemmerich et al. 2000).

**Ctf13p** is an essential subunit of CBF3 but has no known strong characteristics of its own, nor any well-characterised protein domains. It may serve merely to bind the complex together.

The smallest subunit **Skp1p** has a regulatory function: it interacts with the spindle checkpoint via Bub1p (Kitagawa et al. 2003) but also forms part of several other complexes with differing functions and is extremely well conserved throughout eukaryotes (Connelly & Hieter 1996; Seol et al. 2001).

From *in vitro* experiments, the binding of the CBF3 complex to centromeric DNA also requires contacts to bases past the end of the CDEIII consensus. As previously mentioned, this region (CDEIIIc) tends to be AT-rich but has no discernable pattern. Thus the complex binds in a non-sequence-specific manner, perhaps more like the binding of centromeric proteins in higher eukaryotes (Espelin et al. 1997; Pietrasanta et al. 1999).

While the above proteins are the only ones currently known to be involved in binding centromeric DNA, there is clearly a complicated set of interactions between them and the various CDEs. Nevertheless, since only one of these proteins is non-essential (Cbf1p), and the loss of that protein is still heavily detrimental, there is evidently remarkably little redundancy built into the centromere-binding mechanism, considering the vital importance of this process.
1.4.3. Centromere replication and repair.

At the present time only one of the *S. cerevisiae* centromeres, CEN3, is known to overlap an autonomous replicating sequence (ARS308), although CEN12 is directly adjacent to ARS1208. There may be other as-yet-unidentified ARSs that overlap other centromeres, but a detailed study of replication times throughout the *S. cerevisiae* genome (Raghuraman *et al.* 2001) indicated that this is not always—or even usually—the case, since centromeric DNA is on the whole not replicated unusually early, nor all at the same time, though it would appear that it is never replicated late in the S phase.

There is evidence that recombination is often strongly suppressed around centromeres in higher eukaryotes, and this might be expected since sister chromatids are joined at the centromere in prophase I. In humans this suppression is evidently stronger in metacentric chromosomes than in acrocentric ones (Nachman 2002).

The centromeres in *S. cerevisiae*, though substantially different in size and composition, are ‘coldspots’ for meiotic recombination (Gerton *et al.* 2000). However the situation may be complex. One set of studies (Lambie & Roeder 1986; 1988) found substantial repression of meiotic crossing-over when CEN3 sequence was inserted next to markers, but a mutant version of CEN3, with a single-base-pair mutation in CDEIII, actually stimulated crossover events. In a different set of studies (Liebman *et al.* 1988; Symington & Petes 1988) mitotic gene conversion appeared to be unhindered by the presence of centromeres.
If centromeres do inhibit recombination then double-strand break repair by homologous recombination is likely to be hindered, and with it the process of gene conversion that can homogenise sequence polymorphisms.

1.5: Telomere Biology

Most simple bacterial organisms have circular chromosomes, but it has been argued that the original nucleic acid genome at the root of the tree of life, emerging from the ‘RNA world’, was linear and multipartite (Poole et al. 1998). Certainly higher organisms almost always have linear chromosomes, which plainly can attain much greater lengths than circular ones. However, linear chromosomes must have ends, and these ends present a number of challenges. They must be protected from the action of exonucleases, and from double-strand break repair mechanisms, otherwise they will be rapidly degraded as if they were invading foreign DNA, or ‘repaired’ in the same manner as the unwanted loose ends caused by DNA damage, which would cause linear chromosomes to be circularised or fused together (see Section 1.3).

For these reasons telomeres, the ends of linear chromosomes, have a very distinctive sequence structure and associate with a small army of telomere-specific proteins that distinguish them from undesired DNA ends.

Moreover, as a consequence of the 5’-3’ directionality of DNA polymerase and the need for RNA primers, it is impossible to replicate a linear DNA molecule to the very ends (Levy et al. 1992). This ‘end-replication
problem' would inevitably lead to a shortening of chromosomes over the generations, and eventually the loss of important genes were there not mechanisms to periodically extend them.

In most known organisms (*Drosophila* being a notable exception; Levis *et al.* 1993) the telomeres consist of multiple short tandem repeats with a single-stranded 3' overhang. Extension of these repeats is accomplished by a specialised reverse-transcriptase (telomerase) containing an RNA moiety that binds to single-stranded repeats. Telomerase synthesises a single repeat in each round of polymerisation, and may perform several rounds before disassociating from the telomere. The complementary strand is presumably synthesised by conventional DNA polymerases.

Telomeric repeats are usually regular, and indeed the same repeat sequence (TTAGGG) is found in organisms as diverse as humans and *Neurospora crassa* (Schechtman 1990). However among ascomycetous yeasts the repeat unit, though generally TG-rich, varies a great deal in length and composition. *S. castellii*, *S. dairenensis* and *C. tropicalis* are each known to have multiple types of repeats, which may be produced by different segments of a single telomerase template, or by multiple telomerases (McEachern & Blackburn 1994; Chappell & Lundblad 2004). In *S. pombe* the repeat units are highly irregular (Matsumoto *et al.* 1987; Lue & Peng 1997), and they are also irregular in *S. cerevisiae*.

**1.5.1. Telomeres in *S. cerevisiae*.**

The telomeric repeats of *S. cerevisiae* consist of imperfect TG$_{1-3}$ repeats (C$_{1-3}$A on the complementary strand), typically about 300 bp long
(Shampay & Blackburn 1988; Wright et al. 1992) with a single-stranded 3' TG$_{1-3}$ overhang. This overhang is thought to be at least 12-14 nucleotides long for much of the cell cycle (Larrivee et al. 2004) but peaks at >25 nt during S phase (Wellinger et al. 1993). The overall length of double-stranded telomeric repeats is also highly variable, depending on when and to what extent they have been extended.

The telomerase RNA subunit in *S. cerevisiae*, encoded by *TLC1*, contains a template which is complementary to the sequence TGTGTGGGTGTGGT (Dandjinou et al. 2004), and the irregularity of the telomeric repeats probably results from the use of alternative start and stop positions within this template (Forstemann & Lingner 2001).

The telomeric repeats of *S. cerevisiae* are accompanied by some distinctive telomere-associated sequences (TASs) common to many or all chromosome ends. Three distinct types of TAS were originally identified through hybridisation experiments: a Y element, a ‘131’ sequence and an X element (Chan & Tye 1983). These early definitions have since been refined. Y and 131 elements are now regarded as a single unit called Y' (Walmsley et al. 1984). Following Chan and Tye’s observation that “X sequences may actually be composed of a number of smaller repetitive elements and each of the homologous X sequences may contain a subset of these smaller repetitive elements”, X elements have been subdivided into a ‘core’ X motif that is present at all chromosome ends, and four types of imperfect repeats that appear together in different amounts and combinations at each end (Louis et al. 1994). Figure 1.3 shows the distribution of these and position of these elements in the sequenced strain of S288C.
Figure 1.3. Specific and general structure of chromosome ends in S. cerevisiae S288C. Also shown, to the left and at a different scale, are the relative lengths of each chromosome arm. There is no correlation between arm length and the presence or absence of different telomere-associated elements.
1.5.2. Y' elements.

Y' elements appear within, or immediately adjacent to, telomeric repeats, and can be present in multiple tandem copies. These elements range in length from 4.7 kb (VI-L) to 6.9 kb (IX-L and X-L) and are often subclassified as either short type (<6 kb) or long type (>6 kb) although in fact there is a fairly smooth range of intermediate lengths.

Seven of these Y' elements contain lengthy ORFs (YRF1-1/YDR545W, YRF1-2/YER190W, YRF1-3/YGR296W, YRF1-4/YLR466W, YRF1-5/YLR467W, YRF1-6/YNL339C, and YRF1-7/YPL283C), which encode DEAD/DEAH-box ATP-dependent RNA helicases with domains also found in the Translation Initiation Factors encoded by TIF1/YKR059W and TIF2/YJL138C.

Three of these Y' ORFs (YRF1-3, YRF1-6 and YRF1-7) contain short introns. In the remaining Y' elements the ORFs have degenerated to become pseudogenes or multiple separate ORFs.

The helicase proteins (Y'-Help1p) encoded by the intact YRF1 ORFs are highly expressed in telomerase-deficient cells (Yamada et al. 1998), and in some of these mutants Y' elements become greatly amplified in copy number, taking over the role of telomere extension normally accomplished by telomerase (the type I ALT mechanism described below). It is hypothesised that Y'-Help1 expression is normally suppressed by telomeric silencing (Gottschling et al. 1990) but telomere loss in telomerase-deficient cells removes this silencing and the Y'-Help1 protein then promotes Y' amplification in some manner, perhaps by homologous recombination (Yamada et al. 1998). Y' elements also contain their own Autonomous Replication
Sequences (ARSs) and have even been detected as circular extrachromosomal elements (Horowitz & Haber 1985).

Y' elements appear to be highly mobile: the particular configuration varying considerably from strain to strain and even within a given strain (Button & Astell 1986; Jager & Philippsen 1989). There appears to be no correlation between Y' copy number and other features of the chromosome end in question. In behaviour Y' elements seem like a form of selfish DNA, and there are some intriguing connections between Y' elements and retrotransposons, discussed in Chapter 3.

1.5.3. Core X elements and X element combinatorial repeats.

X elements lie between telomeric repeats and the rest of the chromosome. The core X element is more weakly conserved than Y' elements but is found at all chromosome ends in the sequenced strain, varying in length from ~200-470 bp (Louis et al. 1994). Almost all core X elements contain an ARS consensus sequence and a binding site for Abf1p (ARS-binding factor 1). Core X elements improve the segregation efficiency of plasmids carrying telomere repeats (Enomoto et al. 1994).

The repeats at the telomere-proximal end of most X elements ('subtelomeric' or 'X element combinatorial' repeats) are generally TG-rich and highly imperfect, but can nevertheless be loosely subclassified into four types: A-D in order from the telomeres inwards. The order of the repeat unit types is conserved but not all types are present at all ends and the copy numbers of each repeat unit varies. Most of these regions contain binding sites for Tbf1p (TTAGGG-repeat Binding Factor 1), an essential DNA binding protein thought
to be involved in the regulation and repression of telomeric silencing (Fourel et al. 1999). The Tbf1p binding motif (TTAGGG) is the same as the standard telomeric repeat unit found in most other organisms, perhaps an indication that ancestral budding yeast telomeres were more similar to the eukaryotic norm.

To this author, the combinatorial repeats resemble 'tide marks', perhaps left behind at various 'heights' after periodic severe losses of telomere length throughout the evolution of the species. If so then they may represent remnants of previous telomeric repeat types, from more recent types (A) to more ancient ones (D).

In two S. cerevisiae S288C telomeric regions (chr IX-L & X-L, which are an example of subtelomeric duplication; see below) the combinatorial repeats are interrupted by a large insertion which extends into the adjacent Y' elements. The majority of this insertion is identical in sequence to a group I self-splicing intron from the mitochondrial cytochrome b gene Bl4 (Louis & Haber 1991): a case of horizontal transfer from the mitochondrial genome to the nuclear genome.

1.5.4. Telomere-binding proteins in S. cerevisiae.

The S. cerevisiae telomerase core enzyme consists of an RNA subunit (TLC1) and a protein component Est2p. There are also two regulatory components that interact with telomerase and telomeres: Est3p and Est1p. The latter helps to recruit the holoenzyme to its target, and binds to TLC1 (Lin & Zakian 1995; Steiner et al. 1996; Zhou et al. 2000), the single-stranded
telomeric repeats (Virta-Pearlman et al. 1996; Zhou et al. 2000), and to Cdc13p (Qi & Zakian 2000).

**Cdc13p** binds along the single-stranded overhang, thereby ‘capping’ the telomeres, and regulates telomere maintenance through its interaction with other proteins (Lin et al. 2001; Lustig 2001). Cdc13p is well conserved amongst the budding yeasts, but not in higher eukaryotes (Mitton-Fry et al. 2004), although in higher eukaryotes there are functionally-analogous proteins with DNA-binding OB-fold domains indicative of a distant evolutionary relationship (Theobald et al. 2003). In **CDC13** mutants, single-stranded DNA is generated towards the centromere and triggers cell arrest (Booth et al. 2001).

Cdc13p and **Stn1p** each interact with different major subunits (Pol1p and Pol12p respectively) of DNA polymerase α (Qi & Zakian 2000; Grossi et al. 2004), thus tying telomerase action on the single-stranded overhang to synthesis of the complementary lagging strand. These two proteins form a complex with a third essential telomeric protein **Ten1p** (Grandin et al. 2001).

Cdc13p also binds to **Stm1p**, which plays a role in triggering cell death: null mutants actually show increased cell survival under normal growth conditions (Ligr et al. 2001). Stm1p in turn binds to ribosomal components as well as Y' elements and G-quadruplex DNA (Van Dyke et al. 2004). G-quadruplex DNA is an unusual conformation that can be adopted by G-rich DNA (Liu et al. 1995), and there is evidence that such structures do form at telomeres (Henderson et al. 1987), and that secondary structures of this kind might play a role in the ATP and GTP-independent translocation of telomerase along the single-stranded overhang (Shippen-Lentz & Blackburn
DNA helicase Sgs1p, which is required for the type II ALT mechanism (described below) preferentially unwinds this DNA conformation (Sun et al. 1999), and Rap1p (see below) promotes the formation of G-quadruplex DNA \textit{in vitro} (Giraldo & Rhodes 1994).

The single-stranded telomeric overhang is also bound by essential protein Tel2p, the exact role of which remains unclear (Runge & Zakian 1996; Kota & Runge 1998).

There are also various proteins that bind to non-telomeric DNA as well. Rap1p is highly multifunctional. It binds to double-stranded telomeric repeats, but also a broad range of other consensus sequences found at locations throughout the genome (reviewed in Wahlin & Cohn 2002). It would seem to create a barrier against the spread of silent chromatin, but also helps to create silent chromatin at telomeres and silent mating-type loci (Morse 2000). It functions as a transcriptional activator at many promoters, especially for ribosomal genes (Yu et al. 2003). The length of telomeres appears to be largely regulated through a mechanism of Rap1p counting (Marcand et al. 1997; Brevet et al. 2003), though there are other telomeric proteins involved, such as ‘Rap1p-Interacting Factors’ Rif1p and Rif2p.

Rap1p also binds to Silent Information Regulators Sir3p and Sir4p (Moretti & Shore 2001; Luo et al. 2002), which bind to histones and regulate the formation of silent chromatin at telomeres and other locations (Hecht et al. 1995). \textit{SIR3} and origin recognition complex gene \textit{ORC1} are paralogues that arose in the whole genome duplication event, with \textit{SIR3} subsequently undergoing accelerated evolution to adopt its new function (Kellis et al. 2004).
A third Silent Information Regulator Sir2p also binds at telomeres, and at silent mating-type loci and rDNA loci (Strahl-Bolsinger et al. 1997). One of the Sir2p domains is extremely well conserved across multiple kingdoms, and there are multiple homologues of this gene (HST1-4) in *S. cerevisiae* (Brachmann et al. 1995).

The subunits of the Ku heterodimer, Yku70p and Yku80p, are key components in non-homologous end-joining (see Section 1.3), but also bind to telomeres, at the junction between single-stranded and double-stranded DNA (Mimori & Hardin 1986; Falzon et al. 1993). At telomeres they appear to have multiple functions: recruiting telomerase and regulating chromosome length (Stellwagen et al. 2003), and physically tethering clusters of telomeres to the nuclear periphery (Laroche et al. 1998).

The large family of telomere-binding proteins create a unique heterochromatin structure at telomeres that has epigenetic effects on the transcription of nearby genes: the ‘position effect variegation’ of silencing (Gottschling et al. 1990; Aparicio et al. 1991; Laurensen & Rine 1992; Vega-Palas et al. 1998; 2000). Such silencing is by no means even, but rather displays patterns which indicate that telomeric repeats may loop back and interact with core X elements (Pryde & Louis 1997); heterochromatin, and hence silencing, being concentrated around this junction.

### 1.5.5. Alternative Lengthening of Telomeres (ALT).

When key telomerase components are damaged or lost, colonies eventually senesce as subsequent generations lose telomeres until DNA damage checkpoints lead to permanent growth arrest. However a small
proportion of cells recover from senescence by replacing lost telomeres without telomerase.

In *S. cerevisiae*, two main ALT (Alternative Lengthening of Telomeres) mechanisms have been observed: 'type I' and 'type II', both relying on homologous recombination (Lundblad & Blackburn 1993; McEachern & Blackburn 1995; Lundblad 2002).

**Type I** survivors are generally the first to appear and flourish in a senescing population. In these cells Y' elements have proliferated into long tandem arrays at many or all chromosome ends. However most type I survivors eventually senesce again, unless they subsequently recover through the type II mechanism.

**Type II** survivors exhibit a sudden and rapid amplification of telomeric repeats to ~20 times their usual length. While both these ALT mechanisms rely on functional Rad52p and other DNA repair proteins (Chen *et al*. 2001), the type II mechanism also relies on DNA helicase Sgs1p (Huang *et al*. 2001).

Mechanisms analogous to both type II and type I ALT have been observed to occur in human cancer cells (Reddel 2003).

In *S. cerevisiae* cells that fail to avert telomere loss, the nuclease Exo1p plays a key role in triggering cell cycle arrest and death, presumably because its 5'-'3' exonuclease activity generates large amounts of unprotected single-stranded DNA which either triggers damage checkpoints or degrades essential genes (Maringele & Lydall 2002; Pang *et al*. 2003). In *S. cerevisiae* mutants that are unable to maintain their telomeres by any of the above mechanisms, and which additionally lack functional Exo1p and so do not arrest or die in the usual fashion, there appears, at a low frequency, a third
very different mechanism for rebuilding chromosome ends, even after conventional telomeres have vanished. This **PAL** mechanism produces palindromic duplications of sequence, often very extensive, at chromosome ends. It is hypothesised that when end degradation reaches an inverted repeat sequence, a loose single-stranded inverted repeat at the degraded end loops back to base pair with itself, thus providing a primer for DNA synthesis back along the strand (displacing the original strand in its path) until processivity is interrupted. Following resolution and replication, what emerges is a chromosome with an extended region of palindromic duplication. There are palindromic sequences of substantial length in other organisms (there are several in the human Y chromosome) which could have arisen in a similar manner. The PAL mechanism, although it was discovered in *S. cerevisiae*, relies on ubiquitous features of DNA sequence and repair, so it may explain these palindromic genome features in a range of different species, while also suggesting one possible primordial mechanism for maintaining the ends of linear chromosomes before the telomerase system evolved (Maringele & Lydall 2004b; Maringele & Lydall 2004a).

1.5.6. **Subtelomeric duplications.**

In the sequenced strain the subtelomeric regions immediately internal to the telomeric regions commonly contain large blocks of homology shared between two or more chromosome ends, extending for as far as 25kb into the chromosome (Bowman *et al.* 1997; Jacq *et al.* 1997; Louis & Becker, unpublished analysis). The duplication of these regions is indicative of an elevated number of ectopic recombination events in the recent past.
‘Subtelomeric plasticity’ of this kind may be advantageous in an evolutionary sense, providing genomic domains where mutation and recombination can regularly generate variety without disrupting other more important sections of the genome (Louis 1995). Many of the duplicated genes in these regions are involved in secondary metabolism, toxin resistance and cell-to-cell interaction: generally non-essential processes that vary greatly between strains and species of budding yeast. Subtelomeric plasticity has evidently contributed to the adaptive diversity of the entire budding yeast family.
CHAPTER 2: MATERIALS AND METHODS
2.1: Introduction

Almost all of the work described herein is computational in nature. The primary ‘raw materials’ in this work are assorted datasets produced by other researchers, mainly sequence data from a range of yeast strains. The primary ‘methods’ take the form of various third party programs, as well as custom-built scripts written in the programming language Python.

Sources and assessments of these data and software are provided in this section. Bioinformatics is strongly tied to internet resources, and useful web addresses are provided in the body of the text below. In the CD-ROM accompanying this thesis (see Appendix I) there is an electronic version of this document, with active hyperlinks.

Due to the changeable nature of the internet and of the bioinformatics resources therein, no guarantee can be offered that such addresses will remain valid in the long term, although it is to be hoped that they will remain so in the immediate future.

2.2: Sequence and Feature Data

The highly annotated S. cerevisiae genome (the ‘reference genome’ against which all the other yeast genomes are compared) was downloaded from the Saccharomyces Genome Database (SGD: http://www.yeastgenome.org). The most commonly-used (and regularly updated) file was ftp://genome-ftp.stanford.edu/pub/yeast/data_download/chromosomal_feature/saccharomy
ces_cerevisiae.gff, which contains full feature information in the concise GFF3 format as well as the chromosomal sequence data in FASTA format at the end of the file (a recognised standard within the GFF3 format, which is described in Section 2.4).

Most of the work described herein made use of additional whole yeast genome sequencing projects undertaken by Kellis et al. (2003) at the Broad Institute in Cambridge, Massachusetts. All these data were acquired from http://www.broad.mit.edu/annotation/fungi/comp_yeasts/downloads.html.

The sequence data for the genomes of *S. paradoxus*, *S. mikatae*, and *S. bayanus* were incomplete and in multiple contigs, including parts that could not be unambiguously aligned to a single location in the *S. cerevisiae* reference genome. These contigs, in FASTA format, were downloaded from part S1a of the website listed above. Feature annotation files were downloaded from part S2a. These latter files use a customised format (see Section 2.4 below).

Additional genome projects (*S. mikatae* and *S. bayanus* again, as well as *S. kudriavzevii*, *S. castellii* and *S. kluyveri*) were undertaken by Cliften et al. (2003) at Washington University in St. Louis. These data (in FASTA format, compressed with gzip) were downloaded from http://www.genetics.wustl.edu/saccharomycesgenomes/download.html. *S. kluyveri* is also known as *Lachancea kluyveri*. These genomes were more incompletely assembled than those provided by the Broad Institute, with each consisting of a larger number of separate contigs. Moreover no feature annotations were provided, making the data of less use in this work, although
since these genomes are more distantly related to *S. cerevisiae* than the Broad Institute genomes, they were of less direct relevance anyway.

Even more distantly related yeast genomes have also been sequenced and annotated. The Genolevures consortium (Feldmann 2000; http://cbi.labri.fr/Genolevures) undertook partial sequencing of a large number of yeast species. Some of these (*Candida glabrata, Kluyveromyces lactis, Debaryomyces hansenii* and *Yarrowia lipolytica*) were subsequently completed (Dujon *et al.* 2004) and are most readily available through the list of Ascomycota genome projects at NCBI (as are most of the other projects mentioned here, see ftp://ftp.ncbi.nih.gov.genomes/Fungi or http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=PureSearch&db=genome pri&details_term=%22Ascomycota%22%5BOrganism%5D. All have genome annotations and each chromosome is available as a separate GenBank Flat File (see Section 2.4) which contains both annotations and sequence. NCBI also provides feature files in the gff format (one per chromosome) but does not package sequence into those files.

The *Eremothecium gossypii* (also known as *Ashbya gossypii*) genome (Dietrich *et al.* 2004) is complete and annotated and can be downloaded from the same source.

The *Kluyveromyces waltii* genome was also sequenced at the Broad Institute (Kellis *et al.* 2004) but is available from http://www.nature.com/nature/journal/v428/n6983/extref/nature02424-s1.htm. These files follow the same conventions and formats as the equivalent files for the earlier Broad Institute genome projects.
The incomplete *Candida albicans* genome project has been contributed to by multiple sources (Jones et al. 2004) but is collated at http://www.candidagenome.org, a site which is closely patterned after SGD. Sequences and feature files were acquired from http://www.candidagenome.org/download. Supercontig files are available in FASTA format, compressed with gzip, while the chromosomal features are listed in a custom tab-delimited format.

The fission yeast *Schizosaccharomyces pombe*, one of *S. cerevisiae*’s most distant yeast relatives, has also been almost completely sequenced (Wood et al. 2002; there are five ~1000bp gaps) and thoroughly annotated, and its information is centralised at the *S. pombe* GeneDB at http://www.genedb.org/genedb/pombe. Chromosomal sequence can be acquired either through NCBI (in GenBank format) or from the Sanger Centre (in EMBL format) at ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Chromosome_contigs. For each of the three chromosomes, both kinds of file combine sequence and feature information.

### 2.3 Additional Data

The Saccharomyces Genome Database (SGD: http://www.yeastgenome.org), which is maintained by the Department of Genetics at Stanford University, provides an invaluable, well-organised and constantly updated central repository for almost all *S. cerevisiae*-related information. An alternative is Comprehensive Yeast Genome Database
(http://mips.gsf.de/genre/proj/yeast) operated by the Munich Information Center For Protein Sequences (MIPS).

SGD maintains some tools which have been used periodically in the course of this study. At http://db.yeastgenome.org/cgi-bin/gbrowse/yeast, SGD maintains a version of the open source CGI-based Generic Genome Browser (GBrowse; http://www.gmod.org/?g=node/71). SGD’s Advanced Feature Search facility at http://db.yeastgenome.org/cgi-bin/search/featureSearch can be used to retrieve features by type and with various additional search criteria (all ‘dubious’ ORFs on chromosome VI, for example), and the Batch Download facility allows users to download various information about features selected in this way, including coordinates and genomic sequence. This was commonly used to retrieve a list of essential genes (ORFs with the ‘inviable’ systematic deletion phenotype); information that was used in the study of overlapping genes described in Chapter 3 and for the genome-wide analysis of intergenic divergence described in Chapter 4.

SGD also stores Serial Analysis of Gene Expression (SAGE) data, used in Chapter 3, which can be downloaded via FTP from ftp://genome-ftp.stanford.edu/pub/yeast/data_download/systematic_results/SAGE.
2.4 Main Data Formats

The field of bioinformatics suffers from a surfeit of different data formats, and for any given kind of data there is generally no clear universal standard used by all programs and all practitioners. This section, while not exhaustive, describes and assesses the key formats relevant to this work.

2.4.1. Plain text format

A plain text sequence file can contain a single protein or DNA sequence as simple ASCII with no header or additional information. Many programs will accept plain text sequence as input, and the only constraint on the symbols that may be included is what a given program can understand. For example, such files typically contain no line breaks but some programs may be able to read plain text sequence files even if they contain them.

2.4.2. FASTA / Pearson format

The FASTA biological sequence format, also known as the Pearson format, is a simple standard created for the FASTA sequence alignment suite (Lipman & Pearson 1985) and almost universally recognised. Each sequence is preceded by a header line beginning with ‘>’. The associated sequence data follows on preceding lines, often split into lines of 80 characters or less (this is not a strict requirement of the format, but some programs will reject the file if this is not the case).

Either protein or DNA sequence can be listed thereafter, using the relevant IUPAC-IUBMB standard alphabet (see
The format thus supports ambiguous assignments, as well as gap assignments ('-'), though individual programs using the format may not be able to handle these. Multiple headers and sequences can be concatenated within the same file. FASTA formatted sequence data can be included in the GFF3 format described below.

### 2.4.3. GFF3 format

The 'Generic Feature Format 3' (GFF3) assigns a row to each feature. The first eight tab-delimited columns contain standardised information: sequence name (e.g. chromosome name), information source, feature type (generally using standard SOFA ontology terms as described at http://song.sourceforge.net/so.shtml), first coordinate (integer), last coordinate (integer), score (or '.' if irrelevant), strand ('+'/'-' or '.' if irrelevant), and exon phase (or '.' if irrelevant).

The ninth column lists additional attributes separated by semi-colons. A number of standardised attribute tags are defined, all or none of which may be present: ID (should be a unique identifier for the feature), Name, Alias (e.g. alternative gene names), Parent (if a subfeature), Target (indicating alignments), Gap (indicating non-colinearity of alignments), Note (feature description), Dbxref (database reference number) and Ontology_term (standard Gene Ontology codes as described at http://www.geneontology.org/ontology/gene_ontology.obo).

The GFF3 format can be considered flexible in that additional attribute tags can be added as necessary (though such additions will not necessarily
be recognised by other programs). If present, a tag is followed an ‘=’ sign and then the information content itself (comma-separated if, for example, there are multiple entries under ‘Alias’). Thus certain symbols within the information content (e.g. tabs, equals signs, commas) must be converted to the relevant HTML escape code (of the format %xx, where xx is the hexadecimal ASCII value for the symbol) if they are not to break the format.

Comment lines are preceded by the hash (‘#’) symbol, with metadata lines preceded by two hashes. FASTA-formatted sequence data can be appended to the end of a GFF3 file after the metadata tag line ‘##FASTA’.

The GFF3 format strikes a good balance between human- and machine-readability, and has good extensibility and widespread acceptance (it is one of the primary feature annotation formats in the open-source Generic Genome Browser (http://www.gmod.org/?q=node/71). For these reasons it is employed as an output format for data produced by this research (see Appendix I). Additional information on the standard is available at http://song.sourceforge.net/gff3-jan04.shtml.

2.4.4. Broad Institute Forward Feature File formats

The feature annotations provided by Kellis et al. (2003; 2004) for S. paradoxus, S. mikatae, S. bayanus and K. waltii are provided in two different tab-delimited forms. The ‘forward’ file for each new genome lists that genome’s ORFs in order of scaffold number and position on that scaffold, along with the name, where possible, of their S. cerevisiae counterparts (but not the S. cerevisiae coordinates of those counterparts). The ‘reverse’ files are formatted quite differently and list S. cerevisiae ORFs in order, and their S.
cerevisiae coordinates, and providing, where possible, the contig numbers (but not the contig coordinates) for their counterparts in the newly sequenced genome.

Additional tab-delimited information, which is much the same in the two file types though ordered differently, includes ORF descriptions, percentage identity and percentage conservation of homologous proteins, the amount of coverage (some of the new sequence only partially covers the corresponding S. cerevisiae ORF), and assorted book-keeping information from the software that generated it (such as the ‘homology group’ to which a new ORF was assigned, if any). Full details of the information type in each column is provided in the header comments in each type of file.

Both ‘forward’ and ‘reverse’ files have ‘best match’ and ‘all match’ alternative versions.

The ‘reverse’ files may be of greater utility to a human biologist looking to locate the homologue of a particular S. cerevisiae gene, but one will be required to then refer to the ‘forward’ file to determine the contig coordinates of the homologue. For computational parsing the ‘forward’ type contains the most useful information, and the genome-wide analysis scripts developed for this work (see Section 2.6) process this file type.

2.4.5. GenBank Flat File format

The GenBank database is a repository for nucleotide and protein sequences hosted by the National Centre for Biotechnology Information (NCBI) in the US and can be searched using NCBI’s Entrez tool at http://www.ncbi.nlm.nih.gov. NCBI also maintains a non-redundant, cross-
referenced and highly curated sequence database called RefSeq. Sequences returned from either database are in GenBank’s ‘Flat File’ format.

This is information-rich, makes heavy use of multiple data-type tags, and is optimised for human readability. On the other hand it is memory intensive and relatively difficult to parse computationally. Each file can contain a single nucleotide sequence (broken down into six blocks of ten nucleotides per line, with each line starting with a coordinate) plus extensive annotations. Keyword and reference information can be included, as well as feature information that can include protein translations of genes.

A sample record, including field descriptions, is provided by NCBI at http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html.

BioPython includes scripts to directly fetch sequences from GenBank and parse the information therein (see Section 2.6).

2.4.6. EMBL format

The European Molecular Biology Laboratory’s European Bioinformatics Institute (EMBL-EBI) maintains its own nucleotide sequence database (EMBL-Bank), which coordinates and shares information with GenBank and with the DNA Database of Japan (DDBJ). This database can be searched through EMBL-EBI’s webpage at http://www.ebi.ac.uk.

The EMBL format bears some similarities to the GenBank format and stores the same range of information, albeit with shorter two-letter field tags (which are more suited to computational parsing than human reading). A sample record and description of field types is provided at http://www.ebi.ac.uk/help/formats.html#EMBL.
2.4.7. ClustalW's ALN format

ALN files, outputted by ClustalW (see Section 2.5), are designed to store alignment information between two or more nucleotide or protein sequences. The file header begins with 'CLUSTAL' followed by additional information about the version of ClustalW that created the file. Thereafter the file is split into sections of alignment 60 residues long. Each section then consists of a line for each sequence in the alignment (beginning with the name of the sequence, then the sequence itself which may contain the gap character '-', then a coordinate) plus a consensus line (shows matches with '*', gaps with spaces, non-conserved substitutions with '!', and conserved substitutions with ':'). A sample file is provided and described by EMBL-EBI at http://www.ebi.ac.uk/help/formats.html#aln.

2.5: Third-Party Programs

Just as there are numerous formats for biological data, there are numerous pieces of software for processing and viewing that data, designed in many different programming languages and for many different operating systems, accepting different input formats and presenting different user interfaces. Some are web-based utilities while others must be installed on a local machine. Some have a graphical user interface while others are command-line programs. Some (such as BLAST) can send and receive data through 'stdin' and 'stdout' pipes without the need to create intermediate files, facilitating their interaction with custom scripts (see Section 2.6).
This section will summarise and assess several key programs used at various points in this research, and provide URLs for accessing or downloading them. They are free unless otherwise stated.

All computational work described herein was performed on computers running Windows XP with SP2.

2.5.1. ClustalW (v 1.83)

ClustalW is the main utility in the popular Clustal family of multiple sequence alignment programs (Chenna et al. 2003), available at http://www.ebi.ac.uk/clustalw. It can be used through a web interface, or downloadable command-line executables are available from ftp://ftp.ebi.ac.uk/pub/software/dos/clustalw.

ClustalW accepts a number of input formats such as FASTA, EMBL and GCG, and can output alignments in several formats as well. The .ALN format, described in Section 2.4, is the default.

By default ClustalW performs global alignments of the input sequences, but it can be set to look for local alignments instead. Alignments can be calculated by a ‘slow but accurate’ method (Myers & Miller 1988) or a ‘fast approximation’ method (Wilbur & Lipman 1983), but neither are suitable for very large amounts of sequence.

ClustalW also produces phylogenetic tree data (as .DND files) for the input sequences, and can be set to perform bootstrap analysis to determine certainty values for each branch. ClustalW does not display or manipulate phylogenetic trees, however: .DND output files must be viewed in other
programs such as the web-based Phylodendron at

2.5.2. BLAST (v 2.2.14)

The BLAST algorithm (Altschul et al. 1990) has already been described
in Section 1.2. It is useful for searching large datasets for regions of local
similarity to a query sequence (protein or nucleotide), and creates local
alignments with matching sequences. Many on-line database sites (such as
SGD and the databases at NCBI) contain a web interface for BLASTing of
their sequence data, and web-based and stand-alone versions of the BLAST

The BLAST software consists of several DOS-based command-line
executables. Key amongst these, for most purposes, are formatdb.exe,
blastall.exe and bl2seq.exe. The formatdb program accepts FASTA (or
ASN.1) protein or nucleotide sequence files and constructs a BLAST
database of those sequences, which can then be queried using blastall. The
bl2seq program directly compares two input sequences with the BLAST
algorithm, and does not require the prior creation of a database. More
complete documentation accompanies the installation files.

2.5.3. Tandem Repeats Finder (v 4.0)

Tandem Repeats Finder (Benson 1999) is a program for identifying
repeats in DNA sequences, and can find repeat units of any size even if they
are imperfect. The program, at http://tandem.bu.edu/trf/trf.html, is usable
through a web interface but there are also various downloadable executable
versions: for Windows (both standard Windows and command-line DOS versions are available) and for Linux, Solaris and Mac OS X.

Sequences must be inputted in FASTA format, and output is formatted as HTML or plain text tables which include, for each block of repeats, coordinates and statistics such as period size and copy number, nucleotide composition, match and indel percentages (which for imperfect repeats are less than 100%), and a measure of the entropy of the repeat. HTML tables include, for each repeat block, a hyperlink to a secondary output file that displays the sequence and structure of the repeat.

2.5.4. Dotter

‘Dotter’ (Sonnhammer & Durbin 1995) is a command-line-driven DOS program (Linux and Mac OS X versions are also available) which generates dot matrix plots (see Figure 1.1) using a preset sliding window (such that matches smaller than the window size are ignored) to graphically represent similarity between two DNA or protein sequences, or self-similarity within a single sequence. Though now somewhat antiquated it is still available from http://www.cgb.ki.se/cgb/groups/sonnhammer/Dotter.html. The command syntax and option flags are described at this address.

The program accepts either plain text or FASTA formatted sequences, and in the latter case multiple sequences can be included and will be plotted along the same axis, delimited by green borders. Entering the same file on the command line as both query and subject sequence will result in a self-similarity plot. As the length of the input sequences increases linearly (n) the calculation time increases exponentially (n^2), so it is impractical for use with
very large sequences even though the only practical limit is available memory. Nevertheless it remains a useful visualisation tool for exploring sequence duplication and repetition. Small peaks, and large, can be filtered out by variable noise cutoff parameters and a grey colour ramp represents the spectrum between maximum and minimum cutoff values and can be varied in real time once the plot has been calculated and displayed. A crosshair allows selection of specific alignments within the plot, the sequence of which can be viewed in a separate window.

The program allows users to colour-code specific subsequences, to indicate features for example, by loading annotations in a Dotter-specific format (described at the web address above). A parameter-driven zoom feature is documented but certain operating system parameters must be manually adjusted in order for it to work. Dotplot data files can be saved and revisited, and can be calculated as a background process. The only way to capture output as an image is by screen capture. **Figures 4.1 and 4.2** shows processed examples of Dotter output.

### 2.5.5. mVISTA

mVISTA is a web-based alignment visualisation program, part of the VISTA suite of genome-analysis programs (Frazer *et al.* 2004) at [http://genome.lbl.gov/vista/index.shtml](http://genome.lbl.gov/vista/index.shtml). It generates a graphical plot of identity along the length of two or more collinear sequences, as calculated using a sliding window.

mVISTA accepts uploaded multiple sequence data in the FASTA format, or will automatically acquire sequence if given a GenBank accession
number. Annotation data can also be uploaded, either in a custom text format (described at http://genome.lbl.gov/vista/mvista/instructions.shtml) or in GFF format.

The sequences are aligned using AVID (Bray et al. 2003), which is similar to BLAST but performs global alignment, or using LAGAN or Shuffle-LAGAN (Brudno, Do et al. 2003; Brudno, Malde et al. 2003).

The resulting plot can be viewed interactively in the site's Vista Browser or can be downloaded as a pdf file. Figure 5.2 was produced using mVISTA (with some post-processing of the pdf vector file).

2.5.6. BioEdit (v 7.0.5.2)

BioEdit (Tom Hall 1999; no journal citation available) is a useful editor for aligning multiple protein or nucleotide sequences. It is a graphical Windows program written in C++. It is no longer reliably maintained or documented but is still available at http://www.mbio.ncsu.edu/BioEdit/bioedit.html. It interacts with several other programs including ClustalW (providing the correct version of ClustalW is procured), which can be used to generate automatic alignments, but BioEdit also provides tools for manually editing alignments (including ‘drag and slide’ tools) where ClustalW does a poor job. Different residues can be colour-coded as the user desires: a facility that greatly assists sequence visualisation and manual alignment. BioEdit accepts a number of different input formats, including sequence saved to the Windows clipboard.
2.5.7. Mega 3.1

Mega 3.1 (standing for Molecular Evolutionary Genetics Analysis; Kumar et al. 2004) is the latest iteration of a free software suite, available at http://www.megasoftware.net. It can be used to perform manual sequence alignments (but is less intuitive than BioEdit) or automatic sequence alignment (using ClustalW), and can be used to query and BLAST the GenBank database. It is particularly useful for generating and bootstrapping phylogenetic trees from multiple alignments (using a variety of methods including neighbour-joining, minimum evolution and maximum parsimony, and a variety of measures including P-distance, Jukes-Cantor and Kimura 2-parameter scores), and for displaying and manipulating those trees.

Figures 4.8 and 5.3 were produced using this program.

2.5.8. Weblogo

Sequence logos (Schneider & Stephens 1990) are an intuitive way to display protein-binding DNA motifs where different residues contribute to the binding to different degrees. Weblogo (Crooks et al. 2004), at http://weblogo.berkeley.edu/, is a web-based utility for generating logos as bitmap or vector files. The program constructs the logo from multiple pre-aligned homologous sequences, which must be inputted in FASTA, ClustalW ALN, or plain text format (in the latter case, one sequence per line and all sequences must be of the same length). Figure 1.2 contains sequence logos generated by Weblogo.
2.5.9. SPSS 13.0

SPSS (Statistical Package for the Social Sciences; http://www.spss.com/spss) is a commercial statistical analysis package widely used in the sciences. Resembling a spreadsheet program such as Microsoft Excel, SPSS accepts tab-delimited input files, Excel files, and various other formats. It contains functions for nearly any conceivable form of statistical analysis, and was used in this study to perform independent-sample T-tests and Kruskal-Wallis H-tests on the results of the genome-wide analysis.

SPSS can also be used to create a wide variety of visual displays of datasets, and Figures 4.3 to 4.7 were produced in this program.

2.6: Genome-Wide Analysis Scripts

The author is not trained in computer science or software design, but has taught himself during the course of this doctoral work. The scripts written for this work were never designed for general public use, and would undoubtedly appall a trained programmer. Nevertheless they were a necessary step in processing the large amounts of genomic data under consideration, and were perfectly successful to that end. The author initially began writing code in Java but soon found that language unproductive, so moved instead to programming in Python: an infinitely preferable experience.
2.6.1. The Python programming language.

Python (http://www.python.org) was created by Guido van Rossum in 1990, based on the design principle of a clear and elegant human-readable syntax.

One of the most distinctive but contentious features is the use of ‘white space’ (invisible ‘space’, ‘tab’ and ‘return’ characters) as an organisational principle. This contrasts with most other languages (like C++ and Java), in which blocks of code must be surrounded by specific delimiters (in C++ and Java, the brace characters ‘{’ and ‘}’). In these other languages white space, though optional and encouraged as a matter of style and clarity, is functionally ignored. In Python blocks of code are distinguished by white space indentation and, because the language enforces the use of white space in a systematic way, program structure is invariably much clearer to the naked eye, and errors due to missing or incorrect block delimiting are almost unknown.

Python is an interpreted language, like Java. Scripts can be run on any computer that has the Python language and interpreter (which are packaged together) installed on it. Versions of Python are available for all major operating systems, and the same script will work on any Python interpreter (of the same version, and assuming that no operating system-specific methods are present in the script) so that Python scripts are effectively highly cross-platform compatible.

Python is dynamically typed, meaning that variable types are determined when the program is run, rather than when the program is initially compiled, and it is ‘duck typed’, in that variable types need not be explicitly
stated prior to assignment. At run-time Python determines what the variable types must be from the contents of the variable.

Unlike lower-level languages such as C++, Python has automatic memory management and ‘garbage collection’ of variables that are no longer needed, so that programmers do not need to explicitly destroy each and every variable that is no longer required. This reduces the complexity of the code, and reduces the chances of introducing memory leaks and similar bugs.

Python is an object-oriented language, but not rigidly so. Whereas in Java every function must be wrapped in an object class, in Python individual expressions can be run on their own. Python’s interpreter is also interactive, so that individual lines of code can immediately be run and tested at the interpreter’s command prompt. Indeed, because it is an interpreted language, any script can be run and tested at the touch of a button, without the need for formal compilation steps (in which a script is compiled into machine code suitable to the hardware).

These factors combine to make Python easy to learn and easy to use. Scripts can be written quickly, and tested at any stage to ensure that the desired behaviour is forthcoming. The clear syntax makes scripts relatively concise and relatively easy to understand, debug, and expand. Contrast the two simple ‘hello world’ programs in Figure 2.1.

```
public class HelloWorld
{
    public static void main (String args[])
    {
        System.out.println("Hello World!");
    }
}
```

```
print "Hello World!"
```

Figure 2.1. The ubiquitous minimal “Hello World” program in Java (left) and Python (right). Python is far easier to learn and far more productive.
Nevertheless Python is a serious language, particularly popular where rapid application development and prototyping is called for, or for parts of code that require regular maintenance or upgrading. Google, NASA and Industrial Light & Magic all use Python in parts of their IT developments, and the language continues to grow steadily in popularity.

In terms of pure performance, while obviously not as fast or memory-efficient as C or C++ due the overhead of running through an interpreter, Python is in many circumstances as fast as Java or Perl and in some functions faster. It is also generally more memory-efficient than either Java or Perl (Prechelt 2000). This is because most of Python is actually written in C++, and is in a sense a user-friendly interface for that language: the basic data types and operations in Python are actually routed to fast C++ modules.

The core of the Python language is elegant and quite minimalist, but Python ships with a large ‘standard library’ of importable modules (many written largely in C++) to extend the basic functionality. The ‘re’ module, for example, provides functions for regular-expression searching. The ‘math’ module provides operations such as log and sine. The ‘random’ module contains functions for generating pseudorandom numbers and choices.

Because Python is open-source, there are also many external modules that can be downloaded and installed to extend the capabilities of Python. The ‘SciPy’ module (http://www.scipy.org) is a useful example. It contains advanced datatypes such as multi-dimensional arrays, and advanced algebraic and statistical functions of use to the scientific community.
2.6.2. BioPython.

Another external open-source module, developed by some members of the biology community, is BioPython (http://biopython.org) which contains various bioinformatics tools and types for use in Python scripts. For example, it includes data types for DNA and protein sequences, and some functions for manipulating them. There are various parsers for reading different biological file formats (see Section 2.2) and for interfacing with biological databases such as GenBank and programs such as BLAST.

BioPython currently suffers from a number of drawbacks that have limited its use in this work. It lacks functions that were required for the genome-wide analysis described in Chapter 4. The module is large, poorly structured and weakly documented. It sacrifices clarity for versatility and scope, with the result that types and functions are wieldy and difficult to understand, use or modify for more specialist purposes.

Consequently, in the Python scripts described below, BioPython was used only sparingly (such as to interact with BLAST or parse GenBank files): it was easier to design key data types and functions from scratch than to modify the BioPython equivalents, and new code was not written so as to fit into the BioPython framework.

2.6.3. Genome-wide analysis with Python.

Python scripts were developed, using ActivePython (http://www.activestate.com/Products/ActivePython) as a development environment, to perform the genome-wide analysis described in Chapter 4. The code went through many iterations over a period of years, from early
versions written for a single narrow purpose to more modular versions intended to be applicable to a wider range of problems.

The general principle behind the genome-wide analysis was as follows:

**Step 1**: input sequence and annotation data from SGD and the Broad Institute. The genome-wide analysis was restricted to close relatives *S. cerevisiae* (SGD) and *S. paradoxus, S. mikatae* and *S. bayanus* (Kellis et al. 2003). The sequences produced by Cliften et al. (2003) were not used for this analysis as they lacked annotations.

**Step 2**: annotate the 'intergenic' spaces as features in their own right.

**Step 3**: determine which genic and intergenic features have unambiguous homologues in the other species, and the coordinates of those homologous subsequences.

**Step 4**: perform pairwise alignments between homologous features in each pair of species for which homologues of that feature exist, and measure differences from those alignments.

ClustalW was used as an accessory application to perform the alignments. There are frequent differences in the position of start and stop codons in the four *Saccharomyces* species. ClustalW performs global alignments (as opposed to the local alignments done by BLAST), and so gaps at either end of the clustalW alignments are indicative of start/stop codon changes. These gaps can be ignored so that measures of identity are confined to subsequences that are unambiguously homologous and of the same type (coding / non-coding).

Two versions of the genome-wide analysis scripts, an early version designed for Python v2.3 and a redesigned modular version for v2.4, are
described below and included on the CD-ROM accompanying this thesis (see Appendix I). At the time of writing, a new version of Python (v2.5) has been released. Earlier versions of script should be compatible with later versions of Python but the reverse is not true.

2.6.4. Genome-wide analysis script, early version ("GWA early.py").

The script expects the following input files to be present in the same directory as the script: “ScerSGD.fas”, “SparMIT.fas”, “SmikMIT.fas” and “SbayMIT.fas” are FASTA formatted files containing the DNA sequences for each genome; “chromosomal_feature.tab”, “Spar_forward_full.txt”, “Smik_forward_full.txt” and “Sbay_forward_full.txt” contain the corresponding feature details as provided by SGD or Kellis et al. (2003).

The script also interfaces with clustalW.exe, which must also be in the same directory as the script. Note that the “DOS” version of clustalW.exe does not work properly in a Windows XP environment. The script and all necessary input files (2004 versions) and the XP version of clustalW.exe are collected together in the same directory on the CD-ROM accompanying this thesis (see Appendix I). However the script will not run directly from CD-ROM as it attempts to save its data in the same directory, so the entire folder must be copied intact to hard disk.

The script breaks the analysis down into steps, and as each step is completed the data generated so far is saved and a checkpoint value recorded. In this way, if an error occurs in any later phase, causing the program to terminate, it can be restarted from the last checkpoint (assuming that the error has been recognised and rectified) without having to repeat the
earlier steps. To repeat the entire process from the start, the saved files must be removed from the script directory. The steps are as follows.

**Step 0: load sequence information from file, and save each genome as a Python ‘shelf’ object.** The script begins by instantiating a single `FeatureDatabase` object, which in turn instantiates its own `SequenceDatabase` object.

For each of the names listed in `SequenceDatabase.names`, a `SequenceGroup` object is instantiated, each of which creates a `FastAParser` object that reads the appropriate source file and converts each chromosome or contig in the file to a separate `SequenceObject`, which is stored as part of the `SequenceGroup`. Each `SequenceGroup` also records and reports the number of separate sequences and total number of bases that it has loaded, and when complete it saves itself as a Python ‘shelf’ file, in which the structure and data of the object are stored as a file which can then be accessed and altered like a Python dictionary, but stored on disk rather than in RAM. This is, in effect, a form of non-relational database, found to be much faster than interfacing with an SQL database using Python.

**Step 1: load and save *S. cerevisiae* feature details.** The `FeatureDatabase` object instantiates a `FeatureFileParser` to read the feature information for each *S. cerevisiae* feature in “chromosomal_feature.tab”. The parser returns each feature as a `FeatureObject`, which contains the coordinates and other information.

Each new `FeatureObject` is instructed to check itself against the `SequenceDatabase` to ensure that, if it is a gene, it begins with an ATG start codon. If it does not, the script reports the fact but continues anyway. *S. 
cerevisiae S288C contains a few genes that do have unusual start codons, but if a large number are reported then there is an error in the feature annotation file (in all likelihood the feature annotations and sequence data are different versions).

For each new FeatureObject, a FeatureComparison object is instantiated. This will eventually store information about homologues in the other genomes.

The FeatureDatabase then creates a FeatureSequence object for each S. cerevisiae chromosome stored in its SequenceDatabase, and adds each feature to the appropriate FeatureSequence. Each FeatureSequence also records the total length of its chromosome, and the position of the centromere. The FeatureDatabase saves itself (and all its constituent objects and data) at this checkpoint, as it will at the end of all subsequent steps.

**Step 2: annotate S. cerevisiae intergenic regions as separate features.** The FeatureDatabase instructs each FeatureSequence to use its makeIntergenics method. This calculates the coordinates of each gap between features, and stores the result as a FeatureObject, of type “intergenic” and name “L>R” where L is the id of the left-flanking feature and R is the id of the right-flanking feature. In this implementation of the script this is a slow process that can take 5-10 minutes to complete.

**Step 3: load other genome feature files, create FeatureSequence objects for each contig, and calculate intergenics for each.** The FeatureDatabase now loads feature details for the other genomes (from “Spar_forward_full.txt”, “Smik_forward_full.txt” and “Sbay_forward_full.txt”), in each case using a FeatureFileParser with a flag telling it to recognise the
Broad Institute format instead of the SGD tab format. As in step 1, the start
codon of each ORF is checked as that FeatureObject is created.

For each new FeatureObject, the FeatureDatabase looks for an
existing FeatureComparison object with the same name, and if found, the new
FeatureObject is added to the FeatureComparison.

FeatureSequence objects are created for each contig of each genome,
and loaded with the appropriate features. These FeatureSequence objects
are then instructed to determine, on the basis of their feature annotations, if
they are reversed with respect to the homologous sequence in S. cerevisiae.
Thereafter each FeatureSequence calculates its own intergenic regions
(where possible) and adds them as FeatureObjects.

Step 4: calculate distance from centromere for each
FeatureComparison. Absolute distances, in bp (negative if to the left of the
centromere) and relative distances (as a proportion of the chromosome arm
length) are calculated, based on the average of the feature's coordinates, and
stored in each FeatureComparison object.

Step 5: for each intergenic FeatureComparison, perform clustalW
alignment between each pair of features and record the numerical
results. For each FeatureComparison of the "intergenic" type, the
doComparisons method is called. This in turn calls the doClustalWAlignment
method for each pair of features in the FeatureComparison. For each in turn,
the relevant DNA sequences are retrieved from the SequenceDatabase, and
saved together in FASTA format as a 'temp' file in the script directory. The
script then instructs clustalW.exe to align the sequences in the temp file,
maintaining the same order as the original, and clustalW.exe produces output
as ‘temp.aln’ file and an unused ‘temp.dnd’ file. The doClustalWAlignment method parses the ‘temp.aln’ output file and calculates the lengths, number of mismatches, number of gaps etc, storing these numbers in the FeatureComparison object. The ‘temp’, ‘temp.aln’ and ‘temp.dnd’ files are deleted after each alignment has been parsed.

**Step 6: perform clustalW alignment for each gene**

*FeatureComparison.* Step 5 is repeated, but for each FeatureComparison of the “gene” type. Both these steps are extremely time-consuming, taking >72 hours on an average desktop computer.

**Final step: query ‘shelf’ files and extract relevant data.** When the FeatureDatabase has completed, or loaded, it is passed to an OutputTable object, which extracts the desired data from each FeatureComparison object and saves it all as a tab delimited file for analysis in a program such as SPSS.

2.6.5. Genome-wide analysis script, redesigned version (“GWA redesigned.py”).

The previous script was sufficient for its intended purpose, but was poorly designed and inflexible. A new set of code was created from scratch, taking advantage of the author’s improved understanding of Python, and new features introduced to Python version 2.4 such as conditional expressions.

The new code, which went through several iterations, was more modular in design, in the hope of creating a versatile toolbox of the data types, functions and classes that could be applied to a wider variety of tasks.
One of the required capabilities, for example, was the ability to split intergenic regions containing centromeres into intergenic regions on either flank of centromeres, plus the centromeres themselves. This made it necessary to reannotate centromeres, since the SGD annotations are based on experimental footprinting methodologies (Espelin et al. 1997; Pietrasanta et al. 1999) rather than sequence homology, and subsequences derived from those coordinates do not align properly.

New classes were written to interact with programs such as BLAST (using BioPython), ClustalW and Tandem Repeat Finder, and to search DNA sequences using regular expressions. New parser classes were written for formats including FASTA, GFF3, Broad Institute forward feature files, Tandem Repeat Finder output files, and ClustalW .ALN output files. The steps of the earlier script were reconceived as separate classes (AnnotateInterstices, AnnotateOverlaps, AnnotateCentricDistance and so on) acting on basic data types that included DNA, Protein, Feature, Homology and Genome.

Rather than store data in Python 'shelf' files, the decision was taken to store results in a human-readable file format: the increasingly popular GFF3 format. The GFF3 parser was extended to be able to write this format as well as read it.

Simple functions were then written to link these units together into a processing pipeline. The resulting script (included on the CD-ROM; see Appendix I) interacts with formatdb.exe, bl2seq.exe (both BLAST programs), clustalW.exe, and trf400.dos.exe (Tandem Repeat Finder), which must be in the same directory as the script. It also depends on secondary functions imported from roman.py and toolbox.py, which should also be in the same
directory, as should a file called 'essential.txt' which lists the genes annotated with an inviable null mutant phenotype in SGD.

The code is fairly complex but its function in the genome-wide analysis can be broken down into the following basic steps:

**Step 1:** for each genome, retrieve source data directly from on-line sources (such as the "saccharomyces_cerevisiae.gff" file from SGD which combines all feature annotations and all genome sequences in a single file. This resource was made available for the first time in 2005), parse, and store in memory as a *Genome* object, with nested *DNA* and *Feature* objects.

**Step 2:** convert chromosome and contig names into a standardised format of the form "SPEC.[CH].[CONT].[rc]" where SPEC is a four letter species designation ('S cer', 'Spar', 'Smik' or 'Sbay'), CH is the two-digit zero-padded chromosome number if known, and CONT is the four-digit zero-padded contig number if relevant (e.g. "Scer.01.", "Smik..1291." etc).

**Step 3:** for the Broad Institute sequences: determine where possible the chromosome assignment of the contig, based on homology to *S. cerevisiae*, and amend the sequence designation accordingly (e.g. "Smik..1291." is unambiguously homologous to *S. cerevisiae* chromosome I and so becomes "Smik.01.1291.").

**Step 4:** for Broad Institute sequences: reverse complement all contigs where necessary to put them in the same orientation as the homologous *S. cerevisiae* sequence. Append 'rc' to the sequence designation where this has occurred. Revise feature coordinates accordingly (e.g. "Sbay.02.0007." becomes "Sbay.02.0007.rc" and YBR117C coordinates on that 40427 bp contig are adjusted from 34287-36332 to 6141-4096).
Step 5: reannotate point centromere motifs, identified using regular expressions (the pattern used is described in Chapter 4).

Step 6: hyperannotate each genome with additional features such as intergenic regions, and tandem repeats (determined using Tandem Repeat Finder), and hyperannotate all features with additional information such as the names and types of overlapping or embedded features, the distance from the centromere, and whether genes are annotated as essential or not (from a separate SGD-derived file).

Step 7: save each genome as a tab-delimited human-readable GFF3 file, with additional information stored in customised attribute fields.

Step 8: in a separate final step (the do_comparisons function), perform pairwise comparisons using ClustalW, taking each pair of genomes in turn. Where ORFs or intergenic regions are unambiguously homologous, hyperannotate the feature in both genomes with a homology attribute that includes the coordinates of the homologous sequence (e.g. The S. paradoxus homologue of S. cerevisiae YBL033C is at "Spar.02.0365.rc+17129+16092"), as well as calculated measures including P-distance, Jukes-Cantor score and Kimura 2-parameter score ("pd=0.105009633911", "jc=0.113128369607", "k2=0.115374041972"). Re-save the data in GFF3 format, now including the homology attribute.

Querying of the data set can be accomplished by reloading the GFF3 file (e.g. "Sbay = reload("Sbay_complete.gff")"), which parses it back into Genome, DNA and Feature objects in memory.

Feature subsets can then be retrieved by various parameters (e.g. "centromere_set = Sbay.features[type = "centromere"]" or "gene_set =

86
Sbay.features[type = "gene"], and individual feature details can be retrieved using dictionary notation (e.g. "print feature["homology"]", or "print feature["overlaps"]") or as a general report ("print feature.report()”).

More complex queries can be constructed using Python 2.4’s conditional expressions (e.g. “set_of_genes_shorter_than_200_bp = [ feature for feature in Sbay.features["type = gene"] if len(feature) < 200"]”).

2.6.6. Future code: graph-based data structures?

While the redesigned code is certainly a great improvement, there are almost certainly more elegant and versatile solutions that may be explored in the future. Following the discussions in the IT community of new concepts such as the ‘semantic web’ and metadatabase systems such as WinFS (originally intended as the basic file store system for Windows Vista but later removed), a powerful and versatile graph-based data structure may be optimal, in which each individual piece of information would be ‘tagged’ with attributes and metadata that define what that information is and how it may be used. A chromosome sequence, for example, might be tagged with:
type=DNA, species=S. cerevisiae, strain=S288C, chromosome=chr I, date=9.9.06, url=... etc, and new metatags or tags could be added whenever needed. Each new metatag or tag would be a unique object that keeps track of all objects tagged with it. To be useful, such a data structure would be storable in a persistent state (saved to disk as a form of database) but in a manner that allows for fast querying.
2.6.7. Data analysis in SPSS.

For statistical analysis, tab-delimited tables were prepared, with each row representing a single region in a single pairwise comparison. Region details, including coordinates, homology scores, centromeric distances, and regional category types were laid out in each row.

To best compare like with like, intergenic regions containing transposons, ARSs or RNAs in *S. cerevisiae* were filtered out into separate regional categories.

These tables were imported into SPSS, and categories compared using Independent-sample T-tests and Kruskal-Wallis H-tests to determine whether different categories (e.g. intergenic regions containing centromeres, and other intergenic regions) showed statistically different distributions.

2.7: Miscellaneous work

Although most of the work described in this thesis was conducted *in silico*, using pre-existing data, a small amount of additional DNA sequencing was conducted, as well as some simple laboratory tests relevant to Chapter 5.

2.7.1. Strains.

Table 2.1 lists the strains studied in this work. Most of these were the subjects of targeted sequencing of specific centromeric or subtelomeric regions. These were proposed as suitable candidates for this purpose by
Gianni Liti, and were taken from his collection (in the table his strain designation numbers are preceded by 'GL'). This is a subset of the larger collection of Professor E. Louis.

Information concerning the provenance of individual yeast strains is often difficult to obtain, sometimes entirely absent and occasionally even conflicting. The provenance information stated in Table 2.1 is for the most part conflated from various sources, including the records of Gianni Liti and on-line databases of various collections listed in the next section.
### A: Main sequenced yeast species / strains relevant to this work

<table>
<thead>
<tr>
<th>SPECIES</th>
<th>STRAIN DESIGNATIONS</th>
<th>PROVENANCE</th>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae</td>
<td>S288C / GL96</td>
<td>See Mortimer &amp; Johnston (1986) SGD curated sequence</td>
<td></td>
</tr>
<tr>
<td>S. mikatae</td>
<td>IFO1816 / DBVPG7270 /NCYC2970 /GL19</td>
<td>Japan – decayed leaf (Naumov et al. 2000) Sequenced by the Broad Institute. See Section 1.1</td>
<td></td>
</tr>
</tbody>
</table>

### B: Diverse strains with partial centromeric or subtelomeric sequencing (Chapters 4 & 5)

<table>
<thead>
<tr>
<th>SPECIES</th>
<th>STRAIN DESIGNATIONS</th>
<th>PROVENANCE</th>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae</td>
<td>CBS405 / DBVPG6044 /IFO0258 /NRRLY1546 /ATCC10604 /GL60</td>
<td>West Africa, bili wine 1914/1925 (Guitlermond 1914, Naumov et al. 1993)</td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>DBVPG6763 / GL1</td>
<td>Indonesia, lici fruit (A. Vaughan-Martini) (aka S. boulardii)</td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>SK1 / M272 /NCYC1630 /GL17</td>
<td>USA - Pennsylvania, soil beneath Q. alba (Sniegowski et al. 2002) Homothallic diploid, killer strain</td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>YPS128 / GL104</td>
<td>USA - Pennsylvania, soil beneath Q. alba (Sniegowski et al. 2002)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>CBS2980 / DBVPG6037 /GL33</td>
<td>Netherlands - Delft (A. Vaughan-Martini)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>CBS5829 / DBVPG6466 /NRRLY17218 /GL98</td>
<td>Denmark - soil</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>DBVPG6303 / UCD-52-153 /GL31</td>
<td>USA - Davis (A. Vaughan-Martini)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>DBVPG6304 / UCD-51-186 /GL32</td>
<td>USA - Davis (A. Vaughan-Martini)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>DBVPG6565 / GL5</td>
<td>French mayonnaise in USA (A.Vaughan-Martini) (aka S. douglasii)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>DG1768 / GL8</td>
<td>(D. Garfinck) No transposons (G. Liti)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>IFO1804 / GL137</td>
<td>Japan – bark</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>N17 / NA-2C / ATCC96981 /GL26</td>
<td>Russia - Tartarstan (Naumova et al. 2003) (aka S. mangini var. tetrasporus)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>N43 / CBS8437 / GL76</td>
<td>Russia</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>N44 / CBS8438 / GL77</td>
<td>Russia - Temel, 1987 (Naumov et al. 1993)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>Q4.1 / GL36</td>
<td>London - Windsor (A. Burt)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>T21.4 / GL40</td>
<td>London (A. Burt)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>UFRJ60791 / GL20</td>
<td>Brazil - Catalao point, Drosophila (E. Louis)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>UFRJ50816 / GL21</td>
<td>Brazil - Tijuca Forest, Drosophila (E. Louis)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>YPS125 / GL114</td>
<td>USA - Pennsylvania, flux of Q. rubra (Sniegowski et al. 2002)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>YPS138 / GL115</td>
<td>USA - Pennsylvania, soil beneath Q. velutina (Sniegowski et al. 2002)</td>
<td></td>
</tr>
<tr>
<td>S. paradoxus</td>
<td>YPS145 / GL116</td>
<td>USA - Pennsylvania, soil beneath Q. velutina (Sniegowski et al. 2002)</td>
<td></td>
</tr>
</tbody>
</table>

### C: Additional strains used in killer toxin study (Chapter 5)

<table>
<thead>
<tr>
<th>SPECIES</th>
<th>STRAIN</th>
<th>PROVENANCE</th>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. cerevisiae</td>
<td>YDG605 / YDB2</td>
<td>Y55 background haploid. Provided by Duncan Greig.</td>
<td>type 2 killer strain, mating type a, leu2, ura3-1</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>YDG627 / YDB1</td>
<td>Y55 background haploid. Provided by Duncan Greig.</td>
<td>type 1 killer strain, mating type a, leu2, lys2-</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>YEL2036 / YDB3</td>
<td>Y55 haploid, Ed Louis collection</td>
<td>mating type a, ura3-1</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>YEL2037 / YDB4</td>
<td>Y55 haploid, Ed Louis collection</td>
<td>mating type a, his-1</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>KILLER1 / YDB3x1</td>
<td>Y55 diploid, created by the author</td>
<td>type 1 killer diploid</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>KILLER2 / YDB4x2</td>
<td>Y55 diploid, created by the author</td>
<td>type 2 killer diploid</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>CONTROL / YDB3x4</td>
<td>Y55 diploid, created by the author</td>
<td>non-killer control diploid</td>
</tr>
</tbody>
</table>

Table 2.1. Yeast strains used in this study.
2.7.2. Major yeast strain collections.


- Dipartimento di Biologia Vegetale e Biotecnologia Agroambientale (DBVPG, Italy) industrial yeast collection, searchable at http://www.agr.unipg.it/dbvpq/online.html.

- Institute for Fermentation (IFO, Japan) yeast collection, now held at the NITE Biological Resource Center (NBRC), searchable at http://www.nbrc.nite.go.jp/NBRC/NBRCDispSearchServlet?lang=en.

- National Collection of Yeast Cultures (NCYC, UK), searchable at http://www.ncyc.co.uk/search.php.

- Northern Regional Research Laboratories (NRRL, USA) yeast collection now held in the Agricultural Research Service Culture Collection, searchable at http://nrrl.ncaur.usda.gov/cgi-bin/usda/process.html?id=8FMBK8Ms.

- American Type Culture Collection (ATCC, USA), searchable at http://www.lgcpromochem-atcc.com/common/catalog/fungiYeast/fungiYeastIndex.cfm.

- Collection de Levures d'Intérêt Biotechnologique (CLIB, France), searchable at http://genome.jouy.inra.fr/clib/bdd/.

- Microbiology Collection of Yeast Cultures (MCYC, Spain) now held at La Colección Española De Cultivos Tipo, searchable at http://www.cect.org/english/consf.htm.
• Herman J Phaff yeast culture collection at University of California, Davis (UCD, USA), searchable at

2.7.3. Sequencing.
Sequencing was performed by the commercial company “AGOWA Gesellschaft für molekularbiologische Technologie mbH” (in the UK: “AGOWA Genomics”: http://www.agowa.de/). DNA preps for each strain to be sequenced were kindly provided directly to AGOWA by Gianni Liti, and AGOWA designed their own primers based on the specifications provided by the author, i.e. the coordinates of the homologous sequence in the S. cerevisiae genome.

The raw files produced by AGOWA are included on the CD-ROM accompanying this thesis (see Appendix I). The '.abi.seq' files contain the raw sequence of each read in plain text format. For each sequenced region in each strain, two reads were produced, one in each direction. In each case, the bidirectional reads were trimmed and combined using BioEdit. There were no cases in which the bidirectional reads disagreed about a base call within a high-quality section of both reads. The resulting combined reads for each sequenced region are also included in the CD-ROM, along with FASTA files in which the combined and raw reads are placed side by side for comparison.
2.7.4. Testing of killer toxin resistance.

In order to test the toxin resistance of various strains of *S. cerevisiae* and *S. paradoxus* (see Chapter 5), Duncan Greig at University College London kindly provided two haploid Y55-based killer strains (YDG605 and YDG627). To make these diploid, they were crossed with complementary Y55 haploid strains from the collection of Professor E. Louis (prefix YEL).

YDG627 (*leu2*- *lys2*-, mating type α) was crossed with YEL2036 (*ura3*-1, mating type α) to produce a diploid type 1 killer strain, and YDG605 (*leu2*-, mating type α) was crossed with YEL2037 (*his1*-1, mating type α) to produce a diploid type 2 killer strain. Both crosses were performed on minimal media agar plates, to select against haploid auxotrophs. This media was also low pH, to activate killer toxins and select against cells losing the killer virus.

Low pH media is made with 400 ml ordinary growth media (minimal or YEPD) + 100 ml phosphate citrate buffer (consisting of 21.1 g citric acid, 28.4 g K$_2$HPO$_4$, 100 ml water) + ~5 g methylene blue. The buffer can interfere with agar setting, in which case the proportion of agar can be increased accordingly.

To provide a diploid non-killer control strain, YEL2036 and YEL2037 were also crossed on ordinary minimal media agar plates.

A quick and simple procedure was adopted to test resistances. Each strain was grown up in 30 ml liquid YEPD media (low pH in the case of the killer diploids) at 30°C for 24 hours. For each of the test subjects, 5 ml of the liquid culture was spread on a low pH YEPD agar plate and dried. The two killer strain liquid cultures and the control strain liquid culture were centrifuged to a soft pellet and drained, and a portion of each cell mass transferred to a
separate labelled area of each test plate with a clean blunt-ended wooden dowel. The test plates were incubated for 24-48 hours at 30°C, until the presence or absence of a clear halo around the killer colonies was obvious.

A clear halo indicates a lack of test strain growth in proximity to a killer strain, due to the test strain's susceptibility to the secreted killer toxin which defuses out through the agar.
3.1: Introduction

Prior to the discovery of centromere-adjacent supervariation (Chapter 4) and hybrid introgression between *S. cerevisiae* and *S. paradoxus* (Chapter 5), the early stages of this doctoral work were characterised by a period of broad and free-roaming exploration of a number of matters of interest to the author. Indeed without such a period it is doubtful that the main themes of this thesis would have come to light.

This chapter will briefly summarise some of that early work, some of which came to nothing, some of which was superceded by the publication of other groups’ research, and some of which was conducted as a service to the scientific community.

3.2: The Origins of Y’ Elements

As described in Section 1.5, Y’ elements are well-defined telomere-associated features, 4.7-6.9 kb in length, with some unusual properties. They sit within or adjacent to telomeric repeats at many but not all chromosome ends, sometimes in multiple copies. They contain their own autonomous replication sequences (ARSs), and have been detected as circular extrachromosomal elements, although whether this state is an essential part of Y’ propagation or merely a messy by-product remains unclear. They encode (or have in the past encoded) a DEAD/DEAH-box ATP-dependent RNA helicase, and in the absence of telomerase activity (perhaps causing a
general loss of telomeric silencing) these proteins are highly expressed. Some telomerase-deficient cells can maintain telomere length and function through the rapid amplification of Y' elements (perhaps by Y' helicase-mediated homologous recombination; this 'type 1 ALT mechanism' is dependent on Rad52p activity).

The similarity between Y' helicases and translation initiation factors Tif1p and Tif2p means that Y'-like sequence can be found in many organisms, but the sequence downstream of these ORFs is more distinctive. This sequence produces BLAST hits in the genomes of S. paradoxus CBS432, S. paradoxus UFRJ50791 (S. cariocanus), S. mikatae IFO1816, and S. kudriavzevii IFO1802 amongst others, and they can be detected in these species through hybridisation as well. On the basis of BLAST, and also hybridisation (G. Liti, personal communication), S. bayanus does not host these elements.

The origins of these fascinating elements remains unknown, but their activity seems almost parasitic. Therefore the author was driven to investigate whether or not there was any similarity between Y' elements and other 'selfish' DNA elements in yeast, such as plasmids, viruses or transposons.

3.2.1. The 2 micron plasmid.

The 2 micron (2µm) plasmid is a 6.3 kb extrachromosomal element found in most strains of S. cerevisiae. There are typically 40-60 copies in a single cell, which are clustered together in a single group in the nucleus. Each plasmid contains an origin of replication, and four genes: FLP1 which encodes a site-specific recombinase enzyme, REP1 and REP2, the proteins of which
regulate transcription of *FLP1* and autoregulate their own transcription, and *RAF1* which encodes an anti-repressor of *FLP1*.

The *FLP1* protein (Flp1p) promotes recombination between two Flp recombination target (FRT) sequences that are ~600 bp inverted repeats. When this occurs during replication of the plasmid it forces both replication forks to travel in the same direction around the plasmid, generating multiple copies from the single replication origin (Futcher 1986; Volkert & Broach 1986). FLP/FRT recombination has widely been exploited in the targeted transformation of organisms, particularly plants, and as a means of controlling gene expression (e.g. Huang *et al.* 1991; Lyznik *et al.* 1996).

Other circular self-replicating extrachromosomal elements such as rDNA circles or ARS-bearing recombinant plasmids are thought to diminish the lifespan of yeast cells, because unequal partitioning leads to an unhealthy accumulation of these elements in mother cells (Sinclair & Guarente 1997; Falcon & Aris 2003), but 2 micron plasmids do not appear to have this effect, perhaps because of their ability to partition equally between mother and daughter cells (Falcon *et al.* 2005), through complex regulation of Flp1p activity and the interaction between Rep1p and Rep2p and the plasmid's STB partitioning locus (Murray & Cesareni 1986).

Indeed 2 micron plasmids do not appear to confer any significant advantages or disadvantages on their hosts apart perhaps from a slight reduction in growth rates (Futcher & Cox 1983), and are commonly exploited as the basis for genetic vectors (Rose & Broach 1990).
3.2.2. Killer viruses and helper viruses.

Many *S. cerevisiae* strains contain virus-like particles containing RNA genomes (Bruenn & Kane 1978), which replicate as double-stranded RNA. The best understood of these are the M1 and M2 ‘killer’ viruses, and the L-A helper virus (Schmitt & Breinig 2002).

The L-A helper virus RNA is 4.6 kb long. It contains two ORFs, which encode two coat proteins. One is the major coat protein, ~100 copies of which make up the bulk of the viral particle. The second is present as only one copy per particle, and is actually a fusion of the two ORFs, probably through ribosomal frameshifting. It binds ssRNA and is involved in replication of the viral RNA (Icho & Wickner 1989). This virus is found in many yeast strains, but there are various mutually-exclusive allelic versions.

M1 and M2 are 1.8 kb satellite viruses, dependent on the L-A virus for their replication. Like the L-A alleles, these viruses are mutually exclusive: M1 usually outcompetes M2. Both can co-exist with any of the L-A alleles, and where there is correlation between types it is thought to be purely a matter of coinheritance. The M1 and M2 killer viruses (found in yeast killer types K1 and K2 respectively) contain a single ORF which encodes a preprotoxin. This preprotein is processed to produce both a toxin, which the cell secretes and which kills susceptible yeast cells, and a toxin resistance protein which renders a killer strain immune to its own toxin (Young & Yagiu 1978).

All these viruses are generally cytoplasmically inherited, but there is evidence that they can be horizontally transmitted in certain circumstances (el-Sherbeini & Bostian 1987).
3.2.3. Killer plasmids.

In other yeast species a similar killer phenotype is conferred by AT-rich linear DNA plasmids. In *K. lactis*, the plasmid pGKL1 (8.9 kb) encodes 4 proteins: two are toxin subunits (neither with any similarity to the virus toxin), one is a resistance protein, and one resembles a viral DNA polymerase. Much as the killer viruses rely on L-A for their replication, maintenance of the killer plasmid pGKL1 depends on the presence of the larger pGKL2 (13.4 kb), except that pGKL2 produces its own killer phenotype, whereas L-A does not.

pGKL2 contains 10 tightly packed ORFs, largely uncharacterised and with little identity. ORF2, like ORF4 of pGKL1, is homologous to the viral DNA polymerase and is the only point of similarity between the two plasmids. ORF10 may encode a DNA-binding protein. It is hypothesised that one of the other ORFs encodes the terminal protein known to bind to the 5' end of each strand (Kikuchi *et al*. 1984). Both killer plasmids have these terminal proteins (as well as inverted terminal repeats of ~200 bp at each end), and it has been proposed that pGKL2 might produce this protein for both. Thus while the killer viruses depend on L-A for replication, pGKL1 may depend on pGKL2 for end protection (Tommasino *et al*. 1988).

The *K. lactis* killer plasmids, probably of viral origin, maintain their activity when transformed into *S. cerevisiae*, but there is currently no evidence that any *S. cerevisiae* strains naturally contain killer plasmids.

3.2.4. Ty transposons.

The genome sequence of *S. cerevisiae* S288C contains ~50 retrotransposons or pseudo-retrotransposons, and many more partial
retrotransposon sequences, together constituting some 3% of the entire genome. Typically ~5.9 kb long, complete yeast transposons (Ty elements) have ~200-350 bp sequences that are repeated in the same orientation at each end (long terminal repeats or LTRs) and similar “LTR transposons” are found in many fungi, plants and animals. They are clearly derived from retroviruses, contain sequences homologous to retroviral gag-pol genes, and form virus-like particles if replicating. The TYA gene (gag homologue) encodes a coat protein which forms the bulk of the structure of the virus-like particles, while the TYB gene (pol homologue) encodes a polyprotein with protease activity that catalyses its own post-translational processing to produce subunits with reverse transcriptase, integrase and RNaseH functions.

Active transposons are transcribed (like most genes, by RNA polymerase II) as a single large transcript. In most types of Ty element (except Ty5; see below), the two ORFs overlap by a single base pair, so their translation involves a ribosomal frameshift which also regulates the relative expression levels of the two proteins (TYA proteins are ~20 times more common than the TYB polyprotein).

To retrotranspose, the RNA template is reverse-transcribed into DNA, made double-stranded, and reintegrated into the genome with the aid of TYB's integrase activity.

The Ty transposons in *S. cerevisiae* are divided into 5 classes: Ty1-5. Ty1 and Ty2 are quite similar but encode substantially different TYA coat proteins and have other differences in the TYB integrase domain (Kim *et al.* 1998). Ty1 is by far the most common transposon in *S. cerevisiae*, followed by
Ty2. Ty1, and to a lesser extent Ty2, tend to integrate close to tRNA genes, possibly by interacting with the RNA polymerase III complex.

Ty1, Ty2, Ty4 and Ty5 are similar to the copia transposons found in Drosophila, but Ty3 is similar to Drosophila's gypsy transposons. The main differences are a slightly shorter length and a different order of the integrase, reverse transcriptase and RNaseH domains in the TYB gene. Ty3 also habitually integrates upstream of tRNA genes.

Ty4 elements are poorly transcribed, perhaps because they do not integrate at specific sites, and are generally degenerate (Hug & Feldmann 1996). Ty5 elements often insert close to telomeres (Zou et al. 1995; 1996) possibly because of a preference for silenced chromatin (Zou & Voytas 1997). They are highly degenerate in S. cerevisiae, but are evidently highly active and widespread in S. paradoxus (Zou et al. 1995).

3.2.5. No clear structural or functional similarity between Y' elements and other parasitic DNA elements in yeast.

It is clear that there is little functional similarity between Y' elements and their ORFs and the other elements described above. None of the other elements encode RNA helicases, for example. Dot-matrix plots and BLAST queries show no sequence similarity at either the DNA or protein levels (data not shown, but sequences are included on the accompanying CD-ROM; see Appendix I), nor are there even any analogous non-coding structural features within Y' elements, such as direct or inverted terminal repeats.

However Y' elements are usually flanked on either side by telomeric repeats, and where they are not it may be due to subsequent sequence loss.
These repeats may perhaps function in the duplication and translocation of Y' elements, as well as the formation of circular extrachromosomal forms.

3.2.6. Links between Ty transposons and telomere-associated elements.

Although the origin of Y' elements remains unknown, there is certainly evidence of links with transposon activity. Y' amplification in telomerase-negative cells and Ty1 retrotransposition are both impaired at temperatures of 30°C. Y' amplification is dependent on TEC1/ YBR083W, a transcription factor also required for Ty1 retrotransposition. Furthermore Y' RNA associates with Ty1 virus-like particles, and might be reverse-transcribed by the same mechanisms (Maxwell et al. 2004).

There are other curious links between Ty transposons and telomere-associated elements. As mentioned, Ty5 preferentially integrates into budding yeast genomes close to telomeres in S. cerevisiae. Furthermore Ty1 retrotransposition rates are reduced in the absence of Ku proteins, which are involved with DNA repair and also bind to telomeres. Like Y' RNA, Ku proteins appear to associate with Ty1 virus-like particles in vivo and in vitro (Downs & Jackson 1999).

Thus we are presented with a complex picture in which yeast DNA repair processes, telomere structure and maintenance, Y' amplification and the retrotransposition of multiple Ty elements are all intimately linked. If Y' elements are dependent on Ty retrotransposon activity then this association may be an ancient one. Perhaps Y' elements were originally parasites of retroviruses, and entered the yeast genomes along with their host.
With new genome sequence becoming available for a diverse array of yeast (and other) species, we can be optimistic that it will become clearer at what point Y' elements entered the *Saccharomyces* lineage, but perhaps it will fall to the virologists and viral genomics to identify their true ancestor.

**3.3: Analysis of Overlapping Genes**

On July 1st 2003, the *S. cerevisiae* genome, according to SGD, contained 674 pairs of overlapping features, of which 414 were pairs of overlapping nuclear ORFs. Of all these pairs there was only one instance in which both members of the pair were thoroughly characterised: YCR038C (*BUD5*) and YCR039C (*MATα2* or *MATαLPHA2*), and in this case they overlapped by only 4 bp. However in a disproportionately large number of pairs one or both of the overlapping ORFs were uncharacterised (annotated by gene prediction algorithms solely on the basis of DNA sequence). This suggested that the true incidence of overlapping ORFs in the *S. cerevisiae* genome might be much lower than the annotations suggested, and that a number of these hypothetical ORFs might not be genuine.

At this time SGD did not subcategorise ORF features according to the certainty of their classification. Hypothetical or poorly characterised ORFs were often labelled as such in SGD’s ‘description’ field but not in a systematic reliable way. A better measure came from the three GO annotation subcategories (‘molecular_function’, ‘biological_process’ and ‘cellular_component’). ORFs for which all three of these subcategories were
listed as 'unknown' could be considered hypothetical (HY). ORFs for which each of the three subcategories contained functional GO annotations could be considered fully characterised (FC), while ORFs with one or two subcategories listed as 'unknown' could be considered partially characterised (PC).

Table 3.1 summarises the incidence of overlapping ORFs in *S. cerevisiae* in 2003, and the breakdown of pairs by level of characterization. These figures were obtained from SGD’s tab-delimited chromosomal feature file, analysed in Microsoft Excel using various formulae to identify, count and classify overlapping ORFs. It will be noted that in 93.5% of all overlapping pairs, at least one of the ORFs was hypothetical, although hypothetical ORFs constituted only 30.2% of the genome-wide total.

### 3.3.1. A number of hypothetical ORFs mistakenly classified as 'essential'.

Table 3.1 also gives the 2003 incidence of ‘essential’ genes amongst these pairs. These are ORFs for which the null mutant phenotype is listed in SGD as ‘inviable’, based on a systematic large-scale deletion study (Giaever *et al*. 2002). It will be noted that a disproportionate number of ORFs in overlapping pairs are labelled ‘essential’ by this criterion, and that in pairs consisting of one fully characterised ORF and one hypothetical ORF, the hypothetical ORF is annotated as essential in a full quarter of cases.

In a systematic deletion study, the deletion of these hypothetical ORFs would have also wholly or partially deleted the overlapping well-characterised ORF, creating an inviable null mutant phenotype for the hypothetical ORF that
### Individual nuclear ORFs

<table>
<thead>
<tr>
<th>Type</th>
<th>All nuclear ORFs</th>
<th>Overlapping nuclear ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully characterised (FC)</td>
<td>2658 (42.8% of 6215)</td>
<td>207 (25.0% of 828)</td>
</tr>
<tr>
<td>Partially characterised (PC)</td>
<td>1678 (27.0%)</td>
<td>143 (17.3%)</td>
</tr>
<tr>
<td>Hypothetical (HY)</td>
<td>1879 (30.2%)</td>
<td>478 (57.7%)</td>
</tr>
<tr>
<td>Total</td>
<td>6215</td>
<td>828 (13.3% of all 6215 ORFs)</td>
</tr>
</tbody>
</table>

### Overlapping pairs of nuclear ORFs

<table>
<thead>
<tr>
<th>Type</th>
<th>All pairs</th>
<th>Pairs by null mutant phenotype:</th>
<th>Pairs by orientation type:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>One 'essential'</td>
<td>Both 'essential'</td>
</tr>
<tr>
<td>FC+FC</td>
<td>1 (0.2% of 414)</td>
<td>1 (6.7% of 15)</td>
<td>0 (0% of 63)</td>
</tr>
<tr>
<td>FC+PC</td>
<td>18 (4.3%)</td>
<td>3 (20.0%)</td>
<td>1 (1.6%)</td>
</tr>
<tr>
<td>FC+HY</td>
<td>187 (45.2%)</td>
<td>5 (33.3%)</td>
<td>48 (76.2%)</td>
</tr>
<tr>
<td>PC+PC</td>
<td>8 (2.0%)</td>
<td>1 (6.7%)</td>
<td>1 (1.6%)</td>
</tr>
<tr>
<td>PC+HY</td>
<td>109 (26.3%)</td>
<td>4 (26.7%)</td>
<td>8 (12.7%)</td>
</tr>
<tr>
<td>HY+HY</td>
<td>91 (22.0%)</td>
<td>1 (6.7%)</td>
<td>5 (7.9%)</td>
</tr>
<tr>
<td>Total</td>
<td>414</td>
<td>15</td>
<td>63</td>
</tr>
<tr>
<td>Any+HY</td>
<td>387 (90.5%)</td>
<td>10 (66.7%)</td>
<td>61 (90.8%)</td>
</tr>
</tbody>
</table>

Table 3.1. Incidence of overlapping ORFs in the *S. cerevisiae* genome according to SGD on 1/7/03. In only 27 of 414 pairs (6.5% of cases) are both ORFs in the pair fully or partially characterised. In the rest, one or both ORFs are hypothetical and potentially dubious.

is not warranted. We consider this to be an unavoidable flaw in the systematic deletion study: few if any of these hypothetical ORFs are likely to be genuinely essential genes, and they should not be annotated as such.

#### 3.3.2. Overlapping pairs of ORFs tend to be convergent in orientation.

Adjacent or overlapping ORFs can be on the same strand ('parallel') or opposing strands (transcribed in either a 'convergent' or 'divergent' manner). In a random scenario we would expect 50% of gene pairs to be of the parallel type, 25% to be convergent pairs, and 25% to be divergent.

Table 3.1 shows the actual incidence of orientation types amongst overlapping ORF pairs. Cases where one ORF is wholly embedded in the other (on either strand) are categorised separately as these are more complex
occurrences, but in all these cases one partner is hypothetical and therefore dubious. It can be seen that ‘convergent’ pairs are substantially more common than would be expected by chance, but this in fact is expected for biological reasons: promoter sequences are invariably at the 5’ end of ORFs, and in parallel or divergent overlapping pairs such promoter sequences could be embedded in the overlapping partner and would therefore be selectively limited. It makes sense that there would be selection against such orientations in cases of genuine overlapping ORFs.

3.3.3. Supporting evidence from gene expression data.

The genuineness or otherwise of overlapping hypothetical ORFs could also be explored through gene expression data such as SAGE (Serial Analysis of Gene Expression: Velculescu et al. 1997), which is downloadable from SGD at ftp://genome-ftp.stanford.edu/pub/yeast/data_download/systematic_results/SAGE.

SAGE tags are ~14 bp subsequences unique to each gene’s mRNA. A cell’s mRNA population is trapped and reverse-transcribed into cDNA, the cDNAs are cut with enzymes to produce sticky-ended tags, which are concatenated into larger molecules, clonally amplified, sequenced, and analysed. Then the incidence of each tag gives a measure of the mRNA copy number in the source organism. For the purposes of this study, the absolute levels of SAGE tags were less important than the presence or absence of SAGE tag data for a given gene.

Since genes that overlap on the same strand may share the same SAGE tag, only genes that overlap on opposite strands were tested for the
existence of SAGE data (using Microsoft Excel formulae). The results are shown in Table 3.2. Among these opposite-strand overlapping ORFs, a smaller proportion (33.3%) of opposite-strand overlapping hypothetical ORFs had associated SAGE tags than did fully or partially characterised ORFs, as expected. Of those hypothetical ORFs that overlapped fully or partially characterised ORFs, only 21.5% and 25.8% respectively had SAGE tags, suggesting that many of these were not genuine ORFs. In pairs of overlapping hypothetical ORFs, roughly half (51.4%) of the ORFs had SAGE data, suggesting that in most cases only one member of each pair was genuine.

Microarray data from 10 separate experiments (DeRisi et al. 1997; Chu et al. 1998; Spellman et al. 1998; Ferea et al. 1999; Wyrick et al. 1999; Gasch

<table>
<thead>
<tr>
<th>Individual nuclear ORFs</th>
<th>Overlapping ORFs</th>
<th>Overlapping ORFs with SAGE tags</th>
<th>Overlapping ORFs on opposite strands</th>
<th>Overlapping ORFs on opposite strands with SAGE tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully characterised (FC)</td>
<td>207</td>
<td>165 (79.7%)</td>
<td>172</td>
<td>138 (80.2%)</td>
</tr>
<tr>
<td>Partially characterised (PC)</td>
<td>143</td>
<td>89 (62.2%)</td>
<td>115</td>
<td>82 (71.3%)</td>
</tr>
<tr>
<td>Hypothetical (HY)</td>
<td>478</td>
<td>164 (34.3%)</td>
<td>387</td>
<td>129 (33.3%)</td>
</tr>
<tr>
<td>Total</td>
<td>828</td>
<td>418</td>
<td>674</td>
<td>349</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pairs of nuclear ORFs overlapping on opposite strands</th>
<th>Total pairs</th>
<th>Individual ORFs with SAGE tags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FC type</td>
<td>PC type</td>
</tr>
<tr>
<td>FC+FC 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FC+PC 14 (28 ORFs)</td>
<td>10</td>
<td>6 (42.9% of PC type)</td>
</tr>
<tr>
<td>FC+HY 158 (316 ORFs)</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>PC+PC 12 (24 ORFs)</td>
<td>7</td>
<td>7 (58.3% of PC type)</td>
</tr>
<tr>
<td>PC+HY 89 (178 ORFs)</td>
<td>69</td>
<td>23 (25.8% of HY type)</td>
</tr>
<tr>
<td>HY+HY 140 (280 ORFs)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2. SAGE expression data for overlapping ORFs. SAGE data produced by Velculescu et al. (1997), obtained through SGD on 1/7/2003. Genes that overlap on the same strand are likely to share the same SAGE tags, so this analysis is restricted to ORFs that overlap on opposing strands.
et al. 2000; 2001; Lyons et al. 2000; Ogawa et al. 2000; Roberts et al. 2000), mined through SGD’s ‘Expression Connection’ utility, were also studied with regard to overlapping ORFs, but expression data was present for nearly all of the ORFs considered, hypothetical or otherwise. In cDNA microarrays, each spot contains a double-strand cDNA denatured in situ to provide single-stranded probes matching each strand, which would make this data unsuitable for discriminating overlapping ORFs.

3.3.4. Overhaul of ORF classification following comparative studies.

Later in 2003, two comparative genomics studies (Kellis et al. 2003; Cliften et al. 2003) were published in which the genome of *S. cerevisiae* was compared to sequence from closely related species relatives.

Kellis et al. reported ~500 ORFs that were of dubious authenticity based on their lack of conservation, while Cliften et al. reported 515 dubious ORFs. This confirmed the results described above. Both studies also identified some previously unknown ORFs and recommended various changes including the merger of some ORFs and alterations to start and stop codons.

Following these publications, SGD underwent a major revision of its *S. cerevisiae* annotations, and officially subcategorised ORFs as either verified, uncharacterised, or dubious.

Table 3.3 shows a recent re-analysis of overlapping ORFs, using the latest annotation data and the post-2003 ORF categories. If ‘dubious’ ORFs are discounted then there are now only 29 pairs of overlapping ORFs (either
### Individual nuclear ORFs

<table>
<thead>
<tr>
<th>Type</th>
<th>All nuclear ORFs</th>
<th>Overlapping nuclear ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verified (VE) [(\text{\textsuperscript{=FC}})]</td>
<td>4412 (67.1% of 6576)</td>
<td>460 (38.0% of 1212)</td>
</tr>
<tr>
<td>Uncharacterised (UN) [(\text{\textsuperscript{=PC}})]</td>
<td>1349 (20.5%)</td>
<td>144 (11.9%)</td>
</tr>
<tr>
<td>Dubious (DU) [(\text{\textsuperscript{=HY}})]</td>
<td>815 (12.4%)</td>
<td>608 (50.2%)</td>
</tr>
<tr>
<td>Total</td>
<td>6576</td>
<td>1212 (18.4% of all 6576 ORFs)</td>
</tr>
</tbody>
</table>

### Overlapping pairs of nuclear ORFs

<table>
<thead>
<tr>
<th>Type</th>
<th>All pairs</th>
<th>Pairs by null mutant phenotype:</th>
<th>Pairs by orientation type:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>One 'essential'</td>
<td>Both 'essential'</td>
</tr>
<tr>
<td>VE+VE</td>
<td>7 (1.1% of 635)</td>
<td>2 (3.8% of 53)</td>
<td>1 (1.5% of 66)</td>
</tr>
<tr>
<td>VE+UN</td>
<td>14 (2.2%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>VE+DU</td>
<td>447 (70.4%)</td>
<td>48 (90.6%)</td>
<td>60 (90.9%)</td>
</tr>
<tr>
<td>UN+UN</td>
<td>8 (1.3%)</td>
<td>0</td>
<td>1 (1.5%)</td>
</tr>
<tr>
<td>UN+DU</td>
<td>127 (2.0%)</td>
<td>3 (5.7%)</td>
<td>3 (4.9%)</td>
</tr>
<tr>
<td>DU+DU</td>
<td>32 (5.0%)</td>
<td>0</td>
<td>1 (1.5%)</td>
</tr>
<tr>
<td>Total</td>
<td>635</td>
<td>53</td>
<td>66</td>
</tr>
<tr>
<td>Any+DU</td>
<td>606 (95.4%)</td>
<td>51 (96.2%)</td>
<td>64 (97.0%)</td>
</tr>
</tbody>
</table>

Table 3.3. Incidence of overlapping ORFs in the *S. cerevisiae* genome according to SGD on 10/9/06. SGD now classifies ORFs as 'verified', 'uncharacterised' or 'dubious'. A large number of the dubious ORFs were annotated as such following comparative studies by Kellis et al. (2003) and Cliften et al. (2003).

'verified' or 'uncharacterized'). This compares to a total of 27 pairs if hypothetical ORFs were discounted from the original study (see Table 3.1).

In cases where dubious ORFs overlap verified or uncharacterised ones there is now a disproportionate incidence of embedded pairs, very few of which, if any, are likely to be genuine.

However Table 3.3 also shows that there remain a large number of dubious ORFs that overlap verified ones and are still annotated as essential according to the systematic large-scale deletion study performed by Giaever et al. (2002). Indeed, of the 67 dubious ORFs annotated with inviable null mutant phenotypes genome-wide, 66 overlap another feature. These annotations should be reconsidered.
3.4: Annotation of Telomeric Elements in SGD

In 2003, the Saccharomyces Genome Database (http://www.yeastgenome.org) contained no annotations for telomeres or telomere-associated repeats, despite the fact that Professor E. Louis had provided such annotations to the SGD's European counterpart, the Comprehensive Yeast Genome Database (CYGD) which is hosted by the Munich Information Centre for Protein Sequences (MIPS) at http://mips.gsf.de/proj/yeast/CYGD/db/index.html.

The SGD curators, in particular Dianna Fisk, were eager to rectify this omission, so we provided these annotations, based on the most recent available version of the sequence, and together with SGD decided upon a notation scheme, shown in Table 3.4.

The annotations are more detailed than those in CYGD, with separate annotations for telomeric repeats, for the entire telomeric region (including repeats and telomere-associated elements), for Y' elements, core X elements and for the X element subtelomeric repeats. The previous designation of

<table>
<thead>
<tr>
<th>Element type</th>
<th>Notation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telomere (spans all elements below)</td>
<td>TEL**[L/R]</td>
<td>TEL08L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEL08R</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEL12L</td>
</tr>
<tr>
<td>Telomeric repeat</td>
<td>TEL**[L/R]-TR[1-3?]</td>
<td>TEL08L-TR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEL08R-TR1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEL12L-TR3</td>
</tr>
<tr>
<td>Y' element</td>
<td>TEL**[L/R]-YP[1-2?]</td>
<td>TEL08L-YP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEL12L-YP1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEL12L-YP2</td>
</tr>
<tr>
<td>Core X element</td>
<td>TEL**[L/R]-XC</td>
<td>TEL08L-XC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEL12L-XC</td>
</tr>
<tr>
<td>X element combinatorial repeats</td>
<td>TEL**[L/R]-XR</td>
<td>TEL08L-XR</td>
</tr>
<tr>
<td>(formerly 'subtelomeric repeats' -</td>
<td></td>
<td>TEL12L-XR</td>
</tr>
<tr>
<td>see text)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.4. Telomeric annotations provided to SGD in 2003. Where multiple discrete copies of an element exist on the same chromosome end, the different elements are numbered, from the ends inwards.
these latter elements as STRs was considered misleading by the author, since the acronym STR is more widely used to mean simple tandem repeats, and these elements are anything but simple and regular. An alternative designation (X-element combinatorial repeats: XCRs) was proposed and accepted.

As a result of this work, non-coding telomeric elements are now fully annotated in the widely used SGD (since 31st October 2003; see SGD newsletter at http://www.yeastgenome.org/newsletters/2003-10-31.shtml), and coordinates are automatically updated with each new sequence revision. Moreover SGD's advanced feature search (at http://db.yeastgenome.org/cgi-bin/search/featureSearch) makes it simple to search for and retrieve the coordinates and sequence for elements of each type.
4.1: Introduction

As discussed in Section 1.4, centromeres are a vital non-coding region of chromosomes in budding yeast and most other eukaryotes. The point centromeres of *S. cerevisiae* and its close relatives are atypically small and unusual for centromeres in that they are at least partly defined by consensus motifs, but they are no less important in terms of function.

When genome sequence became available for *S. paradoxus*, *S. mikatae* and *S. bayanus* it was expected that the intergenic regions containing their centromeres would be as well-conserved, or better conserved, than other intergenic regions. However a simple dotplot of these regions, conducted out of curiosity using the Dotter program (see Section 2.5), revealed a very high level of variation around the point centromeres, as described below.

Since it was possible that other intergenic regions might be more divergent than previously expected, it was necessary to conduct a genome-wide analysis of the variation of these regions in order to determine whether intergenic regions containing centromeres were indeed more divergent than those without.

4.2: Results

*Table 4.1* describes the intergenic environments of the 16 *S. cerevisiae* point centromeres. The average intergenic region is 578 bp long, though there is substantial variation with some genes being separated by only a few bases and others apparently 7 or more kilobases apart. There may of course be
<table>
<thead>
<tr>
<th>Left flanking ORF</th>
<th>ORF-CEN distance (bp)</th>
<th>Centromere (end size in bp)</th>
<th>CEN-ORF distance (bp)</th>
<th>Right flanking ORF</th>
<th>Total intergenic size (bp)</th>
<th>Orientation type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>YAL001C (TFC3)</strong> *</td>
<td>299</td>
<td>CEN1 109</td>
<td>682</td>
<td><strong>YAR002W (NUP60)</strong> Subunit of the nuclear pore complex (NPC)</td>
<td>1090</td>
<td>&lt; &gt; &gt; (div)</td>
</tr>
<tr>
<td><strong>YBL001C (ECM15)</strong></td>
<td>740</td>
<td>CEN2 108</td>
<td>625</td>
<td><strong>YBR001C (NTH2)</strong> Cuticular subunit of the nuclear pore complex (NPC)</td>
<td>1473</td>
<td>&lt; &gt; &lt; (par)</td>
</tr>
<tr>
<td><strong>YCL001W-B</strong></td>
<td>361</td>
<td>CEN3 108</td>
<td>1191</td>
<td><strong>YCR001W</strong> Dubious ORF; likely to encode a protein based on experimental and comparative sequence data</td>
<td>1660</td>
<td>&gt; &gt; &gt; (par)</td>
</tr>
<tr>
<td><strong>YDL001C (RMD1)</strong></td>
<td>435</td>
<td>CEN4 102</td>
<td>406</td>
<td><strong>YER001W (MNN1)</strong> Alpha-1,3-mannosyltransferase, integral membrane glycoprotein of the Golgi complex</td>
<td>2541</td>
<td>&gt; &lt; (div)</td>
</tr>
<tr>
<td><strong>YFL001W (DEG1)</strong></td>
<td>51</td>
<td>CEN6 109</td>
<td>490</td>
<td><strong>YFR001W (LOC1)</strong> Nuclear protein involved in asymmetric localization of ASH1 mRNA</td>
<td>650</td>
<td>&gt; &gt; &gt; (par)</td>
</tr>
<tr>
<td><strong>YGL001C (ERG26)</strong></td>
<td>425</td>
<td>CEN7 110</td>
<td>95</td>
<td><strong>YGR001C</strong> Putative protein of unknown function with similarity to methyltransferase family members</td>
<td>630</td>
<td>&lt; &lt; &lt; (par)</td>
</tr>
<tr>
<td>**YHL001W (RPL14B) ** §</td>
<td>502</td>
<td>CEN8 109</td>
<td>352</td>
<td>**YHR001W (OSH7) ** §</td>
<td>963</td>
<td>&gt; &gt; (par)</td>
</tr>
<tr>
<td><strong>YIL001W</strong></td>
<td>139</td>
<td>CEN9 117</td>
<td>405</td>
<td><strong>YIR001C (SGN1)</strong> Cytoplasmic RNA-binding protein</td>
<td>661</td>
<td>&gt; &lt; (con)</td>
</tr>
<tr>
<td>**YJL001W (PRE3) **</td>
<td>388</td>
<td>CEN10 110</td>
<td>377</td>
<td><strong>YJR001W (AVT1)</strong> Vascular transporter</td>
<td>875</td>
<td>&gt; &gt; (par)</td>
</tr>
<tr>
<td><strong>YKL001C (MET14)</strong></td>
<td>751</td>
<td>CEN11 109</td>
<td>362</td>
<td><strong>YKR001C (VPS1)</strong> GTPase required for vascular protein sorting</td>
<td>1222</td>
<td>&lt; &lt; &lt; (par)</td>
</tr>
<tr>
<td><strong>YLL001W (DNM1)</strong></td>
<td>672</td>
<td>CEN12 111</td>
<td>442</td>
<td><strong>YLR001C</strong> Putative protein of unknown function; GTPase required for vascular protein sorting</td>
<td>1225</td>
<td>&gt; &lt; (con)</td>
</tr>
<tr>
<td><strong>YML001W (YPT7)</strong></td>
<td>231</td>
<td>CEN13 110</td>
<td>877</td>
<td><strong>YMRO01C (CDC5)</strong> Polo-like kinase with similarity to Xenopus Plk1 and S. pombe Pclp</td>
<td>1218</td>
<td>&gt; &gt; (con)</td>
</tr>
<tr>
<td><strong>YNL001W (DOM34)</strong></td>
<td>142</td>
<td>CEN14 109</td>
<td>754</td>
<td><strong>YNR001C (CIT1)</strong> Citrate synthase</td>
<td>1005</td>
<td>&gt; &gt; (con)</td>
</tr>
<tr>
<td><strong>YOL001W (PHO80)</strong></td>
<td>461</td>
<td>CEN15 110</td>
<td>130</td>
<td><strong>YOR001W (RRP6)</strong> Exonuclease component of the nuclear exosome</td>
<td>701</td>
<td>&gt; &gt; &gt; (par)</td>
</tr>
<tr>
<td><strong>YPL001W (HAT1)</strong></td>
<td>228</td>
<td>CEN16 108</td>
<td>311</td>
<td><strong>YPR001W (CIT3)</strong> Citrate synthase</td>
<td>647</td>
<td>&gt; &gt; &gt; (par)</td>
</tr>
</tbody>
</table>

**Table 4.1. Intergenic regions containing centromeres in S. cerevisiae.** Descriptions of flanking genes taken from SGD.

Notes: Genes annotated as essential are in bold (§ indicates unknown). Presence of introns indicated with * . Orientations of flanking genes in final column: 'div'= flanking genes in divergent orientation; 'con'= flanking genes in convergent orientation; 'par'= flanking genes in parallel orientation ('par' & 'par1' indicate parallel orientation with respect to asymmetric point centromers).
ORFs or other functional sequences in those gaps that have yet to be identified. While intergenic gaps that contain centromeres tend on average to be a little larger (mean 1068 bp) they are also very variable (630 to 2541 bp).

There are occasional difficulties in defining what constitutes the proper boundaries of these intergenic regions. YCR001W, the gene to the right flank of CEN3, is considered a highly dubious ORF, based on comparative studies and a lack of known motifs. If this sequence is in fact not a gene but intergenic then the true flanking ORF (CDC10) is some distance further away. It will be observed that the ORF on the right flank of CEN1 is labelled YAR002W, not 001 as might be expected due to the labelling convention which begins numbering at the centromere for each arm of the chromosome. There is no YAR001: whatever sequence originally bore that number has long since been dismissed as non-coding.

Provided that one bears such issues in mind, some observations can be made about the nature of the flanking ORFs. 4 out of the 32 flanking genes are essential (TFC3, ERG26, PRE3, CDC5), according to the ‘inviable’ null mutant phenotype listed in SGD. Of the 6576 nuclear ORFs in the database as of 20/07/06, 1100 have inviable null mutant phenotypes, so ~5 of the 32 genes would be expected to be essential by chance: the number observed is not significantly different. It must be noted that 2 years earlier, only 2 of the 32 flanking genes were classified as essential (1103 out of 6658 nuclear ORFs genome-wide, so that the expected value out of 32 was still ~5). The small number of flanking ORFs and underlying uncertainty about the accuracy of their annotations prevent us from drawing any reliable conclusions from this particular data.
It can also be observed that of the 32 flanking ORFs, 4 have introns \((TFC3, RPL14B, PRE3, YGR001C)\), indeed one of these \((YGR001C)\) has two introns, although its function is not known. Introns are comparatively rare in the \(S.\, cerevisiae\) genome: only 264 out of 6576 nuclear ORFs are known to have them, and only 11 of those have multiple introns. Thus one might expect by chance that only one of the 32 centromere-flanking ORFs would have introns, and none would be likely to have multiple introns. Thus there are significantly more introns among the flanking ORFs than would be expected \((\chi^2 = 22.8, \text{d.f.}=1, \ p = 1.75 \times 10^{-6})\), although again one must be cautious due to the inherently small size of the dataset.

Intergenic regions can be classified according to the orientation of the ORFs that flank them. Both flanking ORFs can be in the same orientation \((>>\) or \(<<\): parallel), both can be transcribed towards the intergenic region \((><:\) convergent) or both can be transcribed away from the intergenic region \(<>(\): divergent). Since promoters are usually to be found at the 5' end of genes, more promoters would be expected to occupy a divergent intergenic region than would occupy a convergent one, while parallel intergenic regions would be expected to contain an intermediate number of promoter sequences.

As shown in Table 4.1, of the 16 centromere-containing intergenic regions, 9 are of the parallel type (8 expected by chance), while 5 are convergent (4 expected) and 2 are divergent (4 expected). Again such small amounts of data do not lend themselves to rigorous statistical analysis, but it can be seen that there is nothing strikingly unusual about the orientation of centromere flanking genes.
Point centromeres, being asymmetrical, have their own orientation as well. Of the 9 intergenic regions with parallel flanking ORFs, 4 contain centromeres in what might be termed 'the same orientation' (if we treat a centromere as running in the direction of CDEI towards CDEIII), while 5 contain centromeres in the opposite orientation with respect to the flanking ORFs. Again this is no different from what would be expected by chance. Overall it would appear that centromeres occupy rather ordinary intergenic regions (albeit with a possible slight overabundance of nearby introns). However a comparative analysis demonstrates that this is not the case.

4.2.1. Dot matrix plots of centromeres across multiple species and strains reveal exceptional divergence.

An initial ad hoc comparison of the centromeres in *S. cerevisiae* and some of its close relatives was carried out using dot matrix plots. Figure 4.1 shows one such plot, produced using Dotter. Each of the centromeric regions examined appeared to be unusually divergent in each of the six pairwise comparisons between the four species, and this divergence appeared to span a greater extent of the regions than the point centromere alone.

Some additional sequencing was undertaken on our behalf by AGOWA genomics, in order to assess the polymorphism within species, and this also proved to be much higher than expected, especially within *S. paradoxus* (as shown in Figure 4.2).
Figure 4.1. Sample dotplot showing divergence of CEN1 across four closely related *Saccharomyces* species. The CDEII element, which forms the bulk of the point centromere sequence, is visible as a smudge (indicated with arrows) due to its AT-rich composition. However it is clear that there is substantial sequence divergence on either side of the centromere itself. This was particularly surprising in the comparison between *S. cerevisiae* and *S. paradoxus*, thought to be very similar in sequence. The same variation was visible in plots of the other centromeres.
Figure 4.2. Sample dotplot showing divergence of CEN7 across four closely related *Saccharomyces* species, including an additional *S. cerevisiae* strain and two additional *S. paradoxus* strains. Additional sequence reads produced by AGOWA genomics using their own primers. In this plot the positions of the centromere itself and the flanking coding regions have been colour-coded using Adobe Photoshop to better convey where the variation occurs in relation to these sequences. The variation (polymorphism) within *S. paradoxus* strains is particularly striking, with a large deletion present in the type strain CBS432 (indicated with an arrow).
4.2.2. Genome-wide analysis shows intergenic regions containing centromeres are more divergent than most other intergenic regions.

In order to confirm whether the observed centromeric variation was indeed atypical of intergenic regions, a large-scale analysis was conducted using custom scripts to identify and compare all intergenic regions. As described in detail in Section 2.6, intergenic regions were aligned against their homologues using ClustalW, and for each region as a whole the length of the region and the number of mismatches were measured. For the sake of comparison, the same values were also determined for known ORFs which were present and complete in the available sequence.

Figure 4.3 shows a scatter plot of these values for each genic and intergenic region in the S. cerevisiae vs. S. paradoxus comparisons. It is evident that intergenic regions containing centromeres are indeed consistently more divergent than the vast majority of other intergenic regions here analysed. Moreover many of the outliers in the 'other intergenic regions' category prove to have distinctive characteristics that may explain their own high rates of divergence, such as short length or proximity to transposons or ARS elements. (Note that intergenic regions that actually contain such potentially confusing elements were excluded from the analysis, but regions that lie close to—but do not actually overlap—such elements were not excluded).
Intergenic regions containing centromeres
× Other intergenic regions
× Genes (essential)
× Genes (non-essential)

Figure 4.3. Scatterplot showing number of mismatches vs total length for each region compared between *S. cerevisiae* and *S. paradoxus*. Plotted using SPSS. As expected, larger regions tend to contain more mismatches than smaller ones, and intergenic regions are generally more divergent than genes. The large number of points conceals the true distribution: for each of the highly-populated categories there are many more points close to the centre of the conic distribution than there are at the periphery. To illustrate this, regression lines are shown for the intergenic categories, plus 95% confidence intervals based on the predicted mean from the regression line. Intergenic regions containing centromeres (highlighted in red) are consistently far outside the distribution formed by other intergenic regions. See also Figure 4.4.
For a given region, dividing the number of mismatches by the total base pair length of the region gives a simple and robust measure of identity that is commonly called the 'P-distance'. **Figure 4.4** plots the distribution of these values for each region in each of the six pairwise comparisons.

**Figure 4.4** demonstrates that the consistent centromere-adjacent supervariation is observable in all six of the pairwise comparisons. It is perhaps surprising to note that genes annotated by SGD as essential (having inviable null mutant phenotypes) are on average only slightly less divergent than their non-essential counterparts. Note however that the measure here used makes no distinction between synonymous and non-synonymous mutations (since such measures cannot be compared to intergenic regions where there is no such distinction), and that such annotations are in any case in a state of constant flux and may not be entirely reliable.

![Figure 4.4](image-url)
The difference in divergence between centromeric intergenic regions and other intergenic regions is very obvious from the figures. It is also statistically significant to a high degree. Independent-sample T-tests and Kruskal-Wallis H-tests both show, without assumptions about variance, that in each case there is a p-value <0.0005, i.e. a very low probability that the differences observed between these categories is attributable to chance.

**4.2.3. No correlation between divergence and A & T composition.**

Since CDEII is very AT-rich and also rich in A and T doublets or triplets, it was possible that such a composition bias might correlate, genome-wide, with increased divergence in intergenic regions. To test this hypothesis, the A and T composition and the AAA and TTT triplet composition of all intergenic regions and genic regions was measured and plotted against the observed p-distance. **Figure 4.5** shows these results in the *S. cerevisiae* v *S. paradoxus* pairwise comparison. No correlation is observed between these measures of composition and the rate of divergence.

![Figure 4.5](image)

**Figure 4.5. Correlation between composition and divergence between S. cerevisiae and S. paradoxus.** Regions are colour-coded by category using the same scheme as in Figures 4.3 and 4.4. While genes tend to have a lower A & T composition than intergenic regions, and intergenic regions containing centromeres tend to have a slightly higher A & T composition than other intergenic regions, there is no overall correlation between the two factors.
4.2.4. Supervariation not only within the centromere itself but also in flanking sequence.

Although dot matrix comparisons (see Figures 4.1 and 4.2) indicated that centromere-adjacent supervariation is not limited to the point centromere itself but is also present on either side, it was clear that CDEII is quite variable in its own right. In order to determine the sequence divergence within these elements on their own, and the degree to which that divergence contributes to the overall divergence of the entire intergenic region, the genome-wide analysis was refined and then repeated in order to measure the divergence of these regions independently of each other.

Figure 4.6 shows the results of this secondary analysis, using a bar plot format similar to that of Figure 4.4, and establishes beyond doubt that while centromeric DNA is highly variable on its own (almost entirely due to sequence divergence within CDEII), the flanking intergenic regions are also supervariable; indeed these adjacent regions are often more divergent than the centromere itself, particularly so between more distant relatives such as S. cerevisiae and S. bayanus. Independent-sample T-tests show the differences between these centromere-related categories and other intergenic regions are also statistically significant to a high degree (p<0.0005).

4.2.5. Sequence divergence seems to be related to centromere function rather than being a property of centromeric sequence.

There are within S. cerevisiae and its relatives a number of DNA subsequences which resemble the centromere DNA elements, but which do not actually function as centromeres.
Figure 4.6. Variability of centromeric and adjacent DNA compared to other intergenic regions. For each of the six pairwise comparisons, the distributions of each category of region are shown as standard box plots. Intergenic regions that contain centromeric DNA (grey hatched) are consistently much more variable than most other intergenic regions (black). This variation is not limited to the centromeric DNA itself (grey) but also encompasses adjacent intergenic sequence (white hatched). In the category "other intergenic regions", diamond shapes indicate the values of normal intergenic regions that contain pseudo-centromeric DNA sequences (see text). These do so not show abnormal variability.
Centromere DNA elements could be correctly identified in all four species, without generating any false positives, using the following pattern:

\[ WCAYRTG \text{ (CDEI) followed by 78-95 bp with at least 79\% A or T content and at least 34\% AA or TT content (CDEII), followed by RKNYNNNNNWCCGAA (CDEIIIa and CDEIIIb). } \]

To identify 'pseudo-centromeric' sequences, the stringency of this pattern was relaxed to the following pattern:

\[ WCAYRTG, \text{ followed by 68-120bp with } >69\% \text{ A or T content and } >29\% \text{ AA or TT content, followed by NNNNNNNNNNWCCGAA. } \]

Of the pseudo-centromeres thus identified, only a handful were fully intergenic and present in homologous sequence from multiple species. Although the resulting sample size was small, in no case was the divergence in one of these pseudo-centromeric intergenic regions (marked in Figure 4.6 as diamond shapes) found to be exceptional. This is further evidence that the supervariation results from centromere function, rather than particular properties inherent in the sequence itself.

4.2.6. Other measures of sequence identity produce similar results.

The measure of sequence identity used above (P-distance: mismatches/base) is a simple one. More complex measures exist. The Jukes-Cantor correction (Jukes & Cantor 1969) takes into account hidden back-mutations, where a substitution is reversed by a subsequent mutation that restores the original base. The P-distance does not account for this possibility and therefore underestimates the actual mutation rate.
The Kimura 2-parameter score (Kimura 1980) further takes into account the fact that transversion mutations (A or G → T or C) are more severe and less common than transition mutations (A→G or T→C) and adjusts the measure of sequence identity accordingly.

While these more sophisticated measures are necessary to better determine actual substitution rates (as opposed to the apparent mutation rates given by P-distance), they are less important to studies such as this that consider relative substitution rates. In addition to P-distance values, Jukes-Cantor and Kimura 2-parameter scores were calculated for each region in the genome-wide comparison, but the use of these measures does not significantly affect the results: centromere-adjacent sequences are still more divergent than other intergenic regions.

4.2.7. Centromere-binding proteins do not show unusual divergence.

The sequence divergence around centromeres does not significantly affect the short centromere-binding motifs CDEI and CDEIII, and while the AT-rich CDEII element can vary in actual sequence it is evidently constrained by length and composition. It was therefore expected that the proteins that are known to bind to these elements (described in Section 1.4) would not show any abnormal levels of divergence either.

There is no evidence for abnormal amino-acid conservation of these proteins in supplementary data from Kellis et al. (2003), as shown in Figure 4.7. Cbf1p binds to CDEI and is known (like its binding sequence) to be non-essential to centromere function (though it greatly improves the efficiency of that function), and consequently diverges more rapidly than the other
centromere-binding proteins, but not significantly more so than other non-essential proteins. Of the other proteins in Figure 4.7, all of which are essential, Cse4p, Mif2p and Cep3p show average rates of divergence. Skp1p, which has a regulatory function, is the best conserved of these proteins, but it is also well-conserved throughout the eukaryotes (Connelly & Hieter 1996; Seol et al. 2001). Talbert et al (2004) report that MIF2 is under negative selection "consistent with Mif2p interacting with a stable centromere, rather than one that is rapidly evolving".

Whatever process produces the centromere-adjacent supervariation in these species does so without strongly affecting the important motifs and proteins responsible for centromere function.

Figure 4.7. Variation of centromere-binding proteins. Using supplementary data from Kellis et al. (2003). Amino acid conservation values are unknown for some proteins in some species.
4.2.8. There is also an abnormally-high level of centromeric sequence polymorphism within *Saccharomyces* species.

Additional sequencing across certain centromeres in some additional strains of *S. cerevisiae* and *S. paradoxus* (shown in the dot matrix plots of Figure 4.2) demonstrated that there was not only substantial sequence diversity (between species) but also substantial polymorphism (within species) at these loci.

In order to investigate this further, the 630 bp intergenic region containing CEN7 was entirely PCR-sequenced in a larger number of *S. cerevisiae* and *S. paradoxus* strains. These sequence reads were provided by AGOWA genomics using primers of their own design (the presence of well conserved genes on either side of the supervariable region makes it relatively easy to design primers that should work in multiple strains and species, and the moderate sizes of most centromere-containing intergenic regions means that a single read from each direction can cover the whole intergenic span).

The strains were suggested by (and DNA preps kindly provided by) G. Liti, with the intention of broadly covering a range of geographical sub-populations of *S. paradoxus*, and a representative cross-section of the *S. cerevisiae* species. We also included *S. paradoxus* CBS432 in this set, so as to confirm the sequence provided by Kellis *et al.* (2003). A much smaller subset of these strains was also the subject of PCR-sequencing across the 647 bp intergenic region containing CEN16 (see CD-ROM: 06 Generated data\AGOWA sequencing\Centromere sequencing, finished).

For each region in each strain the resulting bidirectional reads were clipped and combined in BioEdit, using the *S. cerevisiae* S288C genome from
SGD as a reference. This process revealed such a degree of polymorphism that most of the strains could be distinguished from each other on the basis of this single centromeric 'fingerprint'.

Figure 4.8 shows a phylogenetic tree, constructed using Mega 3.1 (neighbour-joining method using P-distance) from the sequencing across the intergenic region containing CEN7.

Figure 4.8. Phylogenetic tree of Saccharomyces strains based on intergenic sequence spanning CEN7. Based on ClustalW alignment of sequence between genes ERG26 and YGR001C. S. paradoxus and S. cerevisiae sequencing provided by AGOWA. The geographical origin of each strain is given where known.

Notes: ** Also S. paradoxus N17 and DBVPG65665 (previously known as S. douglasii)

*** Also S. paradoxus DBVPG6303, DBVPG6304 and DBVPG6037

* Also S. paradoxus UFRG50816; both strains also known as S. cariocanus.
Although sequencing errors may have occurred, the tree produced by this sequencing is in strong agreement with other phylogenies produced of these strains (G. Liti, personal communication), and with the known geographical origins of the strains. *S. paradoxus* is known to be a genetically diverse group with distinct geographic subpopulations. Nevertheless it is remarkable that such a detailed phylogeny can be constructed from sequencing of a single short region.

4.2.9. **Polymorphism takes the form of both substitutions and indels and appears to be focussed immediately around the point centromere.**

*Figure 4.9* shows the actual sequence of the CEN7 in multiple strains. Both multiple substitutions and multiple indels are observed, and the polymorphisms tend to cluster around the centromere rather than lying close to the adjacent ORFs.
1.34 bp deletion in European S. paradoxus

Conserved promoter region for ERG26?

Figure 4.9. Intergenic sequence containing CEN7 in multiple S. paradoxus strains and S. cerevisiae S288c. This is the reverse complement of the S. cerevisiae homologous sequence, showing CDEI-III in their canonical order. Conserved residues are shown in grey. S. paradoxus strains: 1= DG1768 & DBVPG6303 & DBVPG6304 & CBS2980; 2= YPS138; 3= YPS125; 4= YPS145; 5= UFRJ50791 & UFRJ50816 (S. cariocanus); 6= CBS8438; 7= CBS8437; 8= IFO1804; 9= CBS432 & DBVPG6565 & N-17; 10= Q4.1; 11= T21.4; 12= CBS5829. 13= S. cerevisiae S288c. Alignment by ClustalW with manual corrections. The corresponding sequence file is on the CD-ROM (06 Generated data\AGOWA sequencing\Centromere sequencing, finished\Fig_4_9.fas).
SPECIAL NOTE

ITEM SCANNED AS SUPPLIED
PAGINATION IS AS SEEN
Figure 4.9. Intergenic sequence containing CEN7 in multiple S. paradoxus strains and S. cerevisiae S288c. This is the reverse complement of the S. cerevisiae homologous sequence, showing CDEI-III in their canonical order. Conserved residues are shown in grey. S. paradoxus strains: 1= DG1768 & DBVPG6303 & DBVPG6304 & CBS2980; 2= YPS138; 3= YPS125; 4= YPS145; 5= UFRJ50791 & UFRJ50816 (S. cariocanus); 6= CBS8438; 7= CBS8437; 8= IFO1804; 9= CBS432 & DBVPG6565 & N-17; 10= Q4.1; 11= T21.4; 12= CBS8292. 13= S. cerevisiae S288c. Alignment by ClustalW with manual corrections. The corresponding sequence file is on the CD-ROM (06 Generated data/AGOWA sequencing/Centromere sequencing, finished/Fig_4_9.fas).
4.3: Discussion

To summarise the results above: in the *sensu stricto* family of budding yeasts the intergenic regions containing centromeres are consistently more variable than other intergenic regions. They are both highly polymorphic within species and highly divergent between them. This 'supervariation' occurs despite the obvious selective pressure to maintain full centromere function, and there is no evidence to suggest that centromere-binding proteins have to adapt to the diverging sequences. Supervariation is not limited to the AT-rich CDEII but is also apparent on either side of the point centromere motifs. This fact, plus the absence of a genome-wide correlation between AT composition and divergence, and the lack of divergence at pseudocentromeric sequences, suggest that supervariation occurs as a result of centromere function rather than the nature of centromeric DNA sequence.

4.3.1. Confirmation by other researchers.

As the main research outlined above was drawing to a close we learned of another research group who had recently independently observed the same phenomenon. Austin Burt’s lab at Imperial College London have been undertaking the detailed sequencing of chromosome III in a large number of *S. paradoxus* strains, and the centromere-adjacent supervariation was observed by Douda Bensasson in the course of her work with these sequences. Their use of a sliding window of sequence identity along the length of chromosome III showed a prominent spike of sequence
polymorphism around the centromere: by far the most variable part of that chromosome (D. Bensasson, personal communication).

4.3.2. Centromere-adjacent supervariation versus other variable regions and rates of neutral mutation.

It would be interesting to compare centromere-adjacent supervariation to polymorphism and divergence in other known variable regions of the genome. The internal transcribed spacer regions (ITSs) in the rDNA repeats at the \textit{RDN1} locus on chromosome XII are non-coding regions that are transcribed and then excised from the 35S pre-rRNA. Such regions are considered strongly selectively neutral, and are therefore commonly used in constructing phylogenies (McCullough \textit{et al.} 1998).

Unfortunately the Broad Institute sequence data does not appear to cover these regions in \textit{S. paradoxus} and \textit{S. mikatae}, and only partially cover it in \textit{S. bayanus}, so it is not currently possible to compare variation in those regions to the centromere-adjacent supervariation here described.

Synonymous mutations in genes (substitutions that do not alter the encoded amino acid) are also used as a measure of the neutral rate of genome evolution, and Kellis \textit{et al.} (2003) have provided this data for many genes in their supplementary information. Unfortunately the methods for calculating such rates are complex and codon dependent, so are not applicable to—or comparable with—substitution rates in non-coding intergenic DNA.
4.3.3. Possible causes of centromere-adjacent supervariation.

At present there is no clear explanation for centromere-adjacent supervariation in the budding yeasts. There are two broad possibilities: that the supervariation results from a lack of any selective pressure in the regions where it occurs, or that it results from a process that actively drives mutation.

There clearly is some selective pressure in these regions: centromere function must be maintained. CDEI and CDEIII are well-conserved. CDEII, while it is apparently somewhat free to vary in exact sequence, is selectively constrained in both length and composition. However, much of the supervariation is in the regions immediately flanking the point centromeres, and these regions might be entirely free of selective constraints, such as promoter sequences.

There are a number of possibilities for active processes that might produce centromere-adjacent supervariation. Tensional stresses induced during chromatid separation might lead to a high incidence of strand breaks around the point centromere. Centromere function might in some way impede the accurate replication of centromeric DNA, or the accurate repair. However, as noted in the introduction to centromere biology in Section 1.4, centromeric DNA does not appear to be replicated unusually early or late, and there is conflicting evidence regarding the efficiency of homologous repair mechanisms in the vicinity of centromeres in yeast.

It is of course perfectly possible, perhaps even probable, that centromere-adjacent supervariation actually arises through a combination of factors: an active mutational process producing a high rate of polymorphism coupled with a lack of selective pressure on the resulting sequence. It may not
be practical or possible to determine the precise genome dynamics at work in these regions, but it is an interesting discovery, and may also have practical application.

4.3.4. Practical applications of the phenomenon.

As Figure 4.8 demonstrates, it is possible to exploit centromere-adjacent supervariation in order to differentiate closely related strains of budding yeast, and to produce phylogenies. In this work a phylogeny was constructed based on the sequence around a single centromere. In some of the strains this sequence was identical, but that may not be the case for other centromeres. Sequencing across several centromeres at once (CENs 7, 16, 6, 9 and 15 all lie in intergenic regions that are less than 750bp long and so may be the easiest to sequence in a single step) may yield a 'centromeric fingerprint' that can distinguish similar strains even more effectively. High resolution phylogenies based on centromeric sequence appear to be reliable, but more work must be done to establish this beyond reasonable doubt, and it will be challenging to determine actual mutation rates unless we better understand the processes that produce centromere-adjacent supervariation.

4.3.5. Possibilities for further investigation.

As more and more yeast genomes become available for analysis, and the quality of sequence, coverage and annotation improves, it will be possible to study centromere-adjacent supervariation in still greater detail. It should be possible to survey the relative frequencies of different kinds of polymorphism,
and compare these to those in other regions such as ITSs and tandem repeats. This may yield clues about the processes responsible.

With more and more sequence available it will be possible to test the efficacy of centromeric fingerprints and their reliability (or otherwise) in producing phylogenies of budding yeast strains.

Improving knowledge of regulatory motifs and their distribution throughout the genomes will allow us to better understand their abundance in the centromeric neighbourhood and correlate this with levels of polymorphism and sequence divergence.

There are experiments already underway in which yeast populations are monitored over time, with isolates from different periods frozen for future reference and comparison. If centromeric polymorphism is sufficiently rapid then it is not out of the question that we may be able to catch it as it occurs, by comparing centromeres from different generations. The identification of mutant strains with accelerated centromere-adjacent polymorphism would go a long way towards determining the processes that are responsible. It is conceivable that one might engineer an assay system for centromere-adjacent supervariation by positioning marker genes in very close proximity.

It will also be interesting to observe whether similar phenomena are observed in more distantly related organisms, such as *K. lactis*, *K. waltii*, *E. gossypii* and further afield still. The mystery of the missing point centromeres in *S. castellii* certainly warrants further investigation. The unusual centromeres of *C. albicans* might also undergo a process of supervariation.

While the centromeres of higher eukaryotes such as *Homo sapiens* are very different in scale and structure, there remains a slim possibility that the
processes underlying centromere-adjacent supervariation will have some counterpart in our own species that the yeast model can help to explain. One never can tell where these things will lead.
CHAPTER 5: HYBRID INTROGRESSION OF

S. CEREVISIAE SEQUENCE INTO

EUROPEAN S. PARADOXUS
5.1: Introduction

As discussed in Section 1.5, subtelomeric regions are especially prone to rearrangements and duplications, and the S. cerevisiae genome contains a number of ‘subtelomeric homology regions’ that are products of this. There is substantial evidence that species exploit this phenomenon to generate phenotypic diversity.

The genome-wide analysis of intergenic divergence described in Chapter 4 revealed one very surprising anomaly at the left end of chromosome XIV in S. cerevisiae and S. paradoxus: a large segment of subtelomeric homology exhibiting much lower rates of sequence divergence than is observed in the rest of the genomes.

This unexpected phenomenon appeared to indicate some form of horizontal transfer of DNA between the species, long after the original speciation event, but further work was required to confirm whether the original sequencing was accurate, and to learn more about the direction, mode and implications of such a transfer.

5.2: Results

Figure 5.1 shows data from the genome-wide analysis of intergenic divergence between S. cerevisiae and S. paradoxus. When the P-distance of each gene and intergenic region was plotted at its position on each
chromosome, a substantial length of the left subtelomere of chromosome XIV proved to be much less divergent than the rest of the genomes.

Early versions of the scripts used to generate these data had earlier introduced a superficially similar anomaly (due to a failure to synchronize the versions of sequence data and annotation data retrieved from SGD, resulting in a misalignment of homologous features) but this error produced a lengthy section of chromosome with much greater-than-average sequence divergence, and was readily corrected. The anomaly in chromosome XIV could not be accounted for by programming error.

Two approaches were taken to confirm the initial result. Firstly, the sequences covering the region in question were aligned and their identity plotted using a different program: mVISTA. The results are shown in Figure 5.2. Secondly, the amino acid identity and conservation values (provided by Kellis et al. 2003) for genes in this region were checked. These are also shown in Figure 5.2.

These approaches confirmed the apparent existence of a 'cross-species transfer region', ~23kb long, which appeared to have passed horizontally between the sequenced strains of S. cerevisiae and S. paradoxus much more recently than when the rest of the genomes diverged.
Figure 5.1. Scatterplot of gene and intergenic P-distances between *S. cerevisiae* S288C and *S. paradoxus* CBS432 along the length of chromosome XIV. ABOVE: The entire chromosome, showing the general distribution of P-distances for intergenic regions (purple and red) and genes (green). Note the presence of some gaps due to missing sequence from the *S. paradoxus* genome project. LEFT: Magnification of the left subtelomeric region, highlighting the extremely low (but non-zero) level of sequence divergence across a region that spans ~23 kb and contains 12 genes. The data points for the rest of the chromosome are collapsed together on the right-hand side of this plot for the sake of comparison.
Figure 5.2. Plot of identity between *S. cerevisiae* S288C and *S. paradoxus* CBS432 in the left subtelomere of chr XIV. Created using mVISTA. The extent of the ‘cross-species transfer region’ is indicated, as is the region sequenced in multiple species to determine the direction and extent of the transfer (see text). ORFs are labelled, and on each major ORF two measures of protein identity are listed in white, using values provided by Kellis et al. (2003). The top figure is the percentage of amino acid identity, while the bottom figure is the percentage of amino acid conservation.
5.2.1. The Broad Institute sequence for *S. paradoxus* was accurate, and the region was passed to European strains of *S. paradoxus* from a strain of *S. cerevisiae*.

Although the *S. paradoxus* sequence in this region was not identical to that of *S. cerevisiae*, there remained a possibility that this phenomenon was due to some form of contamination in the Broad Institute sequencing. In order to check this, a subsection of the region (marked on Figure 5.2) in *S. paradoxus* CBS432 was resequenced by AGOWA genomics to check the accuracy of the existing sequence. At the same time, in order to determine the extent and direction of this horizontal transfer, the same subsection was also sequenced in a number of *S. cerevisiae* and *S. paradoxus* strains. This subsection was chosen in order to cover both genic and intergenic sequence. The resulting bidirectional reads were trimmed and combined using BioEdit.

Figure 5.3 shows a phylogenetic tree based on these sequences, and for the sake of comparison a second phylogenetic tree based on sequence covering CEN7 (see Chapter 4). Both trees were constructed using Mega 3.1 (neighbour-joining method using P-distance). The replicated sequencing of the subsection in *S. paradoxus* CBS432 was in close agreement with that produced by the Broad Institute. That they were not identical is probably an indicator of sequencing error, but the use of two different isolates of this strain (designated OS25 and OS142 in our strain collection; the latter derived from viable spores of the former) might also contribute. Two facts are evident from these phylogenies: firstly the ‘cross-species transfer region’ originates from *S. cerevisiae*, and secondly it is shared by multiple strains of *S. paradoxus*, all of which belong to the European subpopulation.
5.2.2. The cross-species transfer region does not extend all the way to the telomere, and synteny is conserved.

If the Broad Institute sequence assembly is correct then the cross-species transfer region does not extend all the way to the telomere, and there are several genes on the telomere side that show normal rates of divergence between the two species (see Figures 5.1 and 5.2). The cross-species transfer region does not interfere with synteny, which is conserved in both species. The thirteen genes that are wholly or partially contained within the cross-species transfer region are listed, along with their properties, in Table 5.1.
<table>
<thead>
<tr>
<th>GENE NAME &amp; coordinate on chr XIV</th>
<th>GENE FUNCTION</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>YNL332W (THI12) 14832-15854</td>
<td>Protein involved in synthesis of the thiamine precursor hydroxymethylpyrimidine (HMP); member of a subtelomeric gene family including THI5, THI11, THI12, and THI13</td>
<td></td>
</tr>
<tr>
<td>YNL331C (AAD14) 17248-16118</td>
<td>Putative ary-alkyl alcohol dehydrogenase with similarity to <em>P. chrysosporium</em> ary-alkyl alcohol dehydrogenase; mutational analysis has not yet revealed a physiological role</td>
<td></td>
</tr>
<tr>
<td>YNL330C (RPD3) 19302-18001</td>
<td>Histone deacetylase; regulates transcription and silencing</td>
<td></td>
</tr>
<tr>
<td>YNL329C (PEX6) 22633-19541</td>
<td>Peroxisomal membrane AAA-family ATPase peroxin required for peroxisome assembly, contains two 230 amino acid ATP-binding AAA cassettes, interacts with Pex1p</td>
<td></td>
</tr>
<tr>
<td>YNL328C (MDJ2) 22374-22834</td>
<td>Constituent of the mitochondrial import motor associated with the presequence translocase; function overlaps with that of Pam18p; stimulates the ATPase activity of Sac1p to drive mitochondrial import; contains a J domain</td>
<td></td>
</tr>
<tr>
<td>YNL327W (EGT2) 24047-27172</td>
<td>Glycosylphosphatidylinositol (GPI)-anchored cell wall endoglucanase required for proper cell separation after cytokinesis, expression is activated by Swi5p and tightly regulated in a cell cycle-dependent manner</td>
<td></td>
</tr>
<tr>
<td>YNL326C (PFA3) 28346-27336</td>
<td>Palmitoyltransferase for Vac8p, required for vacuolar membrane fusion; contains an Asp-His-His-Cys-cysteine rich (DHHC-CRD) domain; autacylates; required for vacuolar integrity under stress conditions</td>
<td></td>
</tr>
<tr>
<td>YNL325C (FIG4) 31377-28738</td>
<td>Protein required for efficient mating, member of a family of eukaryotic proteins that contain a domain homologous to Sac1p</td>
<td></td>
</tr>
<tr>
<td>YNL324W 31049-31444</td>
<td>Dubious ORF, overlapping FIG4</td>
<td></td>
</tr>
<tr>
<td>YNL323W (LEM3) 31943-33187</td>
<td>Membrane protein of the plasma membrane and ER, involved in translocation of phospholipids and alkylphosphocholine drugs across the plasma membrane</td>
<td></td>
</tr>
<tr>
<td>YNL322C (KRE1) 34234-33293</td>
<td>Cell wall glycoprotein involved in beta-glucan assembly; serves as a K1 killer toxin membrane receptor so confers susceptibility to this toxin.</td>
<td></td>
</tr>
<tr>
<td>YNL321W 34695-37421</td>
<td>Uncharacterised. Protein of unknown function, potential Cdc28p substrate</td>
<td></td>
</tr>
<tr>
<td>YNL320W 37699-39553</td>
<td>Uncharacterised. Hypothetical ORF</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1. Genes wholly or partially within the cross-species transfer region. Descriptions and details from SGD (6/9/2006). Genes with a function that might explain the selection for this region in European *S. paradoxus* are highlighted in bold.

5.2.3. The two ends of the cross-species transfer region are not delimited by any obvious sequence patterns.

In order to determine if the two boundaries of the cross-species transfer region contain any sequence features such as homology or tandem repeats that might explain why the region of elevated homology ends at those points, subsequences covering the two boundaries (*S. cerevisiae* chr XIV 14832-
17248, covering THI12 and AAD14, and S. cerevisiae chr XIV 34695-38553, covering YNL321W and YNL320W) were compared in a dot matrix plot. No such sequence features could be observed (results not shown): the boundaries appear to be arbitrary.

5.2.4. This region overlaps a subtelomeric duplication within S. cerevisiae.

Large subtelomeric duplications are common within S. cerevisiae (and probably within other budding yeast species as well; see Section 1.5). Approximately 17 kb of the left arm of chromosome XIV (1-17248) has strong homology to the left end of chromosome VI (1-15431), up to and including the homologues YNL331C (AAD14) and YFL056C (AAD6). The end of this intraspecies homology region closely corresponds with the start of the cross-species transfer region: YNL331C (AAD14) is the first complete gene, as we move along the chromosome arm from the telomere, to show abnormally high identity between S. cerevisiae and S. paradoxus, and this same gene is the last one in the subtelomeric duplication within S. cerevisiae.

Certain conclusions can be drawn from this about the relative timings of the horizontal transfer and the subtelomeric duplication (see the discussion below).
5.3: Discussion

To summarise the results above: a large (~23kb) section of subtelomeric DNA of *S. cerevisiae* origin has at some point been transferred to the ancestor of multiple European *S. paradoxus* strains, some time after the two species originally diverged. Two observations support the premise that this transfer did not occur recently. Firstly, the transferred region is at least partially diverged, both between *S. cerevisiae* and European *S. paradoxus* and within subpopulations of the latter. Indeed the observed divergence is sufficient to suggest that the transfer may predate human domestication of yeast, though at present there is insufficient sequence data to assert this with any confidence. Secondly, the transferred sequence has propagated through *S. paradoxus* strains to cover a large area of Europe (strain N17 originates from Tartarstan in south-west Russia while strain Q4.1 originates from London).

5.3.1. Horizontal transfer by hybrid introgression

While there are many examples in the natural world of true bacterially- or virally-mediated lateral transfer of DNA between separate species, there are no known examples of this occurring in yeast (without the intervention of human geneticists), nor are there any known bacterial or viral vectors that might mediate such a transfer in yeast (Dujon 2005). Moreover the DNA region involved here is considerably larger than is typical of lateral transfer, and does not extend all the way to the telomere.
On the other hand there is strong evidence that sensu stricto yeast species can mate and form hybrids (de Barros Lopes et al. 2002; Liti et al. 2005; 2006) albeit with initially reduced fertility. It is also possible to find S. cerevisiae and S. paradoxus isolates co-existing in the same natural microhabitats (Sniegowski et al. 2002; Sweeney et al. 2004). Initial hybrid viability might be low but in a sufficiently large natural population there are probably numerous instances of hybrids.

For these reasons we favour a model of hybrid introgression: a rare viable S. cerevisiae / S. paradoxus hybrid formed in Europe, perhaps with a majority of S. paradoxus sequence. Subsequent backcrossing with the parent S. paradoxus population would have removed much of the S. cerevisiae DNA from this lineage, and the cross-species transfer region observed is one remaining relic of this process. While it is possible that this region became fixed in the European population by chance alone, it is also possible that this sequence conferred a selective advantage on the hybrid’s ancestors.

5.3.2. Possible selective advantage conferred by the S. cerevisiae sequence?

The sequencing of a segment partially covering the genes PEX6 and MDJ2 in multiple strains shows that these genes, at least, are present in non-European S. paradoxus, but the sequencing is not extensive enough to confirm if both genes are fully functional in these other strains, nor whether other genes in the region are present and functional. Given the general mutability of subtelomeric sequences in yeast it should not be taken for granted that they are.
Thus it is possible that European *S. paradoxus* acquired from *S. cerevisiae* one or more genes that had been lost or altered in the original *S. paradoxus* ancestor, and that possession of the *S. cerevisiae* version of these genes conferred a selective advantage on the ancestral hybrid.

A survey of the genes contained in the cross-species transfer region (listed in Table 5.1) produces two credible candidates that might be responsible for such an advantage.

*FIG4*, though poorly characterised, is required for efficient mating in *S. cerevisiae*. Therefore it is possible that the acquisition of the *S. cerevisiae* *FIG4* gene improved the mating efficiency of the ancestral hybrid and its ancestors over that of the native *S. paradoxus* subpopulation.

*KRE1* is a cell-wall glycoprotein that also renders cells susceptible to the K1 killer toxin. This toxin is produced by yeasts carrying the M1-dsRNA 'killer virus' that also makes the carrier resistant to the toxins it produces (see Section 3.2). *S. cerevisiae* SK1 is one strain that carries these viruses (its designation standing for 'superkiller 1'; Vodkin *et al.* 1974) and a large number of other killer strains originate from Sicily, amusingly enough (D. Cavalieri, personal communication).

While possession of a mutant form of *KRE1* might confer immunity to the killer toxin it may also produce a selective disadvantage when there are no killer toxins in the environment, in much the same way as a mutant form of haemoglobin can protect against malaria but leaves the sufferer with sickle cell anaemia. Thus the ancestral *S. paradoxus* might have developed killer toxin immunity through a mutation of *KRE1*, but in the absence of subsequent
selective pressure from killer toxins, the functional *S. cerevisiae* version of *KRE1* spread throughout the European population.

In order to determine if there was any correlation between killer toxin susceptibility or immunity and possession of the *S. cerevisiae* sequence, we tested the strains listed in Figure 5.3 for susceptibility to two killer yeast strains (one carrying M1-dsRNA and the other carrying an alternate killer virus called M2-dsRNA) kindly provided by Duncan Greig. The American and Far-Eastern *S. paradoxus* strains tested were indeed all resistant to both the killer strains, as were *S. cerevisiae* strains YPS128 and SK1 (the latter as expected). However within the European *S. paradoxus* strains there was a mixture of responses: CBS432 and Q4.1 were susceptible to the killer toxins (as were the other *S. cerevisiae* strains tested), but *S. paradoxus* DBVPG6565 and N17 were resistant.

The lack of a neat correlation in these strains counts against the hypothesis of *KRE1*-related selective pressure, but cannot dismiss it entirely. Differences in ploidy or other genetic factors might have contributed to differences in resistance. The *S. cerevisiae* version of *KRE1* in European *S. paradoxus* might have become non-functional in some strains at a later date (note that in Figure 5.2 the biggest dip in identity across the entire cross-species transfer region occurs in the middle of the *KRE1* gene). The existence of a range of different killer plasmids and killer toxins could also confuse the picture. And finally it is quite possible that killer toxin resistance is just one of several factors contributing to the putative selective advantage.
5.3.3. Hybrid introgression event occurred after the subtelomeric duplication event in *S. cerevisiae*.

As mentioned in Section 5.2, it is possible to deduce, assuming that the Broad Institute sequence assembly for the region is correct, that the subtelomeric duplication responsible for the extensive homology between the subtelomeres of chr XIV-L and chr VI-L probably preceded the hybridisation event that produced European *S. paradoxus*. Figure 5.4 compares the two possible orders of events. The existence of normally divergent genes at the telomere end of the cross-species transfer region precludes scenario A.

5.3.4. Possibilities for further investigation.

Further sequencing will be required to determine if all known European *S. paradoxus* strains contain the transferred region, or just a subset, and if the same length of transferred sequence is present in all cases. With more complete sequence covering the left end of chromosome XIV in the non-European *S. paradoxus* strains, it will be possible to better study the gene differences that might have resulted in a selective advantage for the *S. cerevisiae* sequence. If *KRE1*, for example, is demonstrably a pseudogene in these other strains then this will lend credence to the theory that killer toxin susceptibility and resistance have driven selection for the *S. cerevisiae* sequence in European *S. paradoxus*. 
Figure 5.4. Possible order of subtelomeric duplication and hybridisation events. The fact that there is normally divergent sequence at the telomeric end of the cross-species transfer region (CTR), as shown in Figure 5.2, suggests that subtelomeric duplication occurred before the hybridisation event that produced the European strains of \textit{S. paradoxus}, not after.

If further evidence supports the theory, it would be intriguing to conduct experiments in which \textit{FIG4}, \textit{KRE1} or other subsections of the \textit{S. cerevisiae} cross-species transfer region are systematically transplanted to non-European \textit{S. paradoxus} strains, and to test for the effect of these transfers on their relative fitness.

Interestingly, it has recently come to the author’s attention that CBS5829 (a European \textit{S. paradoxus} strain from Denmark) is, according to annotations in the CBS database, a killer yeast. On the basis of sequencing across CEN7 (see Figure 4.8) it is very similar to London strain Q4.1 which was found to be susceptible to killer toxins, but unfortunately CBS5829 was not one of the strains for which cross-species transfer region sequence was
acquired. It would be fascinating to correct this oversight, and see which version of \textit{KRE1} is present in this strain.

Additional sequencing will also be necessary to properly determine the rate of divergence following the horizontal transfer and to estimate the timing of the hybridisation event.

The wealth of new sequence information for a huge range of \textit{sensu stricto} yeast strains may well reveal other cross-species transfer regions, which will shed new light on the evidently complex relationships between different yeast species. It can be hoped that with sufficient data we may be able to trace in their entirety the exact sequence of translocations, duplications and horizontal transfers that have shaped \textit{S. cerevisiae} and its relatives, building a comprehensive and unprecedented picture of genome evolution and speciation.
CHAPTER 6: CONCLUSIONS
This thesis has described two discoveries: centromere-adjacent supervariation at budding yeast centromeres (Chapter 4), and hybrid introgression between budding yeast species (Chapter 5). It also delves into other aspects of genome evolution such as whole-genome duplications and subtelomeric duplications (reviewed in Chapter 1) and retrotransposition (in Chapter 3).

Even in these relatively simple organisms, genome evolution proves to be more complex and multifaceted with each new study. Species boundaries are blurred. Large amounts of DNA have entered yeast genomes horizontally, either through hybrid introgression from related species or through true lateral transfer of retroviral DNA.

Some parts of the genomes are much more variable than others. Subtelomeric translocations and duplications generate enormous diversity at chromosome ends, and the resultant phenotypic diversity must greatly improve the ability of these species to diversify and adapt to new ecological and metabolic niches, promoting increasing reproductive isolation and eventually speciation.

Non-coding regions of genome understandably tend to diverge more rapidly than coding regions. In the vicinity of centromeres, this divergence is, for unknown reasons, even more rapid than usual, despite (or perhaps in some sense because of) the immediate proximity of motifs that are unequivocally vital to a yeast cell's survival. This centromere-adjacent supervariation might even in some cases force a radical reinvention of the form and function of centromeres (as would appear to be the case in S. castellii).
This complexity of genome evolution presents some enormous challenges. More complex models of population genetics are required to account for the contribution of nonstandard gene flow. The shaping of genomes by rare and unusual events (such as the hybridisation event in the ancestry of European *S. paradoxus*) is as confounding for generalised models of evolution as cataclysmic events such as meteor strikes are for models of geology based on Lyellian uniformitarianism and gradualism. Accurate phylogenies will no longer be simple trees but networks with different branches for different parts of the genome. To produce such phylogenies with confidence will require more sequence data and more complex underlying calculations.

Eventually, perhaps, it will be possible with sufficient data to reconstruct in great detail the sequence of events that has shaped a given genome, down to the order and approximate timing of subtelomeric rearrangements, transposon insertions and even some smaller-scale mutations. Computationally this is very challenging even in simplified cases (see Figure 6.1) and akin to cryptoanalysis in some ways.

More detailed understanding of the processes that lead to sequence divergence will feed back to the start of the circle: to produce algorithms that can detect homology more accurately by more closely modelling the actual biological processes at work.
Figure 6.1. The challenge of true ‘whole genome phylogenies’. Given two sufficiently related ‘genomes’ X and Y, is it possible to accurately reconstruct the chain of events through which they diverged, and the state of the ancestral genome? Can this ancestral genome then be used for still more ancient reconstructions?

6.1 Future sequencing of yeast genomes.

The genome of *S. cerevisiae* S288C is entirely assembled and almost entirely complete, except for the very ends of some Y' elements and the full length of telomeric repeats, which is in any case variable.

In other yeast genome projects, such as those for *C. glabrata, K. lactis, E. gossypii, D. hansenii* and *S. pombe*, the sequences have been assembled into full chromosomes, but rarely do those sequences extend all the way to each telomere because of the difficulties inherent in cloning and assembling chromosome ends.

The comparison of these distant genomes to *S. cerevisiae* is not without utility, especially when it comes to the characterisation and annotation of their ORFs, but it is the new wave of sequencing projects for closely related yeasts (naturally focussing on close relatives of *S. cerevisiae* S288C, since this genome project provides such a detailed reference) that offer the best
opportunities for understanding the details of genome evolution. Sadly, the
genome sequences generated by the Broad Institute (Kellis et al. 2003; 2004)
and Washington University (Cliften et al. 2003) contain many gaps and are
not fully assembled into chromosomes, and in these projects too the
telomeres are largely undescribed.

The study of telomeres and subtelomeric genome dynamics would
clearly benefit from a concerted effort to fully finish all of these genome
projects, but additional telomere-specific sequencing will certainly be required.

Work is already underway, primarily at the Sanger Centre, to
sequence, to a high degree of coverage, the genomes of a large set of other
*S. cerevisiae* and *S. paradoxus* strains. Like the Broad Institute and
Washington University initiatives, these genome projects are unlikely to be
finished to the chromosome level, at least in the immediate future, but even so
they may provide answers to at least some of the questions raised by this
work, as discussed in Chapters 3-5.

---

### 6.2. Human scrutiny of large bioinformatic datasets.

The field of genetics is increasingly dominated by large-scale projects:
not only the many genome-sequencing projects but also systematic deletion
studies and genome-wide expression studies, all made possible by extensive
automation using robots and computers. Indeed such studies generate such
vast quantities of data that computers are required not only to generate and
collate the data but analyse it as well. But computers can only search for
information that they were programmed to search for. They can of course still
generate interesting and even unexpected results, but only within the narrow
boundaries of their attention, and it still requires a human, familiar with biological science, to recognise that such results are interesting or unexpected.

This work demonstrates that broad analyses can (understandably) overlook significant details that it would take a human browser to detect, and that benefits can be accrued from giving researchers the time and freedom to simply explore. There may be many more such potential discoveries hidden away in the data. It is also a demonstration that bioinformaticians who are biologists first and computer-programmers second can still make a contribution. Indeed any biologist could explore the data and make valuable discoveries, if only they were given the right tools.

6.3. The present and future of bioinformatics.

It is regrettable that the field of bioinformatics remains so anarchic and impenetrable. There are many different programs, using many different complex algorithms, written in many different programming languages for many (but rarely all) different operating systems, presenting a host of very different and often confusing user interfaces, rarely recognising as input more than a handful of the many different data formats, and generating just as bewildering an array of output formats.

It is small wonder that many biologists feel intimidated by the computational tools being produced: the tools are not being produced for them. Most of the minds behind such software are trained in mathematics and computer programming more than biology, or are well-versed in both but no longer able to understand the needs, capabilities and frustrations of the pure
biologist. Because of this, many useful programs remain unexploited by a large section of their intended audience, and huge amounts of data remain unexplored.

Training more biologists in information technology and bioinformatics will only go so far towards solving this problem. Even if they can understand and use a wide range of tools, their productivity is limited by the mutual incompatibility of those tools and formats, and the limits of their individual interests and research briefs.

In the view of this author, the field of bioinformatics would benefit from a Bill Gates: a leader with the vision and capability to put a popular, versatile and unified ‘biological operating system’ on the desktop of every biologist, with an intuitive graphically-driven user interface and a clear syntax for framing complex queries. Though the capabilities of such a program would doubtless be expanded by the community, all such developments would be required to fit a conceptual framework and user interface that has been designed for (and with the input of) ordinary biologists with an average level of computer literacy.

Even if this cannot be created in one fell swoop, perhaps there is hope that the bioinformatics field will someday evolve towards it. There would be a certain gentle irony in that.
The CD-ROM accompanying this thesis (CD-R data cd; should be compatible with Windows XP and Mac OS-X) contains various supplementary files, organised into the following folders:

"01 Thesis electronic copy" contains electronic versions of this thesis, as both Microsoft Word and Adobe PDF documents, with functional hyperlinks.

"02 Source genome data" contains genome sequence data and annotation data (where available) for 15 ascomycetous yeast species sequenced by various different groups, including 2 versions of the *S. bayanus* genome which are not necessarily from the same strain. The files are from various different sources as described in Section 2.2, and in various different formats.

"03 Sequence elements" contains various additional sequences, such as yeast plasmids and viruses, as well as the subsequences of various features such as telomeres, telomere-associated elements, and centromeres from various species.

"04 Software" contains Windows executables for various pieces of bioinformatics software mentioned in this thesis. Most are installation packages, but some are direct executables that require no formal installation, and in case of key web-based applications, internet shortcut files are provided.

"05 Genome-wide analysis scripts" contains some of the key scripts written by the author for the use in this work: early and late versions of the genome-wide analysis scripts, and some supplemental scripts and third-party Python modules required for their function.

"06 Generated data" includes the raw and assembled sequence data from AGOWA sequencing, and output files produced by the genome-wide analysis.
APPENDIX II: PUBLICATIONS AND PRESENTATIONS ARISING
Publications


Conference Posters

Barton DBH, Louis EJ (2003) ">300 Genes That May Not Be Real" (poster); British Yeast Group Meeting 2003; University of Leicester, UK, March 24-26 2003.

Barton DBH, Louis EJ (2004) "Unexpected Divergence around Yeast Centromeres Observed Through Multiple Genome Comparisons" (poster); Yeast Chromosome Structure, Replication and Segregation FASEB Summer Research Conference; Pine Mountain, Georgia, USA; July 10-15 2004.

Conference Talks

Barton DBH (2005) "Multiple Genome Comparisons Show Unexpected Diversity Around Yeast Centromeres, and Hybrid Introgression of S. cerevisiae DNA into European Strains of S. paradoxus" (talk); Comparative Genomics of Eukaryotic Microorganisms ESF-EMBO Symposium; San Feliu de Guixols, Spain, November 12-17 2005.


